INTERNATIONAL JOURNAL

of

COMPUTERS, COMMUNICATIONS & CONTROL

With Emphasis on the Integration of Three Technologies

IJCCC
A Quarterly Journal

Year: 2009 Volume: IV Number: 2 (June)

Agora University Editing House

CCC Publications

Licensed partner: EBSCO Publishing

www.journal.univagora.ro

Publisher & Editorial Office
CCC Publications, Agora University
Piata Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel: +40 259 427 398, Fax: +40 259 434 925, E-mail: ccc.journal@gmail.com
Website: www.journal.univagora.ro
ISSN 1841-9836, E-ISSN 1841-9844

International Journal of Computers, Communications and Control (IJCCC) is published from 2006 and has 4 issues/year (March, June, September, December), print & online.

Founders of IJCCC: I. Dziţac, F.G. Filip and M.J. Manolescu (2006)

This publication is subsidized by:

1. Agora University
2. The Romanian Ministry of Education and Research / The National Authority for Scientific Research

CCC Publications, powered by Agora University Publishing House, currently publishes the "International Journal of Computers, Communications & Control" and its scope is to publish scientific literature (journals, books, monographs and conference proceedings) in the field of Computers, Communications and Control.

IJCCC is indexed and abstracted in a number of databases and services including:

1. ISI Thomson Reuters - Science Citation Index Expanded (also known as SciSearch®)
2. ISI Thomson Reuters - Journal Citation Reports/Science Edition
3. ISI Thomson Reuters - Master Journal List
4. SCOPUS
5. Computer & Applied Science Complete (EBSCO)
6. Vocational Studies Complete (EBSCO)
7. Current Abstracts (EBSCO)
8. Collection of Computer Science Bibliographies(CCSB)
9. Informatics portal io-port.net (FIZ KARLSRUHE)
10. MathSciNet
11. Open J-Gate
12. Google Scholar
13. Romanian Journals CNCSIS (cat.A, cod 849)
14. Information Systems Journals (ISJ)
15. Urlich's Periodicals Directory
16. Genamics JournalSeek
17. ISJ- Journal Popularity
18. Magistri et Scholares
19. SCIRUS
20. DOAJ

Contents

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 104-117

A Novel Fuzzy ARTMAP Architecture with Adaptive Feature Weights based on Onicescu's Informational Energy

Răzvan Andonie, Lucian Mircea Sasu, Angel Caţaron

Răzvan Andonie
Computer Science Department
Central Washington University, Ellensburg, USA
and
Department of Electronics and Computers
Transylvania University of Braşov, Romania
E-mail: andonie@cwu.edu

Angel Caţaron
Department of Electronics and Computers
Transylvania University of Braşov, Romania
E-mail: cataron@vega.unitbv.ro

Lucian Mircea Sasu
Applied Informatics Department
Transylvania University of Braşov, Romania
E-mail: lmsasu@unitbv.ro

Abstract: Fuzzy ARTMAP with Relevance factor (FAMR) is a Fuzzy ARTMAP (FAM) neural architecture with the following property: Each training pair has a relevance factor assigned to it, proportional to the importance of that pair during the learning phase. Using a relevance factor adds more flexibility to the training phase, allowing ranking of sample pairs according to the confidence we have in the information source or in the pattern itself.
We introduce a novel FAMR architecture: FAMR with Feature Weighting (FAMRFW). In the first stage, the training data features are weighted. In our experiments, we use a feature weighting method based on Onicescu's informational energy (IE). In the second stage, the obtained weights are used to improve FAMRFW training. The effect of this approach is that category dimensions in the direction of relevant features are decreased, whereas category dimensions in the direction of non-relevant feature are increased. Experimental results, performed on several benchmarks, show that feature weighting can improve the classification performance of the general FAMR algorithm.
Keywords: Fuzzy ARTMAP, feature weighting, LVQ, Onicescu's informational energy.

1 Introduction

The FAM architecture is based upon the adaptive resonance theory (ART) developed by Carpenter and Grossberg [7]. FAM neural networks can analyze and classify noisy information with fuzzy logic, and can avoid the plasticity-stability dilemma of other neural architectures. The FAM paradigm is prolific and there are many variations of Carpenter's *et al.* [7] initial model: ART-EMAP [9], dARTMAP [8], Boosted ARTMAP [27], Fuzzy ARTVar [12], Gaussian ARTMAP [28], PROBART [21], PFAM [20],

Ordered FAM [11], and μARTMAP [14]. The FAM model has been incorporated in the MIT Lincoln Lab system for data mining of geospatial images because of its computational capabilities for incremental learning, fast stable learning, and visualization [25].

One way to improve the FAM algorithm is to generalize the distance measure between vectors [10]. Based on this principle, we introduced in previous work [2] a novel FAM architecture with distance measure generalization: FAM with Feature Weighting (FAMFW). Feature weighting is a feature importance ranking algorithm where weights, not only ranks, are obtained. In our approach, training data feature weights were first generated. Next, these weights were used by the FAMFW network, generalizing the distance measure. Potentially, any feature weighting method can be used, and this makes the FAMFW very general.

Feature weighting can be achieved, for example, by LVQ type methods. Several such techniques have been recently introduced. These methods combine the LVQ classification with feature weighting. In one of these approaches, RLVQ (Relevance LVQ), feature weights were determined to generalize the LVQ distance function [16]. A modification of the RLVQ model, GRLVQ (Generalized RLVQ), has been proposed in [18]. The SRNG (Supervised Relevance Neural Gas) algorithm [17] combines the NG (Neural Gas) algorithm [22] and the GRLVQ. NG [22] is a neural model applied to the task of vector quantization by using a neighborhood cooperation scheme and a soft-max adaptation rule, similar to the Kohonen feature map.

In [1], we introduced the Energy Supervised Relevance Neural Gas (ESRNG) feature weighting algorithm. The ESRNG is based on the SRNG model. It maximizes Onicescu's IE as a criteria for computing the weights of input features. The ESRNG is the feature weighting algorithm we used in [2], in combination with our FAMFW algorithm .

FAMR is a FAM incremental learning system introduced in our previous work [4]. During the learning phase, each sample pair is assigned a relevance factor proportional to the importance of that pair. The FAMR has been successfully applied to classification, probability estimation, and function approximation. In FAMR, the relevance factor of a training pair may be user-defined, or computed, and is proportional to the importance of the respective pair in the learning process.

In the present paper, we focus on the FAMR neural network, the ESRNG feature weighting algorithm, and the distance measure generalization principle. We contribute the following:

1. We introduce a novel FAMR architecture with distance measure generalization: FAMR with Feature Weighting (FAMRFW), adapting the FAMFW model for the FAMR case.

2. Compared to [2], we include new experiments on standard benchmarks.

We first introduce the basic FAM and FAMR notations (Section 2), and the ESRNG feature weighting algorithm (Section 3). In Section 4, we describe the new FAMRFW algorithm, which uses a weighted distance measure. Section 5 contains experimental results performed with the FAMRFW method. Section 6 contains the final remarks.

2 A brief description of the FAMR

We will summarize the FAM standard architecture and the FAMR learning mechanism, which differentiates it from the standard FAM.

2.1 The FAM architecture

A detailed FAM description can be found in Carpenter's *et al.* seminal paper [7], but more simplified presentations are given in [26] and [19].

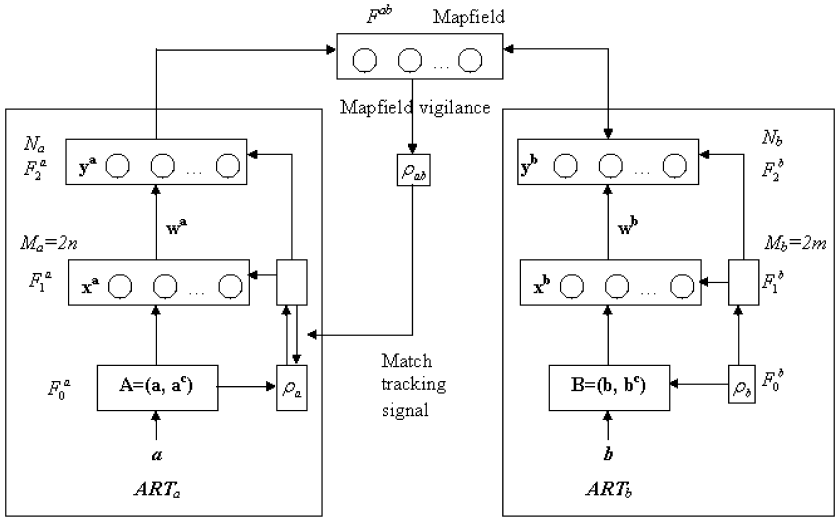


Figure 1: Fuzzy ARTMAP architecture [7].

The FAM architecture consists of a pair of fuzzy ART modules, ART_a and ART_b, connected by an inter–ART module called Mapfield (see Fig. 1). ART_a and ART_b are used for coding the input and output patterns, respectively, and Mapfield allows mapping between inputs and outputs. The ART_a module contains the input layer F_1^a and the competitive layer F_2^a. A preprocessing layer F_0^a is also added before F_1^a. Analogous layers appear in ART_b.

The initial input vectors have the form: $\mathbf{a} = (a_1, \ldots, a_n) \in [0,1]^n$. A data preprocessing technique called *complement coding* is performed by the F_0^a layer in order to avoid node proliferation. Each input vector $\mathbf{a}$ produces the normalized vector $\mathbf{A} = (\mathbf{a}, \mathbf{1} - \mathbf{a})$ whose L_1 norm is constant: $|\mathbf{A}| = n$.

Let M_a be the number of nodes in F_1^a and N_a be the number of nodes in F_2^a. Due to the preprocessing step, $M_a = 2n$. The weight vector between F_1^a and F_2^a is $\mathbf{w}^a$. Each F_2^a node represents a class of inputs grouped together, denoted as a *category*. Each F_2^a category has its own set of adaptive weights stored in the form of a vector $\mathbf{w}_j^a, j = 1, \ldots N_a$, whose geometrical interpretation is a hyper-rectangle inside the unit box. Similar notations are used for the ART_b module. For a classification problem, the class index is the same as the category number in F_2^b, thus ART_b can be substituted with a vector.

The Mapfield module allows FAM to perform associations between ART_a and ART_b categories. The number of nodes in Mapfield is equal to the number of nodes in F_2^b. Each node j from F_2^a is linked to each node from F_2^b via a weight vector $\mathbf{w}_j^{ab}$.

The learning algorithm is sketched below. For each training pattern, the vigilance parameter factor ρ_a is set equal to its baseline value, and all nodes are not inhibited. For each (preprocessed) input $\mathbf{A}$, a fuzzy choice function is used to get the response for each F_2^a category:

$$T_j(\mathbf{A}) = \frac{|\mathbf{A} \wedge \mathbf{w}_j^a|}{\alpha_a + |\mathbf{w}_j^a|}, \quad j = 1, \ldots, N_a \tag{1}$$

Let J be the node with the highest value computed as in (1). If the resonance condition from eq. (2) is not fulfilled:

$$\rho(\mathbf{A}, \mathbf{w}_J^a) = \frac{|\mathbf{A} \wedge \mathbf{w}_J^a|}{|\mathbf{A}|} \geq \rho_a, \tag{2}$$

then the Jth node is inhibited such that it will not participate to further competitions for this pattern and a new search for a resonant category is performed. This might lead to creation of a new category in ART_a.

A similar process occurs in ART_b and let K be the winning node from ART_b. The F_2^b output vector is set to:

$$y_k^b = \begin{cases} 1, & \text{if } k = K \\ 0, & \text{otherwise} \end{cases} \qquad k = 1, \ldots, N_b \tag{3}$$

An output vector $\mathbf{x}^{ab}$ is formed in Mapfield: $\mathbf{x}^{ab} = \mathbf{y}^b \wedge \mathbf{w}_j^{ab}$. A Mapfield vigilance test controls the match between the predicted vector $\mathbf{x}^{ab}$ and the target vector $\mathbf{y}^b$:

$$\frac{|\mathbf{x}^{ab}|}{|\mathbf{y}^b|} \geq \rho_{ab} \tag{4}$$

where $\rho_{ab} \in [0, 1]$ is a Mapfield vigilance parameter. If the test from (4) is not passed, then a sequence of steps called match tracking is initiated (the vigilance parameter ρ_a is increased and a new resonant category will be sought for ART_a); otherwise learning occurs in ART_a, ART_b, and Mapfield:

$$\mathbf{w}_J^{a(new)} = \beta_a \left(\mathbf{A} \wedge \mathbf{w}_J^{a(old)} \right) + (1 - \beta_a) \mathbf{w}_J^{a(old)} \tag{5}$$

(and the analogous in ART_b) and $\mathbf{w}_{Jk}^{ab} = \delta_{kK}$, where δ_{ij} is Kronecker's delta. With respect to β_a, there are two learning modes: *i)* fast learning for $\beta_a = 1$ for the entire training process, and *ii)* fast-commit and slow-recode learning corresponds to setting $\beta_a = 1$ when creating a new node and $\beta_a < 1$ for subsequent learning.

2.2 The FAMR learning mechanism

The main difference between the FAMR and the original FAM is the updating scheme of the w_{jk}^{ab} weights. The FAMR uses the following iterative updating [4]:

$$w_{jk}^{ab(new)} = \begin{cases} w_{jk}^{ab(old)} & \text{if } j \neq J \\ w_{JK}^{ab(old)} + \frac{q_t}{Q_J^{new}} \left(1 - w_{JK}^{ab(old)} \right) & \\ w_{Jk}^{ab(old)} \left(1 - \frac{q_t}{Q_J^{new}} \right) & \text{if } k \neq K \end{cases} \tag{6}$$

where q_t is the relevance assigned to the tth input pattern ($t = 1, 2, \ldots$), and $Q_J^{new} = Q_J^{old} + q_t$. The *relevance* q_t is a real positive finite number directly proportional to the importance of the experiment considered at step t. This w_{jk}^{ab} approximation is a correct biased estimator of the posterior probability $P(k|j)$, the probability of selecting the k-th ART_b category after having selected the j-th ART_a category [4].

Let $\mathbf{Q}$ be the vector $[Q_1 \ldots Q_{N_a}]$; initially, each Q_j ($1 \leq j \leq N_a$) has the same initial value q_0. N_a and N_b are the number of categories in ART_a and ART_b, respectively. These are initialized at 0. For incremental learning of one training pair, the FAMR Mapfield learning scheme is described by Algorithm 1. The vigilance test is:

$$N_b \, w_{JK}^{ab} \geq \rho_{ab} \tag{7}$$

For a clearer presentation, not to create a confusion between vector relevancies and feature weights, we will assume in all our following experiments that relevancies are set to a constant positive value. Since we actually do not use relevances, is this FAMR equivalent to the standard FAM model, as introduced in [7]? The answer is no, because, unlike the standard FAM: *i)* the FAMR accepts one-to-many relationships; and *ii)* the FAMR is a conditional probability estimator, with an estimated convergence rate computed in [4].

Algorithm 1 The t-th iteration in the FAMR Mapfield algorithm [4].

Step 1. Accept the t-th vector pair $(\mathbf{a}, \mathbf{b})$ with relevance factor q_t.
Step 2. Find a resonant category in ART_b or create a new one.
if $|\mathbf{b} \wedge \mathbf{w}_k^b|/|\mathbf{b}| < \rho_b$, for $k = 1, \ldots, N_b$ **then**
  $N_b = N_b + 1$ {add a new category to ART_b}
  $K = N_b$
  if $N_b > 1$ **then**
    $w_{jK}^{ab} = \frac{q_0}{N_b Q_j}$, for $j = 1, \ldots, N_a$ {append new component to $\mathbf{w}_j^{ab}$}
    $w_{jk}^{ab} = w_{jk}^{ab} - \frac{w_{jK}^{ab}}{N_b - 1}$, for $k = 1, \ldots, K-1$; $j = 1, \ldots N_a$ {normalize}
  end if
else
  Let K be the index of the ART_b category passing the resonance condition and with maximum activation function.
end if
Step 3. Find a resonant category in ART_b or create a new one.
if $|\mathbf{a} \wedge \mathbf{w}_j^a|/|\mathbf{a}| < \rho_a$, for $j = 1, \ldots, N_a$ **then**
  $N_a = N_a + 1$ {add a new category to ART_a}
  $J = N_a$
  $Q_J = q_0$ {append new component to $\mathbf{Q}$}
  $w_{Jk}^{ab} = 1/N_b$, for $k = 1, \ldots, N_b$ {append new row to $\mathbf{w}^{ab}$}
else
  Let J be the index of the ART_a category passing the resonance condition and with maximum activation function.
end if
Step 4. J, K are winners or newly added nodes. Check if match tracking applies.
if vigilance test (7) is passed **then**
  {learn in Mapfield}
  $Q_J = Q_J + q_t$
  $w_{JK}^{ab} = w_{JK}^{ab} + \frac{q_t}{Q_J}(1 - w_{JK}^{ab})$
  $w_{Jk}^{ab} = w_{Jk}^{ab}\left(1 - \frac{q_t}{Q_J}\right)$, for $k = 1, \ldots, N_b$, $k \neq K$
else
  perform match tracking and restart from step 3
end if

3 The ESRNG feature weighting algorithm

We use the ESRNG feature weighting algorithm to compute the generalized distance measure in the FAMRFW. Details of the ESRNG algorithm can be found in [1]. Is is based on Onicescu's IE, and approximates the unilateral dependency of random variables by Parzen windows approximation. Before outlining the principal steps of the ESRNG method, we review the basic properties of the IE.

3.1 Onicescu's informational energy

For a discrete random variable X with probabilities p_k, the IE was introduced in 1966 by Octav Onicescu [24] as $E(X) = \sum_{k=1}^{n} p_k^2$. For a continuous random variable Y, the IE was defined by Silviu Guiaşu [15]:

$$E(Y) = \int_{-\infty}^{+\infty} p^2(\mathbf{y}) d\mathbf{y},$$

where $p(\mathbf{y})$ is the probability density function.

For a continuous random variable Y and a discrete random variable C, the conditional IE is defined as:

$$E(Y|C) = \int_{\mathbf{y}} \sum_{m=1}^{M} p(c_m) p^2(\mathbf{y}|c_m) d\mathbf{y}.$$

In order to study the interaction between two random variables X and Y, the following measure of unilateral dependency was introduced by Andonie *et al.* [3]:

$$o(Y, X) = E(Y|X) - E(Y)$$

with the following properties:

1. o is not symmetrical with respect to its arguments;
2. $o(Y, X) \geq 0$ and the equality holds iff Y and X are independent;
3. $o(Y, X) \leq 1 - E(Y)$ and the equality holds iff Y is completely dependent on X.

This measure quantifies the unilateral dependence characterizing Y with respect to X and corresponds to the amount of information detained by X about Y.

3.2 The feature weighting procedure

ESRNG is an online algorithm which adapts a set of LVQ reference vectors by minimizing the quantization error. At each iteration, it also adapts the input vector feature weights. The core of the method is based on the maximization of the $o(Y, C)$ measure.

To connect input vector $\mathbf{x}_i$ with its class j, represented by vector $\mathbf{w}_j$, we use a simple transform. We consider a continuous random variable Y with its samples $\mathbf{y}_i = \lambda \mathbf{I}(\mathbf{x}_i - \mathbf{w}_j), i = 1, \ldots, N$, where:

- λ is the vector of weights;
- $\mathbf{x}_i, i = 1, \ldots, N$, are the training vectors, each of them from one of the classes $c_1, c_2, \ldots, c_M$;
- $\mathbf{w}_j, j = 1, \ldots, P$, are the LVQ determined class prototypes.

Assuming that the M class labels are samples of a discrete random variable denoted by C, we can use gradient ascend to iteratively update the feature weights by maximizing $o(Y, C)$:

$$\lambda^{(t+1)} = \lambda^{(t)} + \alpha \sum_{i=1}^{N} \frac{\partial o(Y, C)}{\partial \mathbf{y}_i} \mathbf{I} (\mathbf{x}_i - \mathbf{w}_j) .$$

From the definition of $o(Y, X)$, we obtain:

$$o(Y, C) = E(Y|C) - E(Y) = \sum_{p=1}^{M} \frac{1}{p(c_p)} \int_{\mathbf{y}} p^2(\mathbf{y}, c_p) d\mathbf{y} - \int_{\mathbf{y}} p^2(\mathbf{y}) d\mathbf{y}. \tag{8}$$

This expression involves a considerable computational effort. Therefore, we approximate the probability densities from the integrals using the Parzen windows estimation method. The multidimensional Gaussian kernel is [13]:

$$G(\mathbf{y}, \sigma^2 \mathbf{I}) = \frac{1}{(2\pi)^{\frac{d}{2}} \sigma^d} \cdot e^{-\frac{\mathbf{y}^t \mathbf{y}}{2\sigma^2}} \tag{9}$$

where d is the dimension of the definition space of the kernel, $\mathbf{I}$ is the identity matrix, and $\sigma^2\mathbf{I}$ is the covariance matrix.

We approximate the probability density $p(\mathbf{y})$ replacing each data sample $\mathbf{y}_i$ with a Gaussian kernel, and averaging the obtained values:

$$p(\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{y}-\mathbf{y}_i, \sigma^2\mathbf{I}).$$

We denote by M_p the number of training samples from class c_p. We have:

$$\int_{\mathbf{y}} p^2(\mathbf{y}, c_p)d\mathbf{y} = \frac{1}{N^2}\sum_{k=1}^{M_p}\sum_{l=1}^{M_p} G(\mathbf{y}_{pk}-\mathbf{y}_{pl}, 2\sigma^2\mathbf{I})$$

and

$$\int_{\mathbf{y}} p^2(\mathbf{y})d\mathbf{y} = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} G(\mathbf{y}_k-\mathbf{y}_l, 2\sigma^2\mathbf{I}),$$

where $\mathbf{y}_{pk}$, $\mathbf{y}_{pl}$ are two training samples from class c_p, whereas $\mathbf{y}_k$, $\mathbf{y}_l$ represent two training samples from any class.

Equation (8) can be rewritten, and we obtain the final ESRNG update formula of the feature weights:

$$\lambda^{(t+1)} = \lambda^{(t)} - \alpha\frac{1}{4\sigma^2}G(\mathbf{y}_1-\mathbf{y}_2, 2\sigma^2\mathbf{I})\cdot(\mathbf{y}_2-\mathbf{y}_1)\mathbf{I}\cdot \cdot(\mathbf{x}_1-\mathbf{w}_{j(1)}-\mathbf{x}_2+\mathbf{w}_{j(2)}),$$

where $\mathbf{w}_{j(1)}$ and $\mathbf{w}_{j(2)}$ are the closest prototypes to $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively.

The ESRNG algorithm has the following general steps:

1. Update the reference vectors using the SRNG scheme.
2. Update the feature weights.
3. Repeat Steps 1 and 2, for all training set samples.

This algorithm uses a generalized Euclidean distance. The updating formula for the reference vectors can be found in [1]; we will not explicitly use this formula in the present paper.

The ESRNG algorithm generates numeric values assigned to each input feature, quantifying their importance in the classification task: the most relevant feature receives the highest numeric value. We use these factors as feature weights in the FAMRFW algorithm.

4 FAMRFW – a novel neural model

The FAMRFW is a FAMR architecture with a generalized distance measure. For an ART_a category $\mathbf{w}_j$, we define its size $s(\mathbf{w}_j)$:

$$s(\mathbf{w}_j) = n - |\mathbf{w}_j| \tag{10}$$

and the distance to a normalized input $\mathbf{A}$:

$$dis(\mathbf{A}, \mathbf{w}_j) = |\mathbf{w}_j| - |\mathbf{A}\wedge\mathbf{w}_j| = \sum_{i=1}^{n} d_{ji}, \tag{11}$$

where $(d_{j1}, \ldots, d_{jn}) = \mathbf{w}_j - \mathbf{A} \wedge \mathbf{w}_j$. In [10] it is shown that:

$$T_j(\mathbf{A}) = \frac{n - s(\mathbf{w}_j) - dis(\mathbf{A}, \mathbf{w}_j)}{n - s(\mathbf{w}_j) + \alpha_a} \tag{12}$$

$$\rho(\mathbf{A}, \mathbf{w}_J^a) = \frac{n - s(\mathbf{w}_j) - dis(\mathbf{A}, \mathbf{w}_j)}{n} \tag{13}$$

A generalization of $dis(\mathbf{A}, \mathbf{w}_j)$ is the weighted distance:

$$dis(\mathbf{A}, \mathbf{w}_j; \lambda) = \sum_{i=1}^{n} \lambda_i d_{ji}, \tag{14}$$

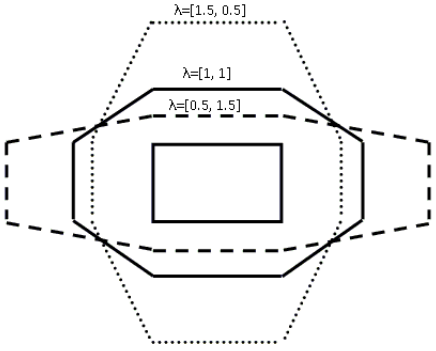
where $\lambda = (\lambda_1, \ldots, \lambda_n)$, and $\lambda_i \in [0, n]$ is the weight associated to the ith feature. We impose the constraint $|\lambda| = n$. For $\lambda_1 = \cdots = \lambda_n = 1$, we obtain in particular the FAMR.

Charalampidis *et al.* [10] used the following weighted distance:

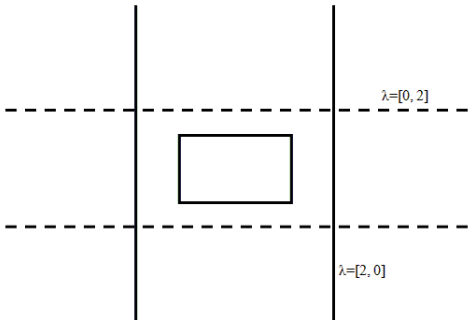
$$dis(\mathbf{x}, \mathbf{w}_j | \lambda, ref) = \sum_{i=1}^{n} \frac{(1-\lambda) l_j^{ref} + \lambda}{(1-\lambda) l_{ji} + \lambda} d_{ji}, \tag{15}$$

where l_j^{ref} is a function of category j's lengths of the hyper-rectangle, and λ is a scalar in $[0, 1]$. In our case, the function $dis(\mathbf{A}, \mathbf{w}_j; \lambda)$ does not depend on sides of the category created during learning, but on the computed feature weights. This makes our approach very different than the one in [10].

The effect of using distance $dis(\mathbf{A}, \mathbf{w}_j; \lambda)$ for a bidimensional category is depicted in Fig. 2(a). The hexagonal shapes represent the points situated at constant distance from the category. These shapes are flattened in the direction of the feature with a larger weight and elongated in the direction of the feature with a smaller weight. This is in accordance with the following intuition: The category dimension in the direction of a relevant feature should be smaller than the category dimension in the direction of a non-relevant feature. Hence, we may expect that more categories will cover the relevant directions than the non-relevant ones.



(a) Bounds for constant weighted distance $dis(\mathbf{A}, \mathbf{w}_j; \lambda)$ for various values of λ. The rectangle in the middle represents a category.



(b) Bounds for constant distance $dis(\mathbf{A}, \mathbf{w}_j; \lambda)$ for the null feature weight. The rectangle in the middle represents the category.

Figure 2: Geometric interpretation of constant distance when using $dis(\mathbf{A}, \mathbf{w}_j; \lambda)$ for bidimensional patterns.

For a null weight feature (Fig. 2(b)), the bounds are reduced to parallel lines on both sides of the rectangle representing the category. In this extreme case, the discriminative distance is the one along the remaining feature dimension. This is another major difference between our approach and the one in [10], where, while using function $dis(\mathbf{x}, \mathbf{w}_j | \lambda, ref)$, the contours of a constant weighted distance are inside

some limiting hexagons. In our method, the contour is insensitive to the actual value of the null weighted feature.

5 Experimental results

We test the FAMRFW for several standard classification tasks, all from the UCI Machine Learning Repository [5]. The experiments are performed on the FAMR and the FAMRFW architectures. The two FAMRFW stages are: *i)* the λ feature weights are obtained by the ESRNG algorithm; *ii)* these weights are used both for training and testing the FAMR.

A nice feature of the FAM architectures and the ESRNG algorithm is the on-line (incremental) learning capability, *i.e.*, the training set is processed only once. This type of learning is especially useful when dealing with very large datasets, since it can reduce significantly the computational overhead. For FAMR training and for both FAMRFW stages we use on-line learning.

5.1 Methodology

For each experiment, we use three-way data splits (*i.e.*, the available dataset is divided into training, validation, and test sets) and random subsampling. Random subsampling is a faster, simplified version of k-fold cross validation:

1. The dataset is randomized.

2. The first 60% of the dataset is used for training and the next 20% for validation (*i.e.*, for tuning the model parameters). The following parameters are optimized using a simple "grid-search" for $\rho_a, \rho_{ab} \in \{0, 0.1, \ldots, 0.9\}$ and $\beta_a \in \{0, 0.1, \ldots, 1\}$. The goal is to allow both fast learning and fast-commit slow-recode. The optimal parameter values are the ones producing the highest PCC and the lowest number of ART_a categories.

3. The network with optimal parameters is trained with the joint training + validation data.

4. The last 20% of the dataset is used for testing. As a result, the percent of correct classification (PCC) and the number of generated ART_a categories are computed.

5. Repeat this procedure six times.

The ρ_a value, optimized during training/validation, controls the number of generated ART_a categories. After training/validation, this number does not change. For $\rho_a > 0$, some test vectors may be rejected (*i.e.*, not classified).

In all our experiments, after the ART_a categories were generated, we set $\rho_a = 0$ for testing. This has the following positive effects:

- All test vectors are necessarily classified.

- We obtain experimentally better classification results, both for the FAMR and the FAMRFW, compared to the ones with optimized ρ_a values. This is shown in Table 1, for all considered classification tasks. The feature weights values in the FAMRFW are the ones mentioned in the following sections.

Table 1: Average PCC test set results using the optimized ρ_a (computed in the validation phase) vs. using $\rho_a = 0$.

	FAMR		FAMRFW	
	optimized ρ_a	$\rho_a = 0$	optimized ρ_a	$\rho_a = 0$
Breast cancer	86.54%	91.22%	91.22%	91.22%
Balance scale	75.86%	76.53%	75.92%	78.13%
Wine recognition	83.33%	84.72%	83.79%	89.35%
Ionosphere	85.44%	88.96%	85.91%	89.43%

5.2 Breast cancer classification

This dataset (formally called Wisconsin Diagnostic Breast Cancer) includes 569 instances. The instances are described by 30 real attributes. The given features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

The FAMRFW generated weights are: $[0.784, 0.816, 0.795, 2.847, 0.784, 0.784, 0.784, 0.784, 0.784, 0.784, 0.784, 0.784, 0.784, 0.785, 0.808, 0.784, 0.784, 0.784, 0.784, 0.784, 0.784, 0.784, 0.829, 0.828, 5.047, 0.784, 0.784, 0.784, 0.784, 0.784, 0.784]$. In Table 2, we observe that the average PCC for the FAMR and the FAMRFW is the same, but the FAMRFW has much less ART_a categories than the FAMR.

Table 2: Classification performance for the Breast Cancer Problem.

Test no.	FAMR		FAMRFW	
	No. of ART_a categories	PCC	No. of ART_a categories	PCC
1	61	93.85%	24	87.71%
2	7	90.35%	7	93.85%
3	10	95.61%	8	91.22%
4	39	85.08%	6	88.59%
5	6	92.98%	6	94.73%
6	6	89.47%	5	91.22%
Average	21.5	91.22%	9.33	91.22%

5.3 Balance scale classification

This dataset was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance. The correct way to find the class is the greater of (left-distance * left-weight) and (right-distance * right-weight). If they are equal, it is balanced. The set contains 625 patterns, with a uneven distribution of the three classes; each input pattern has 4 features.

The ESRNG generated feature weights are $\lambda = [1.002, 1.113, 0.827, 1.058]$. The FAMRFW has better classification accuracy and less ART_a categories than the FAMR (Table 3).

5.4 Wine recognition

The Wine recognition data are the results of a chemical analysis of wines grown in the same region in Italy, but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the 3 types of wines. The dataset contains 178 instances.

Table 3: Classification performance for the Balance Scale Problem.

Test no.	FAMR		FAMRFW	
	No. of ART_a categories	PCC	No. of ART_a categories	PCC
1	95	74.4%	53	75.2%
2	70	80.0%	39	80.0%
3	22	78.4%	54	81.6%
4	75	75.2%	44	85.6%
5	125	71.2%	69	72.0%
6	62	80.0%	107	74.4%
Average	74.83	76.53%	61	78.13%

The ESRNG algorithm produced the weights $\lambda = [0.900, 0.757, 0.659, 1.668, 2.349, 0.702, 1.028, 0.668, 0.774, 0.874, 0.666, 0.701, 1.253]$. The FAMRFW classification results are better, with less generated ART_a categories (Table 4).

Table 4: Classification performance for the Wine Recognition Problem.

Test no.	FAMR		FAMRFW	
	No. of ART_a categories	PCC	No. of ART_a categories	PCC
1	10	88.88%	6	86.11%
2	15	97.22%	10	97.22%
3	32	69.44%	11	86.11%
4	17	83.33%	11	86.11%
5	55	80.55%	39	94.44%
6	12	88.88%	8	86.11%
Average	23.5	84.71%	14.16	89.35%

5.5 Ionosphere

This binary classification problem starts from collected radar datasets. The data come from 16 high-frequency antennas, targeting the free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those passing through the ionosphere. There are 351 instances and each input pattern has 34 features.

The ESRNG generated λ vector is: [0.551,0.520,1.179,1.168,1.301,1.180,0.940,1.272,1.024,0.903, 0.843,0.976,0.870,0.844,0.807,0.877,0.893,1.012,0.994,1.012,0.964,1.061,1.029,1.227,0.978,1.020,0.943, 1.027,1.087,1.032,0.978,1.117,0.999,1.374]. On average, FAMRFW produced much less ART_a categories than the FAMR. This time, the FAMR produced a slightly better PCC (Table 5).

6 Conclusions

According to our experiments, using the feature relevances and the generalized distance measure may improve the classification accuracy of the FAMR algorithm. In addition, the FAMRFW uses less ART_a categories, which is an important factor. The number of categories controls the generalization

Table 5: Classification performance for the Ionosphere Problem.

Test no.	FAMR		FAMRFW	
	No. of ART_a categories	PCC	No. of ART_a categories	PCC
1	28	81.69%	8	90.14%
2	20	81.69%	8	85.91%
3	17	91.54%	7	83.09%
4	9	94.36%	8	88.73%
5	5	90.14%	5	94.36%
6	9	94.36%	5	94.36%
Average	14.66	88.96%	6.83	89.43%

capability and the computational complexity of a FAM architecture. This generalization is a trade-off between overfitting and underfitting the training data. It is good to minimize the number of categories if this does not decrease too much the classification accuracy.

The ESRNG feature weighting algorithm can be replaced by other weighting methods. We have not tested the function approximation capability of the FAMRFW neural network because the ESRNG weighting algorithm is presently restricted to classification tasks. LVQ methods can be extended to function approximation [23] and we plan to adapt the ESRNG algorithm in this sense. This would enable us to test the FAMRFW + ESRNG procedure on standard feature approximation and prediction benchmarks.

Our approach is at the intersection of two major computational paradigms:

1. Carpenter and Grossberg's adaptive resonance theory, an advanced distributed model where parallelism is intrinsic to the problem, not just a mean to speed up [6].

2. Onicescu's informational energy and the unilateral dependency measure. To the best of our knowledge, we are the only ones using Onicescu's energy in neural processing systems.

Bibliography

[1] R. Andonie and A. Caţaron. Feature ranking using supervised neural gas and informational energy. In *Proceedings of IEEE International Joint Conference on Neural Networks (IJCNN2005)*, Canada, Montreal, July 31 - August 4, 2005.

[2] R. Andonie, A. Caţaron, and L. Sasu. Fuzzy ARTMAP with feature weighting. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, Innsbruck, Austria, Febr. 11-13, 2008, 91–96.

[3] R. Andonie and F. Petrescu. Interacting systems and informational energy. *Foundation of Control Engineering*, 11, 1986, 53–59.

[4] R. Andonie and L. Sasu. Fuzzy ARTMAP with input relevances. *IEEE Transactions on Neural Networks*, 17, 2006, 929–941.

[5] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007. University of California, Irvine, School of Information and Computer Sciences http://www.ics.uci.edu/~mlearn/MLRepository.html

[6] I. Dziţac and B. E. Bărbat. Artificial intelligence + distributed systems = agents. *International Journal Computers, Communications, and Control*, 4, 2009, 17–26.

[7] G. A. Carpenter, S. Grossberg, N. Markuzon, J. H. Reynolds, and D. B. Rosen. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Transactions on Neural Networks*, 3, 1992, 698–713.

[8] G. A. Carpenter, B. L. Milenova, and B. W. Noeske. Distributed ARTMAP: A neural network for fast distributed supervised learning. *Neural Networks*, 11, 1998, 793–813.

[9] G. A. Carpenter and W. Ross. ART-EMAP: A neural network architecture for learning and prediction by evidence accumulation. *IEEE Transactions on Neural Networks*, 6, 1995, 805–818.

[10] D. Charalampidis, G. Anagnostopoulos, M. Georgiopoulos, and T. Kasparis. Fuzzy ART and Fuzzy ARTMAP with adaptively weighted distances. In *Proceedings of the SPIE, Applications and Science of Computational Intelligence*, Aerosense, 2002.

[11] I. Dagher, M. Georgiopoulos, G. L. Heileman, and G. Bebis. An ordering algorithm for pattern presentation in Fuzzy ARTMAP that tends to improve generalization performance. *IEEE Transactions on Neural Networks*, 10, 1999, 768–778.

[12] I. Dagher, M. Georgiopoulos, G. L. Heileman, and G. Bebis. Fuzzy ARTVar: An improved fuzzy ARTMAP algorithm. In *Proceedings IEEE World Congress Computational Intelligence WCCI'98*, Anchorage, 1998, 1688–1693.

[13] J. C. Principe *et al*. Information-theoretic learning. In S. Haykin, editor, *In Unsupervised Adaptive Filtering*. Wiley, New York, 2000.

[14] E. Gomez-Sanchez, Y. A. Dimitriadis, J. M. Cano-Izquierdo, and J. Lopez-Coronado. μARTMAP: Use of mutual information for category reduction in fuzzy ARTMAP. *IEEE Transactions on Neural Networks*, 13, 2002, 58–69.

[15] S. Guiaşu. Information theory with applications. McGraw Hill, New York, 1977.

[16] B. Hammer, D. Schunk, T. Bojer, and T. K. von Toschanowitz. Relevance determination in learning vector quantization. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN 2001)*, Bruges, Belgium, 2001, 271–276.

[17] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21, 2005, 21–44.

[18] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15, 2002, 1059–1068.

[19] C. P. Lim and R. Harrison. ART-Based Autonomous Learning Systems: Part I - Architectures and Algorithms. In L. C. Jain, B. Lazzerini, and U. Halici, editors, *Innovations in ART Neural Networks*. Springer, 2000.

[20] C. P. Lim and R. F. Harrison. An incremental adaptive network for on-line supervised learning and probability estimation. *Neural Networks*, 10, 1997, 925–939.

[21] S. Marriott and R. F. Harrison. A modified fuzzy ARTMAP architecture for the approximation of noisy mappings. *Neural Networks*, 8, 1995, 619–641.

[22] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. Neural-gas network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4, 1993, 558–569.

[23] S. Min-Kyu, J. Murata, and K. Hirasawa. Function approximation using LVQ and fuzzy sets. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Tucson, AZ, 2001, 1442–1447.

[24] O. Onicescu. Theorie de l'information. Energie informationnelle. *C. R. Acad. Sci. Paris, Ser. A–B*, 263, 1966, 841—842.

[25] O. Parsons and G. A. Carpenter. ARTMAP neural networks for information fusion and data mining: map production and target recognition methodologies. *Neural Networks*, 16, 2003, 1075–1089.

[26] M. Taghi, V. Baghmisheh, and P. Nikola. A Fast Simplified Fuzzy ARTMAP Network. *Neural Processing Letters*, 17, 2003, 273–316.

[27] S. J. Verzi, G. L. Heileman, M. Georgiopoulos, and M. J. Healy. Boosted ARTMAP. In *Proceedings IEEE World Congress Computational Intelligence WCCI'98*, 1998, 396–400.

[28] J. Williamson. Gaussian ARTMAP: A neural network for fast incremental learning of noisy multi-dimensional maps. *Neural Networks*, 9, 1996, 881–897.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 118-126

Integrated System for Stereoscopic Cognitive Vision, Localization, Mapping, and Communication with a Mobile Service Robot

Cătălin Buiu

POLITEHNICA University of Bucharest
Department of Automatic Control and Systems Engineering
Spl. Independentei 313, 060042 Bucharest, Romania
E-mail: cbuiu@ics.pub.ro

Abstract: This paper describes a stereo-vision-based mobile robot that can navigate and explore its environment autonomously and safely and simultaneously building a tridimensional virtual map of the environment. The control strategy is rule-based and the interaction with robot is done via Bluetooth. The stereoscopic vision allows the robot to recognize objects and to determine the distance to the analyzed objects. The robot is able to generate and simultaneously update a full colour 3D map of the environment that is being explored. The position and type of each detected and recognized object is marked in this 3D map. Furthermore, the robot will be able to use a gripper in order to collect detected objects and carry them to dedicated collecting bins, and so will be able to work in commercial waste cleanup applications. This application represents a successful integration of computers, control and communication techniques in mobile service robot control.
Keywords: control, communication, localization, mapping, mobile robot, stereoscopic vision, virtual reality

1 Introduction

More than 7 million robots will be sold from 2005 to 2008 according to estimations of the International Robotics Federation and of the Economic Commission for Europe of the United Nations. Until 2010 a robust increase of 4% per year in the number of robots is estimated. Many of these are service robots which are used to assist or even to replace humans in tedious, dull, dangerous or repetitive tasks. The same sources estimate that by 2010, service robots will be able to fully assist elder people and people with disabilities, will extinct fires, will explore industrial pipes and more [1].

In ecological applications, service robots are used to collect waste and dangerous items in indoor and outdoor environments. For doing that, the robots must be able not only to perceive and act upon the environment by using a wide range of sensors and actuators, but also to manifest human-like cognitive abilities, such as to localize themselves, to recognize and classify objects, to generate maps of the environment, to learn from experience, to interact in a natural way with humans and other robots or to develop physical and cognitive abilities in a kind of developmental process similar to humans. It is often the case that a team of robots is asked to fulfill such a task. There are already a lot of interesting results obtained in collective robotics, see for example [2] where coordinated control based on artificial vision is investigated and [3] where decentralized formation control of mobile robots with limited sensing is addressed.

The problem of Simultaneous Localization and Mapping (SLAM) consists of estimating concurrently the robot's position and generating a map of its surrounding environment. This is an essential skill for a mobile robot but to this day it has eluded complete and robust solutions because noisy robot dynamics and sensors make solving SLAM a difficult task.

SLAM has been widely used for navigation and typically makes use of laser range-finders or sonars. An advantage of using stereo vision over laser range-finders is the ability to detect obstacles at different heights. The solution presented in [4] is based on learning maps of 3D point-landmarks whose location is estimated using correlation-based stereo and identification is performed using their appearance in images using the Scale Invariant Feature Transform (SIFT) [5]. The authors derive an estimate of the robot's motion from sparse visual measurements using stereo vision and multiple view geometry techniques [6] known in robotics as visual odometry [7], [8], [9].

A number of professional stereo vision systems and related software systems have been developed and have found a number of interesting applications in various domains, from the control of industrial manipulators for assembly and pick-and-place operations, material handling, collision warning and obstacle detection in robotics, people-tracking, environment modeling, autonomous guidance of corn harvesters, digitizing books, to medical applications, such as ophthalmic diagnostics, IR mammography, and robotic laparoscopy.

Stereo vision for navigation has a long history and is frequently exploited for autonomous navigation, but has limitations in terms of its density and accuracy in the far field [10]. If landmarks can be placed in the field of view of the camera, the location of a vehicle can be determined by means of stereo vision [11], and if a solid model of the target object is available, a robotic manipulator will have at its disposal a modeled environment for automatic tasks [12]. In [13] it is presented a stereoscopic vision system for a Khepera miniature robot. The vision system performs objects detection by using the stereo disparity and stereo correspondence.

An adaptive panoramic stereo vision approach for localizing 3D moving objects has been developed in the Department of Computer Science at the University of Massachusetts at Amherst. In the adaptive stereo model, the sensor geometry can be controlled to manage the precision of the resulting virtual stereo system [14]. Other indoor and outdoor stereo vision systems have been developed and tested with satisfactory results and some drawbacks, see [15], [16] and [17].

A novel optical system allows the capture of a pair of short, wide stereo images from a single camera, which are then processed to detect vertical edges and infer obstacle positions and locations within the planar field of view, providing real-time obstacle detection [18].

Very few applications concern the problem of waste collecting service robots acting indoor. The application reported in this paper is part of the bigger ReMaster research project currently under development at the Autonomous Robotics Lab of the POLITEHNICA University of Bucharest, Romania. This project concerns the development of a commercial cognitive service robot to be used in waste cleanup in office buildings. A first prototype (ReMaster One) has been built. Related details on the structure of the prototype and of its cognitive vision system using a monocular vision system are given in [19] and [20]. The acquired expertise has been used to propose the structure and to design a stereoscopic cognitive vision system which is detailed in [21].

The aim of this paper is to present the current phase in the ReMaster project which consists in the design and implementation of an integrated system for stereoscopic cognitive vision, localization, mapping, and communication with the robot. The goals of this system is to allow the robot to recognize and classify various objects and to determine the distance to the objects. Combined with the self-localization ability of the robot, this allows the absolute position of detected objects to be determined and marked on a tridimensional map of the working space that is continously updated. So, stereo vision and SLAM are integrating in order to create a map of the environment, without using any landmarks. The realization of this vision based mapping is the main contribution of this paper.

The paper is structured as follows. Section 2 gives an overview of the system architecture, and of the robot control and communication system, while Section 3 presents the realization of the stereoscopic cognitive vision system. Section 4 describes the way in which the robot is able to generate, maintain and update a tridimensional virtual map of the environment that is being explored. The last section of the paper presents conclusions and some directions for further research and developments.

2 Robot Control and Communication System

The integrated system was implemented on a Koala robot (Fig. 1) which is a mid-size robot designed for real-world applications and capable of carrying larger accessories. It has been chosen for this application, as Koala has the functionality necessary for use in practical applications (like sophisticated battery management), and rides on 6 wheels for indoor and all-terrain operation. It has 16 distance sensors and can be controlled via Bluetooth.

Figure 1: Koala mobile robot (www.k-team.com)

Two commercial webcameras have been mounted on top Koala at the same level (Fig. 2), at a distance of 95 mm between them, and at a height of 170 mm. The cameras are inclined at 10 degrees, and have CMOS 1.3 megapixel sensors (1280*960 pixels images and 640*480 videos), manual focus, and a focal distance of 1/4.8 mm.

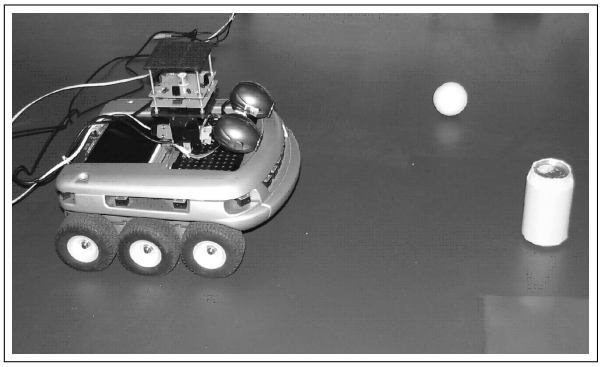


Figure 2: Stereoscopic vision system on-board Koala

On the robot side, there is a Bluetooth 333s module installed and directly connected to the serial interface of the robot. The control program runs on a separate laptop which communicates with the robot by using a Bluetooth connection. The robot can localize itself by using a dedicated and redundant system consisting of a beacon and two transponders ([19], [21]) and an odometry algorithm. The robot is able to navigate in indoor environments consisting of walls and various objects, such as empty cans and bottles.

The control program is implemented in Matlab and is based on simple control rules which allow an obstacle avoidance and waste finding behavior (Fig. 3). The robot will move forward and will be able of a safe navigation and detection of objects in the workspace (Fig. 4). After detecting an obstacle, the robot will stop and action according to the type of the obstacle. If wall, the obstacle will be avoided and the robot will resume moving. If not wall, using the distance sensors on-board the robot, the system will compute the distance to the object and the corresponding angle. Using these two measurements, the system will compute the absolute position of the detected object and will compare this with the stored coordinates of previously detected objects. If the object is new (its absolute position is not in the database), it will get more attention from the robot which will turn so that it is facing the object (Fig. 5).

Now, the system is ready to acquire stereoscopic images of the object. The images are processed and analyzed as explained in the next section. The distance to the object is determined and based on

```
repeat
    move forward
    detect objects
    stop
        if wall then avoid
        else
                determine the object's position using the distance sensors
                if the object is new in the space (not in the database) then
                        turn to the object
                        take stereoscopic images of the object
                        image processing and analysis
                        determine distance to the object using disparity map
                        identify object
                        determine absolute coordinates of the object
                        store object in database
                        mark object on virtual map
                resume moving and avoid object
    continue mission
until mission accomplished
```

Figure 3: Control algorithm

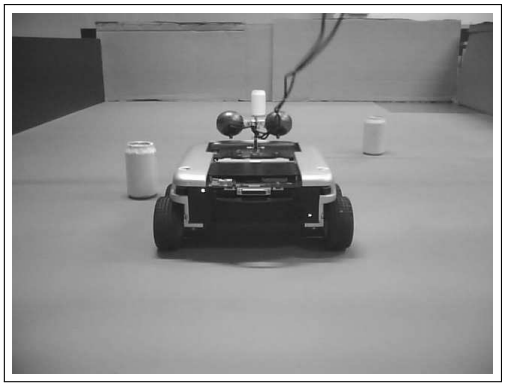


Figure 4: Robot navigating in a test environment

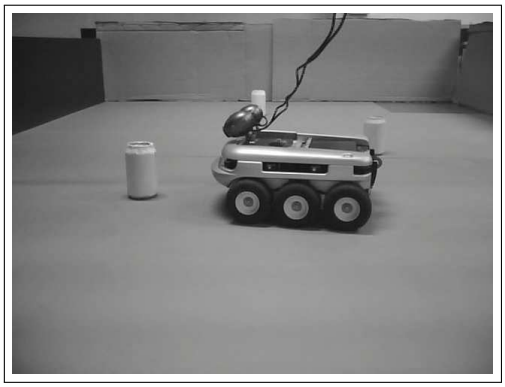


Figure 5: Robot turning to a detected object

the absolute position of the robot, the absolute position of the object is also determined. Then the robot resumes its movement in the workspace after storing the type and coordinates of the object in the database and after having marked its position in the tridimensional map which will be detailed later.

3 Stereoscopic Cognitive Vision system

Stereoscopy is a technique for infering the 3D position of objects from two or more simultaneously views of the scene. Reconstruction of the world seen through stereo cameras can be divided in two steps. First, the correspondence problem means that for every point in one image to find out the correspondent point on the other and compute the disparity of these points. This disparity correlates to distance, and the higher disparity of object pixel means that the object is closer to the cameras. Secondly, there is the triangulation step. Given the disparity map, the focal distance of the two cameras and the geometry of the stereo setting (relative position and orientation of the cameras) compute the (X,Y,Z) coordinates of all points in the images. The system presented in this paper solves both steps as will be described below.

Key advantages of camera based systems include: they offer minimally complex solutions, have very low costs, they are entirely solid state, and colour information can be easily acquired at the same time as range data, helping to build realistic full colour 3D models of the environment. All these advantages are exploited in the application reported in this paper.

Stereo vision provides realtime, full-field distance information, and is useful in many applications in a wide variety of fields, including robotics. There is a number of dedicated software packages, such as Small Vision System for realtime stereo analysis from SRI's Artificial Intelligence Center. Sentience is a volumetric perception system for mobile robots and uses webcam-based stereoscopic vision to generate depth maps, and from these create colour 3D voxel models of the environment for obstacle avoidance, navigation and object recognition purposes.

A "cognitive vision system" is defined in [22] as a system that uses visual information to achieve: recognition and categorization of objects, structures and events, learning and adaptation, memory and representation of knowledge, control and attention. For example, a cognitive monocular vision system for a mobile robot using a CMUcam2+ camera is presented in [20].

The visual system's architecture is presented in Fig. 6. All the visual information processing is done on the same separate laptop. Given the disparity map, the focal distance of the two cameras and the geometry of the stereo setting (relative position and orientation of the cameras), the system is able to compute the coordinates of all points in the images. The distance to the object is used to determine the absolute position of the object which is marked on the 3D map.

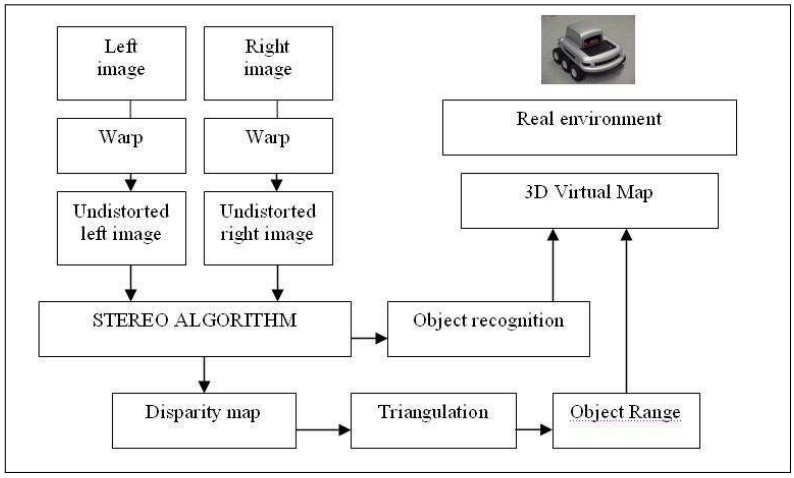


Figure 6: Stereovision system's architecture

Screenshots from our application that present two stereoscopic images of a detected object are given in Fig. 7. These images will be further processed.

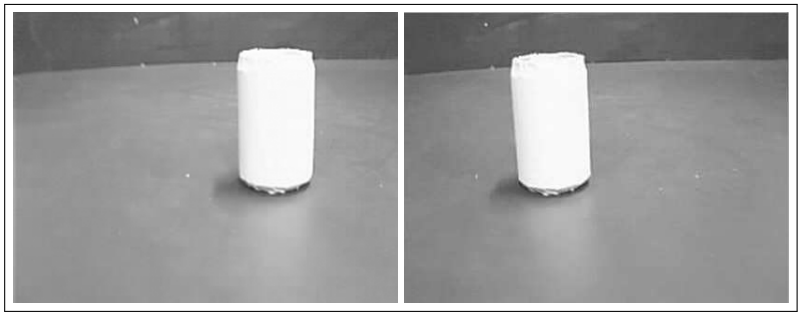


Figure 7: Stereoscopic images of a detected object (left and right hand camera)

Now the images contain relevant data that will be brought in such a form that contours can be extracted. The images are binarized by extracting colour channels corresponding to the colour of detected objects (yellow, in our case). Then, dilatation and errosion filters are applied to the images (see Fig. 8). Then the images are segmented and objects detected (Fig. 9).

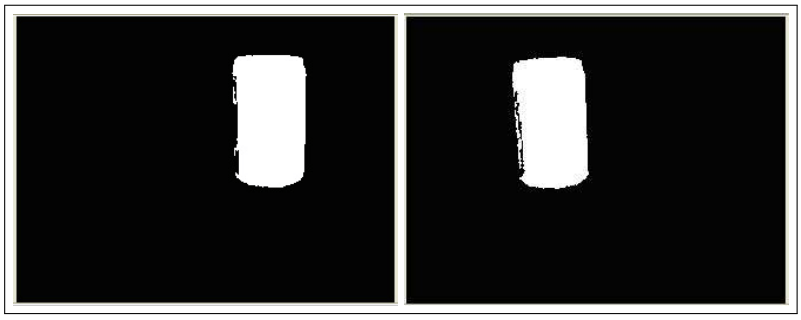


Figure 8: Extraction of yellow colour channel and application of dilatation and errosion filters (left and right hand image)

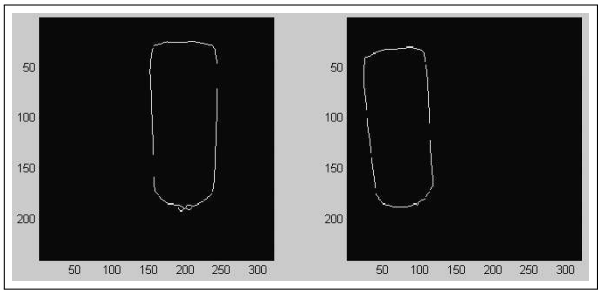


Figure 9: Detected contour for the object in the image

Using simple scalar descriptors, such as area and perimeter, the detected object is recognized and classified as a can, in our case.

4 Generation and Update of a Tridimensional Map of the Environment

Simultaneous Localization and Mapping (SLAM) is an essential capability for mobile robots exploring unknown environments. The robot presented in this paper is using a dedicated self-localization

system based on the use of a Beacon unit and two Transponders (master and slave) [21]. The two transponder units are fixed, while the beacon unit is installed on the robot. Half-duplex bidirectional communication between beacon and transponders is realized by using infrared light and ultrasounds. The system is using a ATmega8 microcontroller with 16MIPS at 16MHz. The localization of the robot is realized by triangulation of the distances to the two transponders. More, odometry algorithms contribute to a more precise localization of the robot in the working space.

The system is able to generate a a virtual map of the explored environment, in which the space, the robot and the objects are modeled as VRML (Virtual Reality Modeling Language) objects. VRML is a standard file format for representing tridimensional interactive vector graphics. It also enables the integration of interactive 3D graphics into the Web.

By using the Virtual Reality Toolbox from Matlab, the system will generate realistic 3D views of the working space and the robot (Fig. 10), and objects (Fig. 11, in which a question mark means an unknown object).

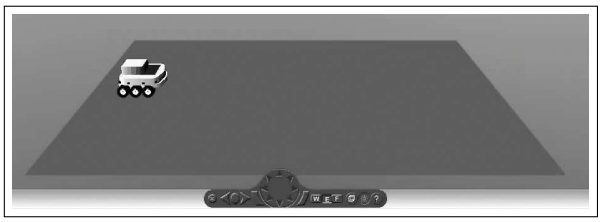


Figure 10: VRML models of the working space and robot

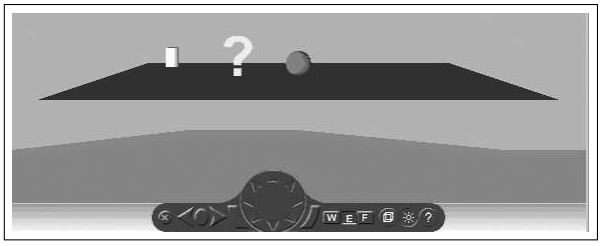


Figure 11: VRML models of objects

The robot will explore the working space according to the control strategy presented above and simultaneously will update the tridimensional map. After an object is detected and the robot turns towards it, both cameras are taking images of the scene. Further, the images are analyzed and the object recognized. The absolute position of the object is also determined and the object marked on the map (Fig. 12). Then the robot resumes its movement in the workspace and associated activities: navigation, search, classification and localization of objects.

The test results show a good and robust functioning of the stereovision system, and although the processing times are not low, this can be improved by the use of an embedded PC with more computing power.

5 Summary and Conclusions

The main research thrust of this paper has been to demonstrate that an integrated system for communication, control, localization and mapping using stereoscopic vision and 3D maps can be designed and implemented for a mobile service robot which will collect waste in indoor environments. This integrated system will be transferred to a more powerful version of the first prototype (ReMaster One) of the commercial waste cleanup robot that is the final aim of the ReMaster project. The new robot will have

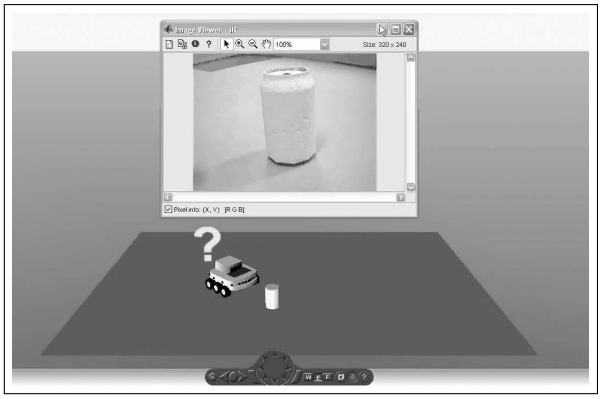


Figure 12: Robot taking pictures of an object, recognizing it as a can, and marking it on the map

a gripper such that the detected objects can be grasped and carried to dedicated bins. Future efforts will address the design and implementation of new navigation strategies based on fuzzy logic. Image processing algorithms based on cellular neural networks are currently under investigation and implementation. More research will be done in what regards the interactive aspects of the robotic system so that the robot will be able to interact with humans and other robots in a natural way.

Acknowledgements

We acknowledge the support of the Romanian Government through the Excellence Research program (Contract 83-CEEX-II-03/31.07.2006) and the work of Cristian Ionita and Laura Antochi to the development of the stereoscopic vision system and virtual map.

Bibliography

[1] C. Buiu (Editor), *Cognitive Robots (in Romanian)*, Editura Universitara, 2008.

[2] C. M. Soria, R. Carelli, R., J. M. Ibarra Zannatha, Coordinated Control Of Mobile Robots Based On Artifcial Vision, International Journal of Computers, Communications, and Control, Vol. I (2006), No. 2, pp. 85-94.

[3] K.D. Do, Bounded Controllers for Decentralized Formation Control of Mobile Robots with Limited Sensing, International Journal of Computers, Communications, and Control, Vol. II (2007), No. 4, pp. 340-354.

[4] P. Elinas, R. Sim, J. J. Little, SigmaSLAM: Stereo vision SLAM using the Rao-Blackwellised particle filter and a novel mixture proposal distribution, In Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), Florida, USA, 2006.

[5] D. G. Lowe, Object recognition from local scale-invariant features, In Int. Conf. on Computer Vision, Corfu, Greece, September 1999, pp. 1150-1157.

[6] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge, UK: Cambridge Univ. Pr., 2000.

[7] D. Nister, O. Naroditsky, J. Bergen, Visual odometry, In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), 2004, pp. 652-659.

[8] M. Agrawal, K. Konolige, Rough terrain visual odometry, In Proceedings of the International Conference on Advanced Robotics (ICAR), August 2007.

[9] K. Konolige, M. Agrawal, Frame-frame matching for realtime consistent visual mapping, In Proceedings of IEEE International Conference on Robotics and Automation (ICRA), April 2007.

[10] M. J. Procopio, T. Strohmann, A. R. Bates, G. Grudic, Jane Mulligan, Using Binary Classifiers to Augment Stereo Vision for Enhanced Autonomous Robot Navigation, University of Colorado at Boulder Technical Report CU-CS-1027-07, April 2007

[11] Wang, L. K., S. Hsieh, E. C. Hsueh, F. Hsaio, K. Huang, Complete pose determination for low altitude unmanned aerial vehicle using stereo vision, In Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pp. 108 - 113.

[12] Lee, S., D. Jang, E. Kim, S. Hong, J. Han, A real-time 3D workspace modeling with stereo camera, In Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pp. 2140 - 2147.

[13] T. Chinapirom, U. Witkowski, R. Ulrich, Stereoscopic Camera for Autonomous Mini-Robots Applied in KheperaSot League, Research Report, University of Paderborn, Germania, 2007

[14] D. R. Karuppiah, Z. Zhu, P. Shenoy, E. M. Riseman, A fault-tolerant distributed vision system architecture for object tracking in a smart room, In B. Schiele and G. Sagerer (Eds.), Springer Lecture Notes in Computer Science 2095, pp 201-219, 2007

[15] S. Florczyk, *Robot Vision: Video-based Indoor Exploration with Autonomous and Mobile Robots*, Weinheim: Wiley-VCH, 2005.

[16] M. F. Ahmed, Development of a Stereo Vision system for Outdoor Mobile Robots, M.S. thesis, University of Florida, 2006.

[17] F. Rovira-Más, S. Han, J. Wei, J. F. Reid, Autonomous Guidance of a Corn Harvester using Stereo Vision, Agricultural Engineering International: the CIGR Ejournal, Manuscript ATOE 07 013, Vol. IX. July, 2007.

[18] W. Lovegrove, B. Brame, Single-camera stereo vision for obstacle detection in mobile robots, In Intelligent Robots and Computer Vision XXV: Algorithms, Techniques, and Active Vision., Proceedings of the SPIE, Volume 6764, pp. 67640T, 2007

[19] C. Buiu, F. Cazan, R. Ciurlea, Developing of a Service Robot to Recognize and Sort Waste, In: 16th International Conference on Control Systems and Computer Science, pp. 298-303. POLITEHNICA Press, Bucharest, 2007.

[20] Ana Pavel, C. Vasile, C. Buiu, Cognitive Vision System for an Ecological Mobile Robot, In Proceedings of SINTES 13, The International Symposium on System Theory, Automation, Robotics, Computers, Informatics, Electronics and Instrumentation, pp. 267-272, Universitaria Press, Craiova, 2007.

[21] C. Buiu, Design and development of a waste cleanup service robot, In Proceedings of the First International EUROBOT Conference, Heidelberg, pp. 194-202, 2008.

[22] A.G. Cohn, D. Magee, A. Galata, D. Hogg, S. Hazarika, Towards an architecture for cognitive vision using qualitative spatio-temporal representations and abduction, In C. Freksa, W. Brauer, C. Habel, K.F. Wender (editors), Spatial Cognition III, Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning, pp. 232-248, Springer-Verlag, 2003.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 127-136

Application of Genetic Algorithms for the DARPTW Problem

Claudio Cubillos, Enrique Urra, Nibaldo Rodríguez

Pontificia Universidad Católica de Valparaíso
Escuela de Ingeniería Informática
Av. Brasil 2241, Valparaíso, Chile
E-mail: claudio.cubillos@ucv.cl, enrique.urra.c@mail.ucv.cl

Abstract: On the Dial-a-Ride with time windows (DARPTW) customer transportation problem, there is a set of requests from customers to be transported from an origin place to a delivery place through a locations network, under several constraints like the time windows. The problem complexity (NP-Hard) forces the use of heuristics on its resolution. In this context, the application of Genetic Algorithms (GA) on DARPTW was not largely considered, with the exception of a few researches. In this work, under a restrictive scenario, a GA model for the problem was developed based on the adaptation of a generic GA model from literature. Our solution applies data pre-processing techniques to reduce the search space to points that are feasible regarding time windows constraints. Tests show competitive results on Cordeau & Laporte benchmark datasets while improving processing times.
Keywords: Dial-a-ride, Passenger Transportation, DARPTW, Heuristic, Scheduling.

1 Introduction

In the research of transport systems, the Dial-a-Ride Problem (DARP) or the customers' transportation problem is largely known [3]. It consists on searching the optimum way to transport a set of customers which are territorially distributed through a locations network, considering diverse constraints, for example the vehicle capacity and the time windows (TW) which are time intervals where a customer can be picked or delivered on the respective location, in a feasibility context.

The problem objective is to optimize the transport system factors (vehicles number, travel costs) and the quality of service for customers (waiting time, travel time). DARPTW (the time windows problem version) is considered a NP-Hard problem, especially because of the time-window constraints [10]. For this reason, the problem is usually solved through heuristics to find good solutions, under its diverse variants. In fact, the time windows restrictions make the problem highly non-convex, making it difficult to find feasible solutions.

One of the tools considered in this context are the Genetic Algorithms (GAs). After their comparison in the scientific scenario thanks to John Holland on the 70's decade [9], these algorithms have been successfully accepted by their efficiency to solve problems of diverse complexity, and to date there is a large number of proposed GA models that considers the canonical GA problems, particularly the linkage concept [8].

In this work, the application of GAs on DARPTW is extended considering two elements: on the one hand, the implementation of the LLGA model [8] with and adequate adaptation in the context and on the other hand, the use of data pre-processing techniques (namely precedence table of events and incompatible clients' list) for aiding the GA to avoid infeasible solutions from the time windows perspective.

The paper is structured as follows. Section 2 explains the DARPTW problem for then in section 3 tackling other research in the field. Section 4 details the implemented GA and section 5 the experiments and its results. The main conclusion of the work are drawn in Section 6.

2 The DARPTW Problem

DARPTW is a multiple objective optimization problem, because there are two critical factors to be optimized: on the one hand the total transportation costs and on the other, the quality of service offered to customers (minimizing their insatisfaction with the service). There is a set of transportation requests from customers that are known in advance and do not change during algorithm execution, defining the problem as static. Each request defines a time window for the customer delivery and a time window for the customer pickup. The upper bound (Latest Time) and the lower bound (Early Time) of the time window are supplied. A solution is considered infeasible when the vehicle arrives outside the time window bounds, defining these as hard time windows.

To execute the service, there is a homogeneous vehicles set with the same load capacity that cannot be exceeded. The passengers are picked and delivered by the same vehicle. A vehicle can enter on inactivity times or slacks only without passengers on board. Additionally, there is only one depot (single depot), a particular location where vehicles start and end their travels. In this work, maximum route duration is not considered.

3 Related Work

As mentioned previously, there are only few researches where the GAs are considered for DARPTW. There are more works which provide relevant elements in this context on VRP (Vehicle Routing Problem), a generic case of DARP. An example is Thangiah's work[11], which describes GIDEON, a GA based heuristic to solve VRP with time windows. This mechanism uses a cluster first - route second strategy, starting with the assignation of customers to vehicles and after improving the best solution by a post-optimizing process. For the system implementation, a GA software called GENESIS is used, where the individuals are represented by bit strings. The client clusters/sections are obtained from an individual by splitting him on K divisions of B bits. Each division is used to compute the size of a sector. The individual quality is obtained through the cost function evaluation when all computed clients are served, regarding their derived sector divisions. In this work, B = 3 is used, bigger values showed less satisfactory results. For the testing, the parameter values for the population size, crossover rate and mutation rate were 1000, 0.5 and 0.001 respectively. A set of 56 instances were tested using the Solomon's benchmark data, widely known in this context. In the results, 41 instances showed improvement in comparison to other heuristics developed by Solomon and Thompson.

Regarding DARPTW, in [1], a cluster first - route second strategy is also used. The mathematic model used in their work is a generalization of the one used in this work. This generalization is justified by the use of soft time windows, so the objective function considers additional elements related to quality of service. The individual is based on client clusters. A matrix is used, where the row number equals the available vehicles and the columns equals the clients and depots number. If an element on the matrix equals to 1, then the respective client is assigned to the respective vehicle. Additionally, a row (vehicle) represents a specific route. Because the absence of standardized benchmark data set in DARPTW, in contrast to VRP case, a set developed by Cordeau & Laporte is used [5], that contains 20 random instances. This set considers instances from 24 to 144 customers. The obtained results are compared also with a Cordeau & Laporte research [4], with similar improvement.

Finally, there is the Cubillos work [6], where the objective is to develop a specific GA model to solve the DARPTW problem, considering it as a deceptive problem, using the GA to solve the full problem, in contrast to a universal solver GA and other researches where the GA solves only a part of the problem. To accomplish this, an adequate framework was developed, that considers all critical GA elements. At the same time, the considered DARPTW instance was very specific in contrast to other studies, for example, there were only outbound customers. A bus-passenger representation for each gene on an individual was used. The solution decode is done by an ordered lecture, where the first occurrence of a customer is

always a pickup and the second one is always a delivery, this obligates the representation to consider only two genes per customer. The vehicle associated with the pickup is the one actually considered in the solution and the one of the delivery can be different but not considered. The final results showed an improvement on the solution quality factor, in contrast to the vehicle quantity, when compared to previous research.

It is important to highlight that none of the above solutions considers any technique for improving the search over the feasible solution space, especially regarding the time-windows constraints. This is especially true when applying the crossover and mutation operators, reason why the present work improves actual GA solutions by trying to avoid or minimize the infeasibilities and subsequent reparations after the operators through the use of pre-feasibility tables or schemas.

4 Implementation

In this section, the implemented GA framework and each of its elements will be described in the following.

4.1 Preliminary feasibility schemas

As stated before, the time window constraints make the problem search space highly non convex, meaning that it is very easy to move from a feasible solution to an unfeasible one when searching through meta-heuristics, being the only solution to roll-back or to repair the unfeasible solution.

In the present work an approach has been developed to minimize these problems by reducing a priori the feasible planning combinations considering the time windows restrictions.

When processing the dataset of requests, a preliminary precedence table of events is built, where each event (pickup or delivery) is associated a list of other events which need to be inserted before that event on a route (if assigned to the same vehicle) to preserve time windows feasibility. In Figure 1, for example, are shown 4 clients (A, B, C and D) with their time windows for pickup ($+$) and delivery ($-$). An obvious relation is that the pickup $A+$ must precede its delivery $A-$. Then, the pickup of C ($C+$) cannot be before the pickup of a ($A+$) as it would be impossible for the same vehicle to serve both events within their time window bounds as the last ends before the other stars.

In the most general case, an event Y must precede an event X when $ET_X + DRT_{XY} > LT_Y$, where ET_X corresponds to the Early Time (time windows lower bound) of event X, DRT_{XY} the direct ride time from the location of X to the location of Y and LT_Y to the Latest Time (time windows upper bound) of event Y. In this way, it is possible to build a list of precedence events for each event.

On the other hand, there is an incompatible clients' list that defines, on a level of passenger assignation (clusters), which clients cannot be transported by the same vehicle, for the time windows feasibility of a solution. This list can be obtained from the pairs of events that precede each other simultaneously: if an event X is preliminary to Y and at the same time Y is preliminary to X, then it is impossible for both to be part of the same route and, consequently, its respective clients cannot be transported by the same vehicle.

In practical terms, the case involved is when a couple of events are too far the one from the other and their time windows are too close to each other making it impossible for a single vehicle to go from one location to the other within their time windows constraints.

4.2 Initial Population Generation

It is based on the client incompatibility previously described. The mechanism is concerned that incompatible clients will not be assigned to the same vehicle, first generating a base feasible solution that

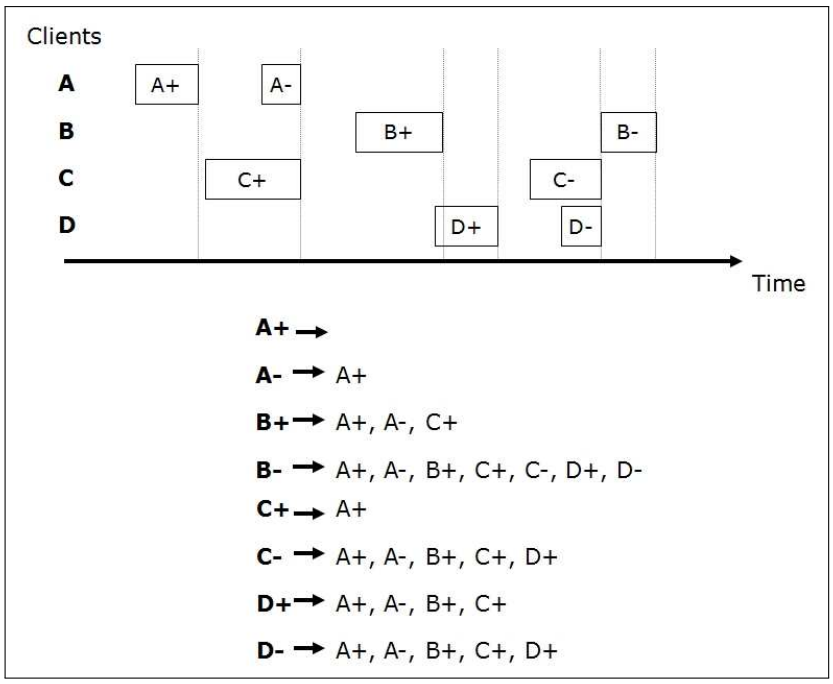


Figure 1: Example of preliminary precedence table of events, where each event has other events that must precede it when they are on the same vehicle

contains all conflicting clients on separated vehicles and afterwards generating a final feasible solution as a result of the insertion of the remaining clients over the base feasible solution.

The implemented mechanism does not assure the generation of a feasible solution on the first try. In this context, the randomness on the insertion heuristic and the vehicle selection for the clients, facilitates the possibility of restart the process when the incapacity of continue generating a feasible solution is detected.

4.3 Genotype and Crossover

A model like LLGA [7] has been considered. The incorporation of the locus on the genes allows to order them in different ways representing an identical solution. A gene is composed by the locus, that corresponds to the client number on this implementation, and the vehicle assigned to him. Figure 2 shows the clients A, B, C, D and the vehicles V1, V2, V3 to which are assigned.

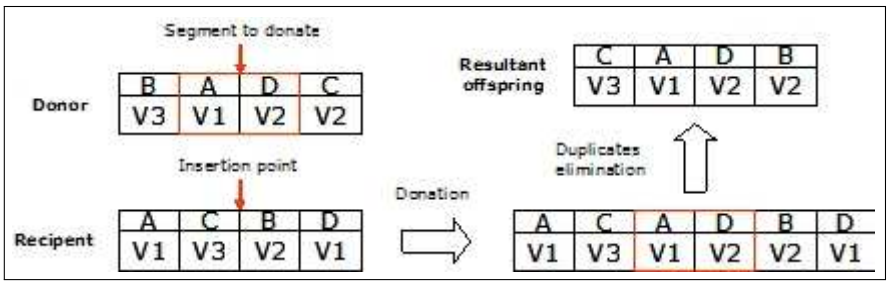


Figure 2: Crossover operation for the implemented genotype

This representation tackles the assignment of clients to vehicles (clustering) while the scheduling itself (route construction) is carried out through a greedy insertion heuristic explained in the next subsection.

On crossover, (see Figure 2) a father assumes the donor role, giving an own gene segment of its structure, on a random insertion point of the other father that assumes the recipient role, where the duplicated genes are deleted. In this case, the duplicated clients from the recipient are deleted. This generates a change on the clusters from the chromosome external layer, but also triggers the modification of the routes from the chromosome internal layer.

Because this process is not exempt from unfeasibility problems, in the basis of the randomness of the internal processes that support the crossover (insertions, eliminations, etc.) and the own crossover factors (insertion point, donated segment), a restart of the whole process is considered, until a feasible crossover has been done.

4.4 Route Scheduling

It regards the greedy insertion heuristic used for route generation, plus feasibility evaluation procedures. This mechanism is based on "MADARP" model shown on [6], applying the concept of time windows intersection and the use of pre-calculated data for a direct evaluation.

Figure 3 (a) shows a portion of a schedule, containing the pickups and deliveries of clients A and B on a first block, then a slack (that is, vehicle idle time without passengers onboard) and a second block serving clients C and D. Within the first block, the evaluation of client X is carried out by evaluating all the possible permutations of the new client in the sequence, that is, (n+1)(n+2)/2 with n the number of events already present in the block.

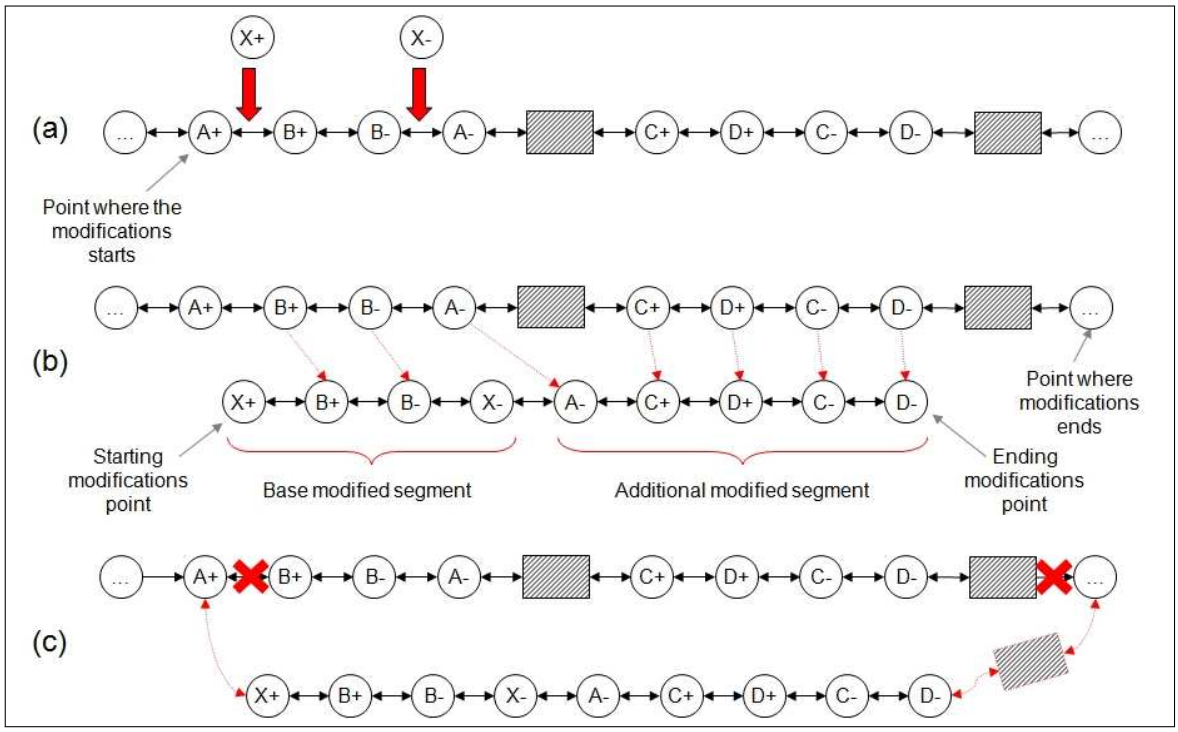


Figure 3: Crossover operation for the implemented genotype

In addition, the heuristic also considers the possibility of merging blocks as Figure 3 shows. The insertion of client X makes the slack time after $A-$ to disappear, resulting in a bigger block.

4.5 Phenotype or Evaluation

It is worth highlighting that there are two generic elements evaluated on a solution: The transport system efficiency and the quality of service offered to the clients. On the first element, the considered factors are the vehicle travel time, the slacks length and the vehicle quantity. On the second element, the excess ride time and the wait time are considered. A weight is assigned to each of these factors that directly influence the evaluation result.

4.6 Selection

For the proposed models, a tournament selection is used, from where a population subset is obtained, and an individual from these are selected as a winner. On this operator, the selection pressure can be managed through the tournament size. In this implementation, when a winner is obtained from the tournament, it is not removed from the original population, allowing them be the winner on future tournaments and facilitating the convergence.

4.7 Mutation

In contrast from other models described in literature, two mutation operators have been developed in this implementation: a cluster mutation and a route mutation. Each of these has its own probability, hence an individual can be affected by one, both or none of them on the generation advance. The cluster mutation consists in moving a client from one vehicle to another and the probability is applied to each client in the solution as Figure 4 (a) shows.

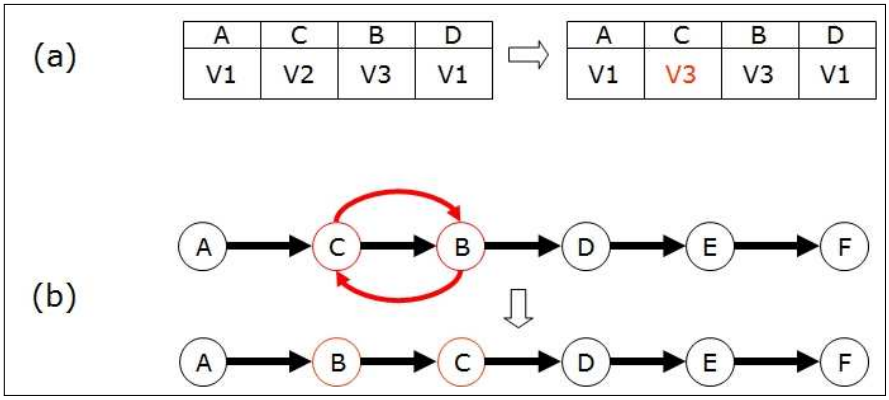


Figure 4: Crossover operation for the implemented genotype

The route mutation consists on interchanging the order of an event's pair from the route and the probability is applied to each route in the solution (see Figure 4 (b)). Because this operator is not exempt from unfeasibility problems and the mutation probabilities are small, when all possibilities has been evaluated and a feasible result has not been obtained, the process is discarded for a client or a route according to the respective mutation.

5 Experiment Settings

In literature, a common benchmark dataset used to evaluate heuristics for the DARPTW problem corresponds to the ones provided by Cordeau & Laporte [5]. These sets are divided on two subsets: One is used in [3] for an Branch-and-Cut algorithm, while the other one is used in [4] for a Tabu Search algorithm and in [1] for a cluster-first route-second GA. These are used on this research both for the GA

calibration stage and final testing stage. For the evaluation, the benchmark data was divided on three groups according to the number of clients: small sets (pr01, pr02, pr11, pr12 and pr17), medium sets (pr03, pr05, pr15 and pr19) and big sets (pr16).

On the calibration stage, the following parameters were considered: 3 sets (a small one, a medium one and a big one); Population size: 100; Maximum generation number: 10000; Crossover ratio: 0.35, 0.45 and 0.75; Cluster mutation ratio: 0.0025, 0.005 and 0.0075; Route mutation ratio: 0.025, 0.05 and 0.075; Tournament size: 2.

These give a total of 108 possible execution combinations. For each combination, 3 runs were done, generating a total of 324 executions on the calibration stage. The problem parameters used were: Route duration factor: 8; Slack time factor: 1; Vehicle quantity factor: 0; Excess ride time factor: 2; Wait time factor: 0; Ride time factor: 4.

Finally, the parameters associated to the best results were: Crossover ratio: 0.45, cluster mutation ratio: 0.005 and route mutation ratio: 0.075 for the Small set; Crossover ratio: 0.35, cluster mutation ratio: 0.075 and route mutation ratio: 0.075 for the Medium set; and Crossover ratio: 0.75, cluster mutation ratio: 0.025 and route mutation ratio: 0.025 for the Big set.

With these parameters, the final tests were executed. In this stage, a bigger number of runs per test and a bigger generation number were used. For each instance set, 20 runs and 15.000 generations were considered. A total of 200 tests with the fore-mentioned characteristics were carried out.

5.1 Obtained Results

The results were compared with Jorgensen et al. [2] research (cluster-first route-second GA) and Cordeau & Laporte [4] research on Tabu Search. Table 1 shows the results obtained by our work while Tables 2 and 3 show the results obtained by the previously mentioned researches. Because the compared models do not have the same characteristics, the comparison was done on the basis of time units of two critical factors: On the one hand, the total route duration that is associated with the transport system resources optimization, and on the other hand, the total client travel time that is associated with the offered quality of service.

LLGA Model					
Set	Route duration		Ride time		CPU Time (Min.)
	Best	Avg.	Best	Avg.	
pr01	955,25	1012,32	524,59	546,56	1,36
pr02	1839,06	1836,05	838,41	889,05	4,08
pr03	2787,18	2810,59	1597,95	1634,1	7,96
pr05	4068,05	4118,03	2935,48	3007,97	18,43
pr11	902,18	908,09	449,91	454,4	1,58
pr12	1503,34	1503,71	744,93	766,21	4,49
pr15	4057,08	4118,34	3152,67	3160,79	22,09
pr16	4658,35	4665,77	2348,48	2377,45	17,48
pr17	1223,68	1227,46	612,4	612,4	3,13
pr19	3427,06	3441,71	2515,53	2597,1	25,42

Table 1: Summary of the results obtained by our LLGA model

It is important to mention that our solution considers two restrictions not covered by both, Jorgensen et al. and Cordeau & Laporte researches. These are the time windows as hard constraints and the incapacity of vehicles to transport clients when they are on slack times. Despite this more restrictive scenario our solution performed well compared to them.

Jorgensen et al.					
Set	Route duration		Ride time		CPU Time (Min.)
	Best	Avg.	Best	Avg.	
pr01	1039	1041	310	447	5,57
pr02	1994	1969	1330	1367	11,43
pr03	2781	2779	2894	3081	21,58
pr05	4274	4250	4837	5099	58,23
pr11	928	907	549	630	5,46
pr12	1710	1719	1300	1214	11,72
pr15	4336	4296	4720	4615	58,93
pr16	5227	5309	6397	6134	81,23
pr17	1316	1299	784	990	8,29
pr19	3676	3679	5358	5362	44,66

Table 2: Summary of the results obtained by cluster first- route second GA of Jorgensenet al. [2]

Cordeau & Laporte			
Set	Route duration	Ride time	CPU Time (Min.)
pr01	881	1095	1,9
pr02	1985	1977	8,06
pr03	2579	3587	17,18
pr05	3870	6154	46,24
pr11	965	1042	1,93
pr12	1565	2393	8,29
pr15	3596	6105	54,33
pr16	4072	7347	73,7
pr17	1097	1762	4,23
pr19	3249	5581	51,28

Table 3: Summary of the results obtained by Tabu Search of Cordeau & Laporte [4].

In terms of solution quality, our LLGA model clearly performed better than the Jorgensen et al. solution. A closer look at the tables shows that for all the evaluated datasets our LLGA model presented lower average times for vehicles' route duration, clients' ride times and also CPU time, being this last one specially important.

Regarding Cordeau & Laporte results, based on a Tabu Search solution, on most cases their solution presented best times for the route duration but not for the ride times or the CPU times. In addition, on three cases (pr02, pr11 and pr12) our LLGA solution performed better also for the route duration.

5.2 Results Discussion

As exposed in the results section, our LLGA implementation presents better results for the clients' travel time with respect the other two studies. This is mainly due to the use of time-windows as hard constraints, making it easier to enforce the optimization on this factor. Then, when focusing on the vehicles' route duration our solution showed better times regarding the cluster-first route-second GA of Jorgensen et al. while obtaining worse results when compared to the Tabu-based solution. This can be understood as the cost of having good average ride times for clients, as there is a trade-off relation among both variables.

The improvement obtained in our solution is mainly explained because of the use of the so-called precedence table of events and the incompatible clients list. Both elements provide a way to reduce the search-space when considering the time-windows restriction.

In this sense, it is important to remember that the DARPTW problem behaves as a deceptive problem when solved through GAs, mainly because of the difficulty for the GA operators to find the appropriate building blocks for constructing feasible solutions. This fact is especially true when facing the DARP with TWs, as the time-windows do impose additional restrictions on the clients' requests that make the region of feasible solutions highly non-convex.

This means that it is very easy to move within this region from a feasible to an infeasible solution. Furthermore, depending on the tightness or looseness of the time-windows such region may look like a small group of feasible points spread over a big region of infeasible solutions.

Under such an scenario, what the precedence table does is to make such region more convex by pre-processing which sequences of events (pickup & delivery) are feasible within the "sea" of infeasible sequences due to incompatibilities in their respective time windows, making it impossible for a vehicle to serve both events in that sequence and satisfy their time intervals.

A similar situation happens with the incompatible clients' list. In this case, both events of the client are considered, his pickup and delivery, detecting the cases in which it is not possible for a vehicle to serve both clients and their time-windows constraints no matter the sequence used and no matter which other clients are assigned. This is important for identifying requests that are too close in time (either at pickup or delivery) while being too far geographically, avoiding putting them together.

In this way, the initial population and GA operators move more around feasible solutions. This reduction follows a similar principle as in Constraint Satisfaction Problems and their techniques to reduce the domain of variables.

From other point of view, it can be seen as a search with memory, as in Tabu search. However, in this case the memory is used for making the algorithm to "remember" which portions of sequences are feasible in order to reduce effort instead of remembering the solutions found so far to avoid local optima. Another important issue is that the CPU time in all cases is lower, being sometimes even the half of it or even less (e.g. pr5, pr15, pr16 and pr19). It is worth highlighting that the LLGA tests were done with a 2.66 GHz Intel Pentium 4 CPU, while the Cordeau & Laporte tests were done with a 2.0 GHz Intel Pentium 4 CPU and the Jorgensen et al. tests were done with a 2.0 GHz Intel Celeron CPU. Although the hardware configurations are dissimilar, they do not completely justify the time improvement. Undoubtedly, the precedence table of events and the incompatible clients' list have caused this

diminishing.

6 Conclusions

The proposed model has shown interesting results according to the comparisons made, highlighting the times lowering, undoubtedly the weaker GA factor. This work will allow researchers to develop comparisons of highly restrictive models, unlike most works in literature where a bigger number of constraints are relaxed.

There are two important elements of the developed model on this work that must be remarked. On the one hand, the use of pre-feasibility schemas has represented an interesting support tool for the GA behavior. Although it depends strictly on the problem, the GA behaves better when the search space is bigger and complex, as in many NP-hard problems while having a convex feasible region of solutions.

Bibliography

[1] K. Bergvinsdottir, *The Genetic Algorithm for solving the Dial-a-Ride Problem,* Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2004.

[2] R. M. Jorgensen, J. Larsen, K. B. Bergvinsdottir, Solving the Dial-a-Ride problem using genetic algorithms, *J. Oper. Res. Soc.*, Vol.58, pp. 1321-1331, 2006.

[3] J. Cordeau, A Branch-and-Cut Algorithm for the Dial-a-Ride Problem, *Operations Research*, Vol. 54, pp. 573-586, 2006.

[4] J. Cordeau, G. Laporte, A tabu search heuristic for the static multi-vehicle dial-a-ride problem, *Transportation Research Part B*, Vol. 37, Issue 6, pp. 579-594, 2003.

[5] J. Cordeau, G. Laporte, *Benchmark datasets* accessed on march-2007, available online at: http://neumann.hec.ca/chairedistributique/data/darp/. 2007.

[6] C. Cubillos, *MADARP: Multi-Agent Framework for Transportation Systems,* Thesis for the academic degree of Dottore di Ricerca in Ingegneria Informatica e dei Sistemi (ciclo XVII), Politecnico di Torino, 2005.

[7] C. Cubillos, N. Rodriguez, B. Crawford, A Study on Genetic Algorithms for the DARP Problem, *Springer Berlin-Heidelberg LNCS*, vol. 4527, pp. 498-507, 2007.

[8] G. Harik, D.E. Goldberg, Learning linkage, *Foundations of Genetic Algorithms*, vol. 4, pp. 247-262, 1996.

[9] J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to biology, control and artificial intelligence,* MIT Press, ISBN 0-262-58111-6, 1998.

[10] M.W.P. Savelsbergh, The General Pickup and Delivery Problem, *Transportation Science*, Vol. 29, pp. 17-29, 1995.

[11] S. Thangiah, Vehicle Routing with Time Windows using Genetic Algorithms, *Application Handbook of Genetic Algorithms: New Frontiers*, Vol. II. Lance Chambers (Ed.). CRC Press, 1995.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 137-147

Presentation of an Estimator for the Hurst Parameter for a Self-Similar Process Representing the Traffic in IEEE 802.3 Networks

Ginno Millán, Gastón Lefranc

Pontificia Universidad Católica de Valparaíso
Escuela de Ingeniería Eléctrica
Avda. Brasil #2147. Valparaíso - Chile
E-mail: ginno.millan@gmail.com and glefranc@ucv.cl

Abstract: The hypothesis for the existence of a process with long term memory structure, that represents the independence between the degree of randomness of the traffic generated by the sources and the pattern of traffic stream exhibited by the network is presented, discussed and developed. This methodology is offered as a new and alternative way of approaching the estimation of performance and the design of computer networks ruled by the standard IEEE 802.3-2005.
Keywords: Computer networks, IEEE 802.3-2005 standard, network traffic, self-similar process.

1 Introduction

The positioning and consolidation of Ethernet as a predominant standard in the local and extensive coverage computer network field against some traditional technologies, such as Frame Relay, DQDB and ATM, are facts that are explained by its main characteristics, i.e. compatibility and interoperability among some Ethernet equipment from different speeds, high performance, scalability and self-configuration capacity, independence IP addressing and, undoubtedly, the usual scale economy.

Ethernet, initially at 3 Mb/s, has evolved from 10 Mb/s to 10 Gb/s in the last twenty-two years. The latter does not consider the IEEE 802.3ba standard, which specifies the Ethernet at 40Gb/s and 100 Gb/s since 2007. Furthermore, it has also evolved from using simple bridges developed for the network bridging with identical physical protocols and average access to N*10 Gb/s switches [1], [2]

Two very interesting, striking and critical aspects are involved in this continuous evolution underwent by the Ethernet networks. The first is related to the complete abandonment of the origin of the half-duplex shared system, which has yielded to full-duplex links. On the other hand, the second involves its extension, as Ethernet has evolved from LAN range distances to WAN range coverages [3]. Furthermore, although both changes have gradually occurred, they result quite critical from the Ethernet point of view, as both represent the disappearance of the mechanism controlling the media access regulated by the CSMS/CD protocol and also assume a strong and radical change in the transmission media, which tends to the massive and total use of the optical fiber to support the ever demanding application of enormous bandwidths.

In general, the local area networks, particularly Ethernet, were developed as high capacity interconnected networks; in contrast to WAN network technologies that are based on switching and transmission flow generally lower than those available in the LAN networks. However, the evolution of the different technologies used in both environments now converges in solutions based in Ethernet and its different specifications. Therefore, the current Ethernet networks are switched and almost totally made from full-duplex links, but they also incorporate the multiplexing according to the IEEE 802.1Q standard and bear transmission distances similar to those involved in the conventional WAN links. This degree of evolution

is largely ascribed to great development achieved by the Ethernet switches, as there has not only been an increase in the operational transparency and simplicity degree but also a direct effect in the incorporation of additional functionality to the switching, which, from the standard point of view, involves the extension of the format of the original frame and the incorporation of labeling for VLAN and the priority establishment for service classes, the size increase of the CSMA/CD carrier signal, the incorporation of burstiness packets to try to compensate the network velocity loss caused by the extension bits of the carrier and, most of all, in the complete abandonment of the restraint and solution conflict mechanism given by the exponential backward algorithm [4], [5].

On the other hand, the migratory tendency towards Ethernet networks without shared media is confirmed through the incorporation of the IEEE 802.1X, IEEE 802.1w standards (RSTP currently included in the IEEE 802.1D standard) and the IEEE 802.1s standard (MSTP; currently included in the IEEE 802.1Q standard). These establish the dedicated full-duplex links as an essential requirement for the correct operation. It is important to point out these dedicated links are not only necessary to obtain the best benefits from the network, but also to ensure the logical link control (LLC), simplifying the protocols and making possible the rapid convergence mechanisms in layer two.

One example of the latter is the IEEE 802.3e standard (Ethernet at 10 Gb/s), which does not involve the use of half duplex links in the specifications as the IEEE 802.3z (Ethernet at 1000 Mb/s). The use in the IEEE 802.3z is strictly related to the compatibility with the equipment database previously installed and its aim is to work as a platform in the migration or technological transition processes. It is important to remember that the IEEE 802.3z is the latest specification in the IEEE 802.3 standard giving native support to this device communication method.

It seems quite interesting the diffusion and flooding dominated in the traditional Ethernet networks as the basic and valid mechanisms to establish the presence or absence of stations; nowadays, however, a minimum diffusion of the frames is sought due to the performance degradation and the exhaustive control of the traffic flow, the same reasons that must be avoided in the WAN networks.

New technologies are similarly incorporated to the large covering networks, which are typical of the LAN environments due to their robustness and great price/assistance relation. These are well established in access and metropolitan environments and, increasingly, in WAN environments [6].

All these arguments guarantee a reassessment of the study of the benefits of the CSMA/CD access control mechanisms, in terms of the impacts of the performance in the current commuted environments, i.e. considering the performance as the useful information quantity that the network is able to transport in relation to the real quantity of transported bits, as well as a characterization of the nature of the traffic under study in terms of a performance pattern capable of describing the temporal evolution as well as the implications on the previously defined performance.

The performance parameter is considered as an active form of measuring the benefits from a network, as this is one of the crucial aspects in the global analysis of the communication systems, considering its impacts in the final users. On the other hand, the traffic characterization among the networks is considered, because the performance observed in the performance parameter depends on it, thus becoming a key factor for the characterization.

Furthermore, a new approach to carry out the modeling processes of the Ethernet networks is justified, as in terms of the evolution previously stated, it is inferred that the natural successor of the IEEE 802.3u standard (Fast Ethernet) must be the IEEE 802.3z standard (Gigabit Ethernet), the latter will give way to the access and WAN environments, the IEEE 802.3ae (Ethernet at 10 Gb/s) and IEEE 802.3ba (Ethernet at 40 Gb/s and 100 Gb/s) standards, respectively. The impacts associated to these technological migrations must be adapted and properly evaluated, sized up and classified in terms of the impacts on the installed equipment databases. This must be previously defined before any adoption.

However, the existence of self-similar traffic patterns is accepted on the empirical fact that these are characterized by the constant presence of package traffic bursts through different time scales and that the characteristic property of the self-similar processes is the Long Range Dependence (LRD), which

is observed when the trunking level increases [7]. Then, the fractal performance of this type of traffic does not coincide with the performance traditionally modeled through the Poisson processes, which are characterized by the absence of bursts and a low variability that is reflected in the temporal independence between the samples. In short, these are processes that show a short-term temporal dependence rejecting the relation between temporarily distant processes, i.e. these are null memory processes that, therefore, do not considered as valid the presence of any pattern representing the sent traffic.

Then, considering the same arguments and the fact the self-similar phenomena show the same aspect or performance when visualized with different enlargement degrees or different scales of a certain dimension, and also that the temporal series showing self-similarity with respect to time are the object of interest in computing networks, a degree of self- similarity must be established for such series, which must be expressed using a parameter representing the de-growth speed of the self-correlation function. This responds to the fact that a time series is self-similar when the aggregated series involves the same self-correlation function as the original series. The latter is achieved using the Hurst parameter, H, which can be estimated through different methods, being the Whittle the one with the highest statistical rigor.

The Whittle estimator calculation may be carried out from different algorithms, all of which need the underlying stochastic process. In this paper a variation of this estimator is proposed, which allows the obtaining of the self-similarity degree with an acceptable commitment relation between the cost of the computing model (higher inconvenient when obtaining the Whittle estimator no matter the chosen method) and the estimation quality.

This proposal formally involves the modification of the Whittle local estimator, or semi-parametric Gaussian estimator, of the memory parameter in the short-term standard process stated in [8]. Then, all the advantages of the original technique are expected, which shows all its main attributes as an alternative to the regression technique of the periodogram logarithm shown in [9]. It is expected that under less restrictive suppositions, an asymptotic efficiency profit form is shown. Therefore, an analysis of the asymptotic performance of the original semi-parametric Gaussian estimator must be carried out on the memory parameter in processes with cyclic or seasonal memory, thus allowing divergences or spectral zeros of asymmetric type. Then, modifying the original algorithm, the consistency and asymptotic normality needed for the characterization of the traffic flows will be obtained.

2 Problem Definition

The traffic analysis based on the queuing theory has resulted in great help to the network design and the system analysis when carrying out capacity planning and performance prediction [10]. However, there are some cases in the real world in which all predicted results from a queuing analysis significantly differ from the performance observed in reality [11]. In this sense, it must be remembered the validity of the analysis based on the queuing theory depends on the Poisson nature of the data traffic, and when dealing with Poisson processes, the representation of the length of each arrival and the time between each frame arrival are represented by independent and exponentially arranged random variables. Therefore, these are null memory models. Thus, being this the case, of models in those that the probability of an arrival in an instant, is independent of the instants of previous arrivals, property that is not completed in the nets of commutation of packages. However, it is acknowledged the objective of these suppositions correspond to relatively simple models, from an analytical point of view.

According to the results obtained in [12], different authors have studied the existence of a temporal dependence and the establishment of the great impact involved in the assistance of a queuing system outstands. Therefore, there is plenty of literature on entry traffic models showing more or less complex correlation structures, which are used in cases when the telecomputing system model under study allows an adequate analytical treatment. In any case, all these models, mainly of Markovian nature, reject the correlation from a given temporal separation, even when it may be arbitrarily increased at the expense of making the model complicated with additional parameters.

In [13] and [14] is shown that after several measurements on an Ethernet network, the traffic shows a self-similar or fractal nature (understanding traffic as the frame quantity in the network by a time unit). This becomes apparent with the existence of a long-term correlation.

The self-similar characteristic of the traffic in WAN networks is shown in [15] and [16]. Furthermore, the fractal nature of the data flow from the protocols involved in the signaling system 7 (SS7) in the common channel signaling networks is shown in [17].

The self-similar nature of the traffic due to the WWW is shown in [18], using the experimental evidence, the possible causes and origin are shown.

The self-similar nature in the traffic of the variable bit rate (VBR) is shown in [19] and then in [20].

In [21] is shown that the probability distribution that is followed by the queuing size of a multiplexor shows a Weibull-type asymptotic fall when using certain self-similar process as the entry traffic. Then, in [22] it is shown this fall is even slower and hyperbolic when other self-similar processes are used.

Reference [23] shows the effective band width estimated through Markov models, where the queue size distribution show an exponential fall, highly underestimates the loss rate of cells in different orders of magnitudes in ATM networks.

An outstanding long-term positive correlation in the added traffic is shown in [24], from the analysis of a voice and data multiplexer, where some high delays were obtained, higher than those obtained using the Poisson models.

In reference [25] the number of entries is shown, counted in adjacent time intervals that result in the superposition of multiple independent and homogeneous voice sources susceptible to be treated through updating process models. This is a complex process that includes strong correlations and involves a significant impact in the assistance of the studied system.

The interesting thing about this and other studies is that they highlight the impact the long-term temporal dependence have on the assistance of the communication networks, which is intrinsic to the most diverse traffic types, in contrast to other models that do not include this due to analytical simplicity (renewal processes, for example) or that show a more analytically complex correlation structure (such as the Markov or generally self-regressive models). All these are known as short range dependence models (SRD). Furthermore, the main problem these traditional models involve is that they require a great number of parameters to characterize the strong correlations existing among different traffic types in a network. In this study, it is clearly stated that while the parameter number increases, the analytical complexity also increases (not necessarily in a lineal form) and that there is also a difficulty to give a physical interpretation to all the parameters and to estimate them from empirical data.

The self-similarity and fractability characteristics help to describe a phenomenon where certain object property is preserved in relation to the temporal and spatial scalability and in a self-similar and fractal object. Thus, the magnified parts are similar to the form of the complete object, where the similarity is measured in some proper form. Then, the simpler form of self-similarity is obtained by reconstruction through the iteration of a certain procedure. If this process is indefinitely repeated for each new segment, any portion of the object can be magnified to exactly reproduce a bigger portion, no matter how small the portion. This property is known as “exact self-similarity” [26]. No exact self- similar characteristics are intended to be observed in a highly random process such as the one involving the packets arrival to a data network; however, the traffic observed as sample traces from a stochastic process are considered and the similarity is only restricted to certain specific statistics of the temporal series though adjusted in scale. Furthermore, the exact self-similarity in abstract mathematical objects and approximate self-similarity for each considered specific execution are also discovered. No exact self-similarity is expected from the network traffic under study; however, a self-similar stochastic behavior is expected and, according to this, the second order statistics may be used to capture the process variability in order to determine the self-similarity rate. In fact, the scale invariance may be defined in terms of the self-correlation function, as the polynomial decrease of this function (in place of the exponential) represents a long-term dependence manifestation similar to the self-similarity, and constitutes the view from which all self-similar

processes must be interpreted through the development of this investigation.

On the other hand, in relation to the fundamental problem of the self-similar process analysis, or more precisely, to the temporal series showing LRD, i.e. the estimation of the Hurst parameter (H), the methods involved in the literature can be classified in the following two groups:

1. Geographical methods of lineal regression. Used to estimate a statistic $T(x)$ behaving asymptotically for a determined group of x values and, therefore, are based in obtaining the straight line that better adjusts (for that group of x values) to the $\log(T(x))$ against the $\log(x)$ using the least square lineal regression, thus obtaining the H parameter value directly from the gradient value of that particular straight line.
2. Methods based on Maximum Likelihood Estimators (MLE) for H. These help to minimize the differences between the periodogram of the concerning series and its theoretical spectrum.

The methods involved in the first group are relatively simple and algorithmically fast to implement. However, their main disadvantage is that an asymptotic performance must be first estimated from a finite sample number, which makes the estimation of the H parameter directly and considerably dependant from the right selection of the x value group. Therefore, the graphic representations result crucial for the confirmation of correspondence between the x values and the lineal performance area and the proper adjustment of the straight line for the represented points. Furthermore, it is important to point out these methods only allow a precise estimation of H, as the resulting confidence intervals involves a high cost for the computing resource and long processing periods, both as a result of the use of the intensive graphic-type methods. The MLE-based methods, on the other hand, though more complex and a higher computing cost, result more flexible and efficient from the statistical inference point of view, as they allow the obtaining of confidence intervals for the H estimated values. Therefore, these are the most common methods.

This investigation will deal with MLE-based methods to solve the confidence intervals directed to obtain the first approximations for an H value, which will be then adjusted using the proposed method. In any case, the methods from the first group will be thoroughly analyzed and the results will be contrasted with those resulting from the MLE and the adjustment suggested by the proposed method. Then, and as the analysis starts using MLE, it is important to stress these are methods designed to minimize the differences between the periodogram of the series and the parametric model suggested for the theoretical spectral density. Besides, the exact estimation of the MLE is quite expensive form the computing point of view, so likelihood Gaussian functions are generally used (Gaussian MLE). However, even when these kind of functions are considered, the computing costs results quite high, thus the approximations based on Gaussian MLE are generally used and the most widely used approximation is the Whittle approximation; therefore, the suggested method is based on the modification of one of the variables for the result.

3 Justification and Proposition of an Efficient Whittle Estimator

Being $f(\lambda, \theta)$ the parametric form of the spectral density of a Gaussian stationary process, X_t where, $\theta = (\theta_1, \ldots, \theta_M)$ is the parameter vector to estimate. Then, being $I(\lambda)$ the periodogram of samples defined by

$$I(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} X_t e^{jt\lambda} \right|^2 \tag{1}$$

The approximated MLE o Whittle is the vector

$$\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_M) \tag{2}$$

minimizing the function

$$Q(\theta)\widehat{=}\frac{1}{2\pi}\left\{\int_{-\pi}^{\pi}\frac{I(\lambda)}{f(\lambda,\theta)}d\lambda+\int_{-\pi}^{\pi}\log[f(\lambda,\theta)]d\lambda\right\} \tag{3}$$

in practice, the estimation of the Whittle estimator is done choosing an adequate scale parameter θ_1, complying with

$$f(\lambda,\theta)=\theta_1 f(\lambda,\theta^*)=\theta_1 f^*(\lambda,\eta) \tag{4}$$

thus annulling the second addend of (3), i.e.

$$\int_{-\pi}^{\pi}\log[f(\lambda,\theta^*)]d\lambda=\int_{-\pi}^{\pi}\log[f^*(\lambda,\eta)]d\lambda=0 \tag{5}$$

where $\eta=(\theta_1,\theta_2,\ldots,\theta_M)$ y $\theta^*=(1,\eta)$

In [27] is shown the scale parameter is given by $\theta_1=\frac{\sigma_\varepsilon^2}{2\pi}$, being the mean-square prediction medium (MSPE).

On the other hand, the discreet version of the Whittle estimator is suggested in [28], which is approximated to 3 using a Riemann addition in the frequency range given by $\lambda_k=2\pi N^{-1}k$, with $k=1,2,\ldots N^*$ (being N^* the integer part of $(N-1)/2$). Then, the function to be minimized is given by the expression

$$\widetilde{Q}(\theta_1,H)=\frac{4\pi}{N}\left\{\sum_{k=1}^{N^*}\frac{I(\lambda_k)}{f(\lambda_k,\theta_1,H)}+\log[f(\lambda_k,\theta_1,H)]\right\} \tag{6}$$

The estimated H parameter is obtained through the selection of the adequate scale parameter, $\hat{H}$, a value that minimizes the following expression

$$\widetilde{Q}^*(H)=\widetilde{Q}(1,H)=\sum_{k=1}^{N^*}\frac{I(\lambda_k)}{f(\lambda_k,1,H)}=\sum_{k=1}^{N^*}\frac{I(\lambda_k)}{f^*(\lambda_k,H)} \tag{7}$$

where it is verified that

$$f^*(\lambda,H)=\frac{1}{\theta_1}f(\lambda,\theta_1,H)=\frac{2\pi}{\sigma_\varepsilon^2} \tag{8}$$

The following disadvantages of the Whittle estimator stand out in its conventional forms, from this point of view:

1. Determination of the parametric form of the spectral density.
2. High estimation period due to graphic method use.

The Central Limit Theorem for self-similar processes result quite useful from the perspective of the application of the Whittle estimator to processes where it is not possible to assure anything in relation to the spectral density, as it is a good approximation for non-Gaussian series, which allow the application of aggregated series to any result from the pure self-similar Gaussian processes, as shown in [29]. Then, it seems quite interesting this theorem allows the supposition that for a temporal series of N size, whose self-correlation shows a LRD hyperbolic fall, if m and N/m are sufficiently big and the variance is finite, then the FGN process is a good approximation for the aggregated sequences of the series, even when it does not represent a Gaussian approximation [30].

The latter is the base for the variant of the Whittle estimator known as the added Whittle estimator, which gives a more robust and less biased form of the Whittle estimator when no information on the

exact parametric form of the spectral density is available. In other words, a shorter representative series is obtained given by the expression

$$X_k^{(m)} = \frac{1}{m} \sum_{i=km}^{(k+1)m-1} X_i \quad 0 \leq k \leq [N/m] \tag{9}$$

And then the Whittle estimator is used, considering the Fractionary Gaussian Noise (FGN) as the parametric model of its spectral density. However, in spite of the fact of using a shorter series considerably reduces the computing cost, the observed problem is that the variance of the estimator increases and so does the self-similarity degree, thus reducing the pattern representativeness degree.

Another problem associated to the method is the impossibility to a priori know the value of the appropriate m. However, in this last sense, a method to represent the estimation of the H parameter are shown in [31], which were obtained for different m values, and finding a region where the graphics are approximately plain.

The local Whittle estimator is shown in [32], which in contrast to the Whittle estimator, represents a semi-parametric estimator that only specifies the parametric form of the spectral density for those frequencies close to zero, i.e.,

$$f(\lambda) \sim G|\lambda|^{1-2H} \tag{10}$$

when $\lambda \to 0$

When replacing $f(\lambda, H)$ given by (10) in (2), and integrating up to the $\frac{2\pi M}{N}$ frequency, with $\frac{1}{M}+\frac{M}{N} \to 0$ when $N \to \infty$, the following is obtained

$$Q(G,H) \widehat{=} \frac{1}{M} \sum_{j=1}^{M} \left[\frac{I(\lambda_j)}{G\lambda_j^{1-2H}} + \log(G\lambda_j^{1-2H}) \right] \tag{11}$$

When replacing the G constant by its estimation, given by

$$\widehat{G} = \frac{1}{M} \sum_{j=1}^{M} \frac{I(\lambda_j)}{\lambda_j^{1-2H}} \tag{12}$$

the function to be minimized is obtained, i.e.,

$$R(H) \widehat{=} Q(\widehat{G}, H) - 1 \tag{13}$$

$$R(H) \widehat{=} \log \left[\frac{1}{M} \sum_{j=1}^{M} \frac{I(\lambda_j)}{\lambda_j^{1-2H}} \right] - (2H-1) \frac{1}{M} \sum_{j=1}^{M} \log(\lambda_j) \tag{14}$$

Nevertheless, the problem still remains, as the selection of the M value results critical and from this value the bias and variance depend, and yet again the bias and variance problem shows. However, as long as M increases, the estimated value for H will quickly converge to the real H value but the spectrum form will continuously fall apart from (6) and the SRD effects will be even greater, thus increasing the bias. Then, as in the previous method, one must choose to represent the estimated value of H against m and find the plain region of the graphic. As it has been observed from the methods about the Whittle estimator, all of them require the minimization of an expression, (6) or (14). The most obvious way of carrying out such minimizations is to assess these expressions for a certain number of equidistant H values (q), which will depend on the selected resolution. However, it must be pointed out that as lightly higher number of samples are needed for the algorithm to be immediately translated into a very high computing cost.

Then, a reduction through an algorithm is suggested to reduce the computing cost, which is aimed to decrease the number of points to be assessed. Therefore, the convex characteristic of the function in all the domain $[0.5, 1($ will be considered also that the minimal is then unique. Then, a searching method by bi-section used on the function derivative will have to allow the number of evaluated points to be around $\log 2(q)$, which will be of advantage in a significant sample saving, and at the same time it does not settle in relation to the bias and variance commitment.

Therefore, considering the assessment of the derivative in a H_i point may be approximated by a difference coefficient for a sufficiently small increase, h, i.e.,

$$Q'(H_i) \approx \frac{Q(H_i + h) - Q(H_i)}{h} \tag{15}$$

for $h \to 0$

It is derived the main work hypothesis is related with the establishment of a self-similarity degree based on the Whittle estimator; however, due to the complexity involved in the bias and variance commitment behind all models, it is then suggested that in a reduced point spectrum it is possible to give an answer with an acceptable level of commitment. Therefore, in order to obtain confidence intervals, and taking advantage of the convexity of the function to minimize in the whole dominium $[0.5, 1($, which implies the presence of a unique minimum, it will be considered that when estimating a unique H parameter, if $\widehat{H}$ represents the value minimizing the $Q(H)$ function and that H_0 represents its real value, then

$$(\widehat{H} - H_0) \to N(0, \sigma_H) \tag{16}$$

where the σ_H parameter will be defined by:

$$\sigma_H^2 = \frac{4\pi}{N} \left\{ \int_{-\pi}^{\pi} \left[\frac{\partial \log[f(\lambda, H)]}{\partial H} \right]^2 d\lambda \right\}_{H=H_0}^{-1} \tag{17}$$

Then, the estimation of the derivative given by (14) may also approximate through a difference coefficient for a sufficiently small increase, h. However, in this case, the chosen value for θ_1 different, because a cancellation is produced. Then,

$$\sigma_H^2 = \left\{ \sum_{k=1}^{N^*} \left[\frac{\partial \log[f(\lambda_k, \theta_1, H)]}{\partial H} \right]^2 \right\}_{H=\widehat{H}}^{-1} \tag{18}$$

Where the [33] it is suggested in the end that

$$\sigma_H^2 = \left\{ \sum_{k=1}^{N^*} \left[\frac{\log[f(\lambda_k, \theta_1, \widehat{H} + h)]}{\partial H} - \frac{\log[f(\lambda_k, \theta_1, \widehat{H})]}{\partial H} \right]^2 \right\}^{-1} \tag{19}$$

4 Conclusions

The hypothesis on the existence of a process with a temporary structure of long-term memory was presented, discussed and developed, which was representative of the independence between the randomness degree of the traffic generated by the sources and the pattern of the traffic flow shown by the network. The latter has been shown to be considered as a new form and alternative to cover the performance and network design estimation topics that are ruled by the IEEE 802.3-2005 standard.

The traditional models based on the Poisson processes, or even more general, based on short-term dependency processes are unable to describe the performance of he current data networks, particularly

those related to the switched Ethernet networks according to the IEEE 802.3-2005 standard. Consequently, it is necessary to redefine the study of the load systems, considering self-similar entry processes as a result of the self-similar traffic demand imposing new requirements in the network design, especially in what the buffering strategies are concerned. All methods traditionally used to assess the Whittle estimator show the disadvantages of the need to know the parametric form of the spectral density, as well as a high computing cost resulting from the intensive use of graphic methods. It is estimated these problems may be solved as far as it is feasible to introduce an algorithm that reduces the number of points to be assessed. This does not only mean reduced costs in computing processing, but also a new alternative to be considered in the study of the self-similar or fractal traffic consideration on the benefits of a network.

A convex function to carry out the minimization of the generalized function of the local Whittle estimator is suggested for a delimited domain, which involves the advantage to include a sole minimum that is completely individualizable, and thus using a bi-sectional searching method applied on the derivative of the function, should allow the determination of a point around which all values fluctuate. Therefore, this allows the availability of a plain region in which the value of the H parameter is perfectly approximated by a difference coefficient. The latter must be translated into savings in the computing costs and processing time that endorse the new proposed model.

The development of the simulations is being tackled through literature search as a first stage, which must clearly include the simulation techniques and statistical analysis for long-term dependency series, as it is considered it is not enough to be able to deduce the results from some aggregations, but there must be a standardization of the procedures to specifically study the long-term dependencies. This area requires urgent attention, as most of the operational costs involved in the computing capacity directly depend on it.

Bibliography

[1] Metcalfe, M. and Boggs, R. Ethernet: Distributed Packet Switching for Local Computer Networks.*Communications of the ACM*, Vol. 19 N°7, 1976.

[2] Ibańez, G. Contribución al Diseño de Redes de Campus Ethernet Autoconfigurables. Ph.D. thesis, Dept. Ing. Telemática, Universidad Carlos III de Madrid, Madrid, España, 2005.

[3] Ibańez, G. Contribución al Diseño de Redes de Campus Ethernet Autoconfigurables. Ph.D. thesis, Dept. Ing. Telemática, Universidad Carlos III de Madrid, Madrid, España, 2005.

[4] García, J., Ferrando, S., and Piattini, M., *Redes para Proceso distribuido*, Madrid, Ra-Ma, pp. 127-160, 1997.

[5] Zacker, C., *Redes. Manual de Referencia*, Madrid, McGraw-Hill, pp. 275-341, 2002.

[6] Halabi, S.,*Metro Ethernet. The Definitive Guide to Enterprise and Carrier Metro Ethernet Applications*. Indianapolis, Cisco Press, pp. 1, 2003.

[7] Leland, W., Taqqu, M., Willinger, W., and Wilson, D., On the Self-similar Nature of Ethernet Traffic, *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1-15, 1994.

[8] Robinson, P., Gaussian Semiparametric Estimation of Long-range Dependence,*Annals of Statistics*, Vol. 3, no 1995b, pp. 1630-1661, 1983.

[9] Geweke, J., and Porter-Hudak, S., *The Estimation and Application of Long Memory Time Series Models*. J. Timer Ser. Anal. 4, pp. 221-238, 1983.

[10] Stallings, W., *Internet y Redes de Alta Velocidad. Rendimiento y Calidad de Servicio*. 2nd ed., Madrid, Pearson Prentice Hall, pp. 224-225, 2004.

[11] Stallings, W., *Internet y Redes de Alta Velocidad. Rendimiento y Calidad de Servicio*. 2nd ed., Madrid, Pearson Prentice Hall, pp. 224-225, 2004.

[12] Kleinrock, L., *Communication Nets*, New York, McGraw-Hill, 1972.

[13] Leland, W., Taqqu, M., Willinger, W., and Wilson, D., On the Self-similar Nature of Ethernet Traffic,*Computer Communications Review*, Vol. 23, pp. 183-193, 1993.

[14] Leland, W., Taqqu, M., Willinger, W., and Wilson, D., On the Self-similar Nature of Ethernet Traffic, *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1-15, 1994.

[15] Klivansky, S., Mukherjee, S., and Song, C., *Factor Contributing to Self-similarity over NFSNet*, Georgia Institute of Technology, 1995.

[16] Paxon, V., and Wilson, D., Wide-area Traffic: The failure of Poisson Modeling, *IEEE/ACM Trans. Networking*, Vol. 3, no. 1, pp. 266-244, 1995.

[17] Duffy, D., Mcintosh, A., Rosenstein, M., and Willinger W., Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks, *IEEE Journal on Selected Areas in Communications*, Vol. 12, pp. 544-551, 1994.

[18] Crovella, M., and Bestavros, A., Self-similarity in World Wide Web Traffic: Evidence and Possible Causes, *IEEE/ACM Trans. Networking*, Vol. 5, no. 6, pp. 835-846, 1997.

[19] Garret, M., and Willinger, W., Analysis, Modeling and Generation of Self-similar VBR video Traffic, Proc. *ACM SIGCOMM'94*, pp. 269-280, London, 1994.

[20] Beran, J., Sherman, R., Taqqu, M., and Willinger, W., Long-range Dependence in Variable-bit-rate Video Traffic, *IEEE Trans. Communications*, Vol. 24, no. 2, pp.1566-1579, 1995.

[21] Norros, I., A Storage Model with Self-similar Input, *IEEE Tans. Queueing Systems*, Vol. 16, pp. 387-396, 1994.

[22] Likhanov, N., Tsybakov, B., and Georganas, N., Analysis of an ATM Buffer with Self-similar "Fractal") Input Traffic, Proc.*IEEE INFOCOM'95*, pp. 985-992, Boston, MA, 1995.

[23] Elwaid, A., and Mitra, D., Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High-speed Networks, *IEEE/ACM Trans. Networking*, Vol. 1, no. 3, pp. 329-343, 1993.

[24] Sriram, K., and Whitt, W., Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data, *IEEE Journal on Selected Areas in Communications*, Vol. 4, pp. 833-846, 1986.

[25] Heffes, H., and Lucantoni, D., A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance, *IEEE Journal on Selected Areas in Communications*, Vol. 4, no. 6, pp. 856-868, 1986.

[26] Bravo, J., and Marrone, L., Tráfico Autosimilar. Algoritmo Algebraico para Asignación Dinámica del Buffer, *Revista INGENIUS*, Ed. 2, Facultad de Ingenierías, Universidad Politécnica Salesiana, Ecuador, 2007.

[27] Beran, J., *Statistics for Long-memory Processes, in Generalized Additive Models (Monographs on Statistics and Applied probability)*, New York, Chapman & Hall, 2000.

[28] Geweke, J., and Porter-Hudak, S., *The Estimation and Application of Long Memory Time Series Models. J*. Timer Ser. Anal. 4, pp. 221-238, 1983.

[29] Taqqu,M., and Teverovsky, V., *A Practical Guide to Heavy Tails: Statistical Techniques Ans Applications*. http://citeseerx.ist.psu.edu, 1983.

[30] Taqqu,M., and Teverovsky, V., *A Practical Guide to Heavy Tails: Statistical Techniques Ans Application*. http://citeseerx.ist.psu.edu, 1983.

[31] Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1994), On the Self-similar Nature of Ethernet Traffic, *IEEE/ACM Trans. Networking*, Vol. 2, no. 1, pp. 1-15.

[32] Robinson, P., Log-periodogram Regression of Time Series with Long-range Dependence, *Annals of Statistics*, no. 23, pp. 1048-1072, 1995.

[33] Millán, G., and Lefranc, G., Proposición de un Estimador del Parámetro de Hurst para un Proceso Autosimilar Representativo del Grado de Aleatoriedad del Tráfico Registrado en Redes IEEE 802.3-2005. *XVIII Congreso de la Asociación Chilena de Control Automático (ACCA)*, 2007.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 148-157

Applying RBF Neural Nets for Position Control of an Inter/Scara Robot

Fernando Passold

Pontifical Catholic University of Valparaíso
College of Electrical Engineering
Avenida Brasil 2147, Valparaíso, Chile
E-mail: fernando.passold@ ucv.cl

Abstract: This paper describes experimental results applying artificial neural networks to perform the position control of a real scara manipulator robot. The general control strategy consists of a neural controller that operates in parallel with a conventional controller based on the feedback error learning architecture. The main advantage of this architecture is that it does not require any modification of the previous conventional controller algorithm. MLP and RBF neural networks trained on-line have been used, without requiring any previous knowledge about the system to be controlled. These approach has performed very successfully, with better results obtained with the RBF networks when compared to PID and sliding mode positional controllers.
Keywords: manipulator robots, position-force control, neural networks.

1 Introduction

This paper describes and discusses practical results obtained with the use of computational intelligence techniques, specifically artificial neural networks, applied to the position control of a real manipulator robot. The neural controller proposed in this work was applied in a real scara robot installed at the Industrial Automation Laboratory of the Federal University of Santa Catarina, Brazil. The work is the result of a cooperation between the Automation and Mechanical Engineering departments. The robot was manufactured by the Institute of Robotics (IfR) of the ETH (Swiss Federal Institute of Technology, *http://www.ifr.mavt.ethz.ch/*). Differently from most industrial manipulator robots, this one has an open architecture, which allows the implementation of any type of control law. The main purpose was to evaluate new algorithms for position/force control, since the robot is also equipped with a force sensor. Fig. 1 shows the Inter/Scara robot.

Manipulator robots are a type of non-linear and time variant systems. Conventional controllers, such as PD and PID, are used among other advanced and robust controllers that require some knowledge about the dynamic model of the system under control. In the case of manipulator robots, it is difficult to obtain some parameters, as the inertia matrix and mass centers of any joint, with sufficient accuracy. Therefore, either adaptive controller are required to overcome these inaccurate parameters or control laws based on Lyapunov functions are developed to guarantee some kind of stability [1]. Both approaches require some knowledge about the system. This work explores a computational intelligence technique, namely artificial neural networks, to deal with this situation. In particular, control algorithms based on neural networks and fuzzy logic techniques are considered intelligent control approaches that do not require any previous knowledge about the system to be controlled. The main goal of this work is to explore an effective computational intelligence technique to control this complex system using a structure as simple as possible, with preference to one that requires less change in the previous conventional controller that has installed in the system, do not overload the main processor and is robustness against disturbance and load effect variations.

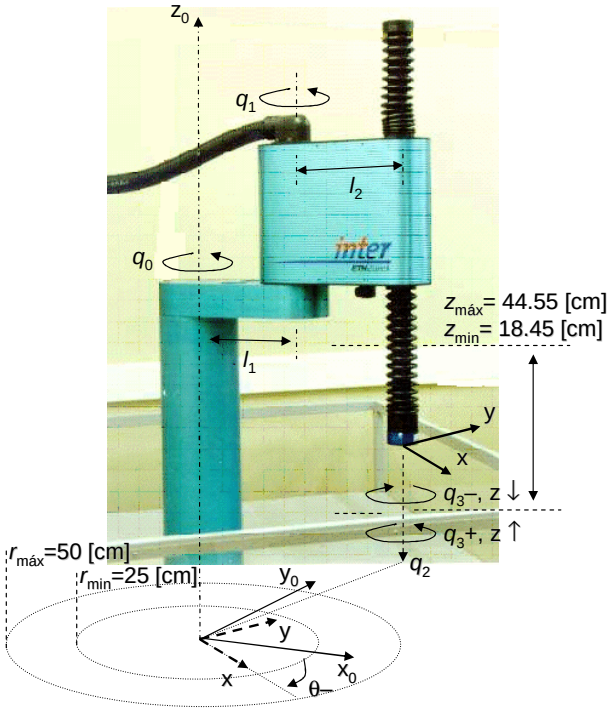


Figure 1: Inter/Scara robot and its coordinate system.

2 Neural Controller Types

Artificial Neural Networks (NNs) have been applied to several cases of control systems, showing special adequacy when we have to deal with complexity, non-linearity or uncertainty [13]. The neural approach is interesting, notably in the cases where:

a) Mathematical models of the process are poor or do not even exist and linear models are inadequate to represent the system with sufficient accuracy;

b) The system works satisfactorily with an existing controller but its performance deteriorates substantially when high performances (e.g. speeds) are required, so non-linear controllers become necessary.

NNs have proved their ability to approximate arbitrary nonlinear maps and the availability of methods for adjusting their parameters on basis of input and output data makes them particularly attractive in identifiers and controllers [2]. Narendra also comments that it seems to be valuable to keep linear and non-linear components working in parallel, where the neural networks represent the non-linear components [2]. He also mentions the brief learning time periods and the increase of accuracy that result from this kind of integration. In the control systems area, a few neural models have been proved to be more suitable than others, mainly:

1) Multilayer Perceptron networks (MLP), and;

2) Radial Base Function networks (RBF).

Among several ways to apply NNs in a control scheme, we can cite: (i) inverse identification (requires undesirable off-line training), (ii) reference model structure, (iii) internal model control, (iv) predictive control (uses two NNs, one of them trained off-line to identify the system), (v) optimal control (also requires two nets: the first one is used to quantify the state space of the system and the next acts as a

classifier). Katič and Vukobratovič [3] and Morris and Khemaissia [4] discuss two learning architectures that seem to be the most appropriate and promising: a) Feedback-error learning architecture; and b) Adaptive learning architecture. The feedback error learning approach is characterized by the NN inserted in the feedback path to capture the nonlinear characteristics of the system. The ANN weights are tuned on-line with no off-line learning phase and, when compared with the adaptive technique, we do not require any knowledge of the robot dynamics, linearity in the unknown system parameters or the tedious computation of a regression matrix. Thus, this approach is model-free and represents an improvement over adaptive techniques [5].

2.1 Multilayer Perceptron Net

This network consists of a set of sensory input units (source nodes) that constitute the input layer, one or more hidden layers and an output layer. The input signal propagates through the network in a forward direction, on a layer-by-layer basis [6]. Multilayer perceptrons could be trained in a supervised manner using the back-propagation algorithm. Back-propagation is based on the error-correction learning rule and uses a least-mean-square error algorithm to adjust its connection weights. The error back-propagation learning consists of two stages: first, a forward phase, when the input vector is applied to the sensory nodes of the network, and its effect propagates through the network layer by layer. Finally, an output set is produced as the current response of the network. During the forward phase the weights of the network are kept unchanged. In the second phase, the backward phase, an error signal is propagated backward through the network against the direction of synaptic connections and the weights are adjusted to make the current response of the network move closer to the desired response based on a steepest descent algorithm, or back propagation weight update rule, also called generalized delta-rule [6].

2.2 Radial Base Functions Net

The basic structure of the RBF network consists of three layers. Different from the MLP networks, the layers here have different tasks. The first layer is passive and only connects the model to the real world. The second layer is a unique hidden layer. It performs a non-linear transformation from the input vector space to the internal vector space. The last layer is the output layer and transforms internal vector space into output vector space in a linear manner. There are several algorithms available to train the network [6, 7, 8]. These two types of neural nets can be universal function approximators [6, 8].

3 Controller Proposed

The controller proposed here uses a conventional PD or PID that performs in parallel with an artificial neural network trained on-line. This kind of architecture for neural controllers is known as feedback error learning because the net uses the output signal generated by the conventional controller as its own error signal that is back propagated for learning purposes ([2, 5]). Fig. 2 shows the architecture applied in this case.

Both types of NNs tested deal with the same input data vector: $x_{NN} = \left[\begin{array}{ccc} q & \dot{q} & \ddot{q}_d \end{array} \right]$, where q is the vector of current joint positions; $\dot{q}$ is the vector of current joint speeds (obtained through numerical differentiation); $\ddot{q}_d$ corresponds to the desired joint accelerations (like in other manipulators, there is no accelerometer available for each joint, so the desired acceleration computed by the path generator was used). These inputs were bounded into its maximum and minimum possible operational values for this robot and then scaled between the neural input range -0.9 up to +0.9 only for the MLP nets. Note that, different from MLP nets, the RBF nets do not need a scale procedure sice that they could deal directly with rough data, but it was necessary to organize the input data into three different classes: 1) joint positions, 2) joint speeds and 3) joint accelerations – see fig. 3. The idea behind organizing the data

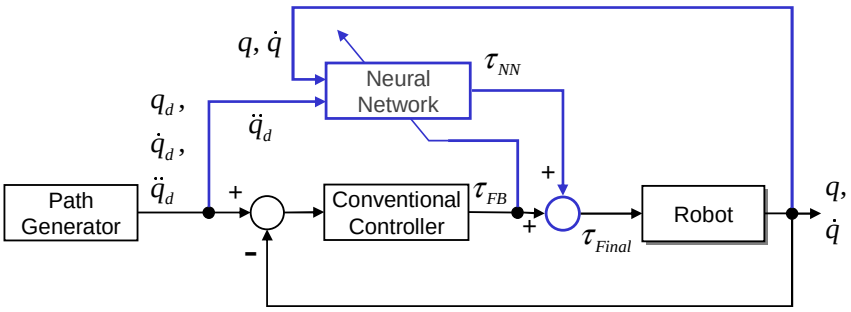


Figure 2: Feedback-error-learning neural controller used.

in different classes comes from previous skills with fuzzy inference systems. RBF networks could be compared with fuzzy inference systems [9]. Each class of input data could be understood as linguistic operations of fuzzy systems. Each class of data is mapped using m Gaussian functions and could be compared to the m membership functions that will be used in a fuzzy system. And finally, the rules and the way they are evaluated in a fuzzy system were performed by the output layer of a RBF network. Each synaptic connection of the output layer could be compared to the fuzzy IF-THEN-rules. The overall outputs are derived from the weighted sum done by the output layer [10]. Hence, m Gaussian functions were created to categorize each vector of each class of input data, as can be seen in fig. 3. It could be argumented that the massive amount of data required for the input layer (3 classes $\times$ 4 d.o.f. $\times$ m Gaussian functions) is a drawback of this approach, but this solution was related to the final application in mind in this case. Motions in the plane XY were done by the first two joints of the robot. Height (Z) and final orientation (θ) is performed by the last two joints of this robot but there is a mechanical coupling between them (a ball-screw-spline system), i.e., changes only in the final orientation of the robot result in a small change in the final height reached by the robot. That, represents an extra challenge to develop an effective controller to this kind of robot.

The centers x_i of the m desired Gaussian functions are fixed, based on the range (minimal and maximal values) of the input vector. That, allows defining the maximum Euclidian distance, d_{max}, between each Gaussian centers as: $d_{max} = (x_{max} - x_{min})/(m-1)$ and then fixing the standard deviation (or spread) of each Gaussian function to be used according to:

$$\sigma = \frac{d_{max}}{\sqrt{2m}} \tag{1}$$

The traditional back-propagation algorithm expanded with momentum term was used to adjust in real-time the weights of the MLP and RBF networks [6]. The addition of the momentum term to the delta rule traditionally used to update the weights of the net (based on the method of the descending gradient of the error signal), speeds up this algorithm, it eliminates decurrent oscillations of the calculation of the gradient and prevents the net to get paralyzed into a point of minimum local (and not global) in its surface errors [6]. Both networks end with 4 neurons, each one to evaluate the torque needed to command each joint motor of the robot. The final torque sent to each joint is defined as:

$$\tau_{Final} = \tau_{FB} + \tau_{NN} \tag{2}$$

where τ_{FB} is related to the torque evaluated by the conventional feedback controller that performs in parallel with one of these networks. PD and PID controllers working in the joint space have been used. The equation for the PID used is given by:

$$\tau_{FB} = \hat{B}(q)\left[K_p\tilde{q} + K_i\int \tilde{q}\,dt + K_d\dot{\tilde{q}}\right] \tag{3}$$

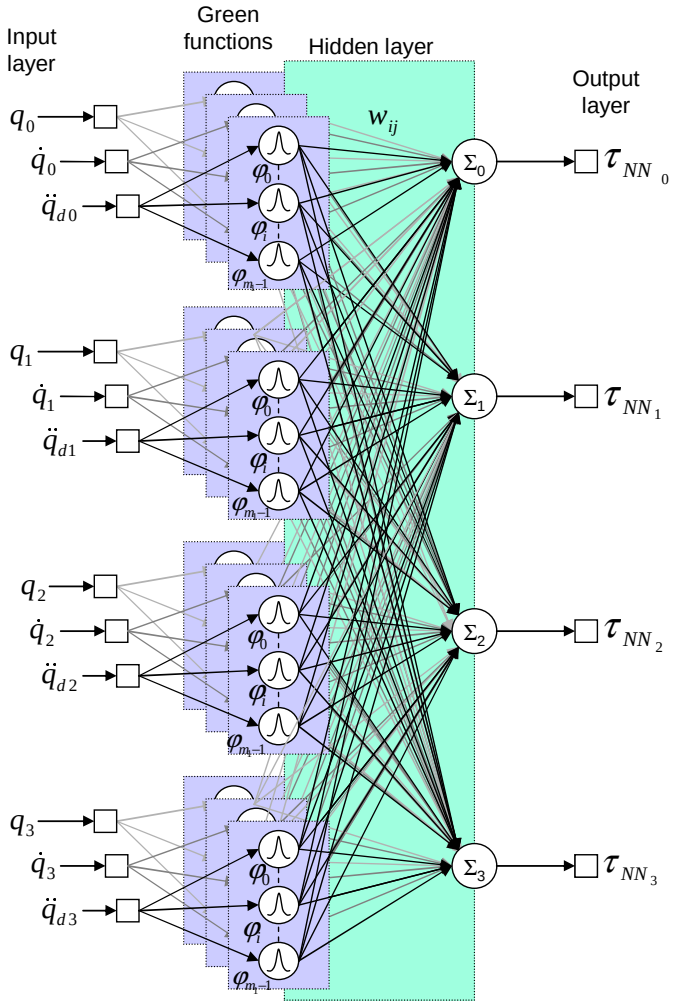


Figure 3: Structure of the RBF net implemented.

where $\hat{B}(q)$ refers to the inertia matrix of the robot (estimated); $\tilde{q} = q_d - q$ represents the error between the desired and the actual joint position; $\dot{\tilde{q}}$ refers to the velocity error; K_p is the vector of proportional gains for each joint; K_i refers to the integral vector gains and K_d is the derivative gain vector for each joint. To get a PD action over the system, the K_i vector was not used (equal to zero).

4 Experimental Results

The proposed controllers were implemented to the Inter/Scara robot (fig 1). The first two links of the Inter/Scara robot acts like an XY planar robot, and each one has 0.5 meters of length and its mechanical transmissions use harmonic drives (HD) to reduce motor-joints frictions at a minimal level. The last two joints use a ball-screw-spline mechanical scheme that allows movement in Z direction and the definition of the final orientation of these robot (angle θ). This robot could be manipulated in the Z direction between 18.45 (cm) and 44.55 (cm). Tab. 1 shows the Denavit-Hartenberg parameters among others of this robot.

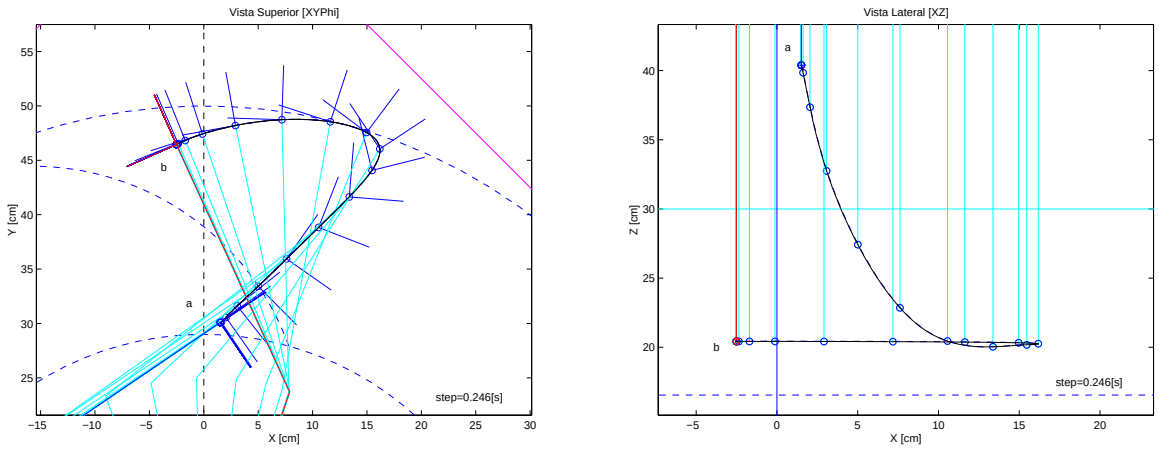
The Inter/Scara uses the XO/2 real-time object oriented operational system (*www.XO2.org*) developed by the Institute of Informatics of the Swiss Federal Institute of Technology (ETH). It includes a high level object oriented programming language called Oberon which is a kind of successor of Modula 2 and Pascal (see *http://www.oberon.ethz.ch.* The robot runs a XO/2 version over a PowerPC 200 MHz

Table 1: Denavit-Hartenberg parameters of Inter/Scara robot.

Joint	α_i	a_i	d_i	q_i	m_{l_i}	τ_{max_i}
0	0	250	665	q_0	$\cong 6.3$	333.0
1	180^o	250	0	q_1	$\cong 19.5$	157.0
2	0	0	q_2	0	*	877.0
3	0	0	0	q_3	*	16.7
	(degrees)	(mm)	(mm)		(Kg)	(Nm)

Note: m_{l_1} includes the last 2 joints(*).

equipped with 16 Mbytes of memory, which communicates with its I/O devices using an industrial VME bus (67 MHz). The user interface is done through a TCP/IP connection with a PC running the Oberon System 3 for Windows Win32 2.3 (http://www.oberon.ethz.ch/) over the Windows 2000 Pro. The user develops the whole control system of the robot (including text command interface with the user, initialization and security functions) and through a cross compiler the execution programm is downloaded to the CPU of this robot (up to 4Mbytes of code). The algorithms for the proposed neural controllers were implemented as a real-time task running within a sampling rate of 1 millisecond. In each millisecond is evaluated the action of conventional controller, also it is evaluated the forward phase of the neural net and still the backward phase when the trajectory error, $\tilde{q} > 0.0001(\text{rad}|\text{m})$. The controller was tested over the trajectory shown by fig. 4. Table 2 shows the joint positions, speeds and accelerations developed for each joint. All the four joints were moved simultaneously.



(a) XYθ view. (b) XZ view

Figure 4: Trajectory used for the tests.

The PD/PID controller were digitally implemented using a 1(ms) sampling rate, and also an anti wind-up scheme was introduced with it to limit the integrative values up to 10000. Table 3 shows the parameters used for the tested PD/PID controllers. Their gains was settled initially based on Ziegler-Nichols parameters and then better adjusted using trial-and-error method.

Fig. 5 shows the output torques developed by different controllers for the joint 3 (the last and faster). Note the different performances developed by the PD+MLP1c (MLP with one single hidden layer), PD+MLP2c (MLP with two hidden layers), PD+RBF(5) (RBF with 5 Gaussian functions) and PID+RBF controllers tested.

Table 2: Parameters of the trajectory used for the tests.

Parameter	Joint 0	Joint 1	Joint 2	Joint 3
q_a	2.45	-1.85	-0.25	-1.57
q_b	1.25	0.75	-0.46	0.00
	[rad]	[rad]	[m]	[rad]
$\dot{q}_{máx}$	-0.71	1.19	-0.22	0.85
	[rad/s]	[rad/s]	[m/s]	[rad/s]
$\ddot{q}_{máx}$	0.85	1.00	0.47	-0.93
	[rad/s^2]	[rad/s^2]	[m/s^2]	[rad/s^2]

Table 3: PD/PID gains used.

	Joint 0	Joint 1	Joint 2	Joint 3
K_p	4900	12100	90000	14400
K_d	140	220	600	240
K_i	478	1200	9200	1410

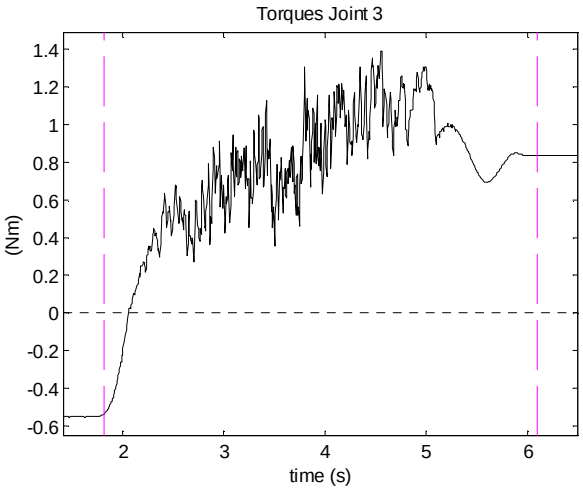
Fig. 6 shows the trajectory error during all the time. During the initial one-third of the robot configuration change period the NNs are in their learning time and the conventional controller still predominates in the joint control. But even before the end of this period of time, it could be seen that the NNs output torque takes predominance over the final torque evaluated (fig. 6). It could be seen that the NN learns the dynamic behavior of the system and then does the dynamic compensation that results in higher performance compared to a conventional controller.

The best results for the MLP nets were achieved with learning rate $\eta = 0.035$ and momentum term $\alpha = 0.5$. Related to the RBF nets, the best learning parameters founded were: $\eta = 0.005$ and $\alpha = 0.5$. Note that a PID performing with a RBF net, allows the better performance (fig. 5(d) and 5(e)) followed by the PD+RBF, PD+MLP and finally, the single PD. It was noted that the use of two hidden layers for the MLP NN does not imply in a better performance and moreover adds a small residual memory effect in presence of a disturbance on the system (this behavior has been observed in tests where an elastic string was placed in the middle of a linear trajectory).

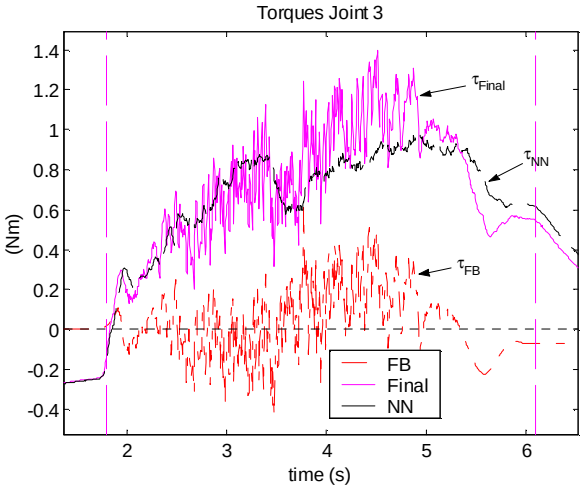
Experimental results have also demonstrated that the addition of more than 5 Gaussian functions in the RBF NN controller, slightly increase the performance, but at the expenses of a significant higher computational cost.

5 Conclusions

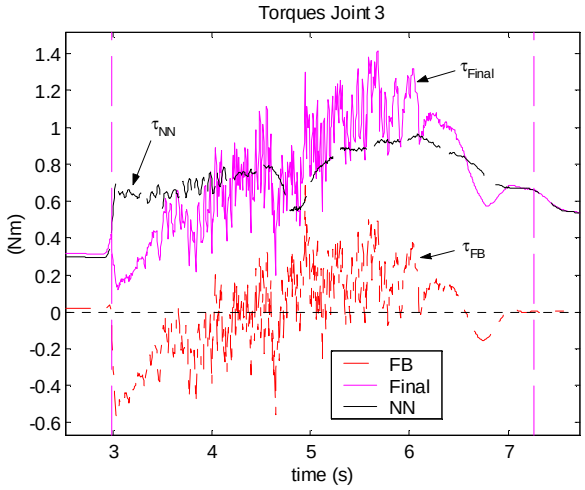
This paper has presented a practical and successful application of a neural controller performing in parallel with a conventional controller in the position and trajectory following control of a real robot. The use of a conventional controller performing in parallel with the NNs is advantageous to maintain the robustness of the system when the NN become saturated (due to high learning rates) and it is important to force the readjustment of the synaptic weights of the NN used when the robot changes its configuration. As soon as the NN captures the dynamic behavior of the system, the final torque is given quite totally by the NN and a higher performance could be achieved. Both the MLP and RBF ANN perform very well,



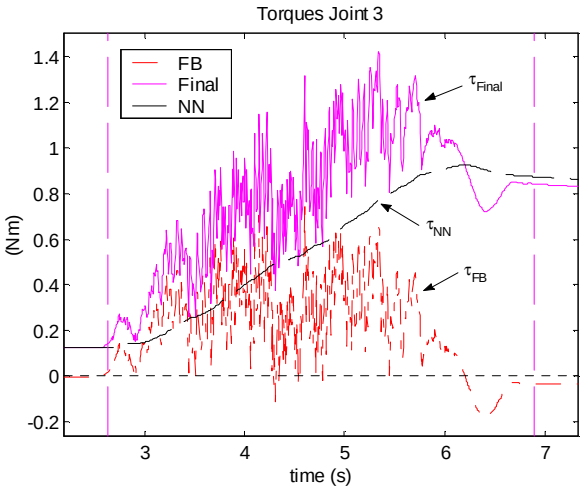
(a) PD controller output torques.



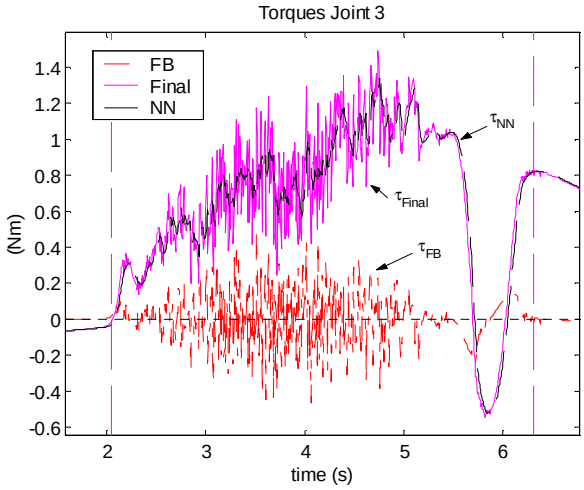
(b) PD + MLP1c controller outputs torques.



(c) PD + MLP2c controller outputs torques.



(d) d) PD + RBF(5) controller outputs torques.



(e) PID + RBF(5) controller outputs torques.

Figure 5: Controllers output torques.

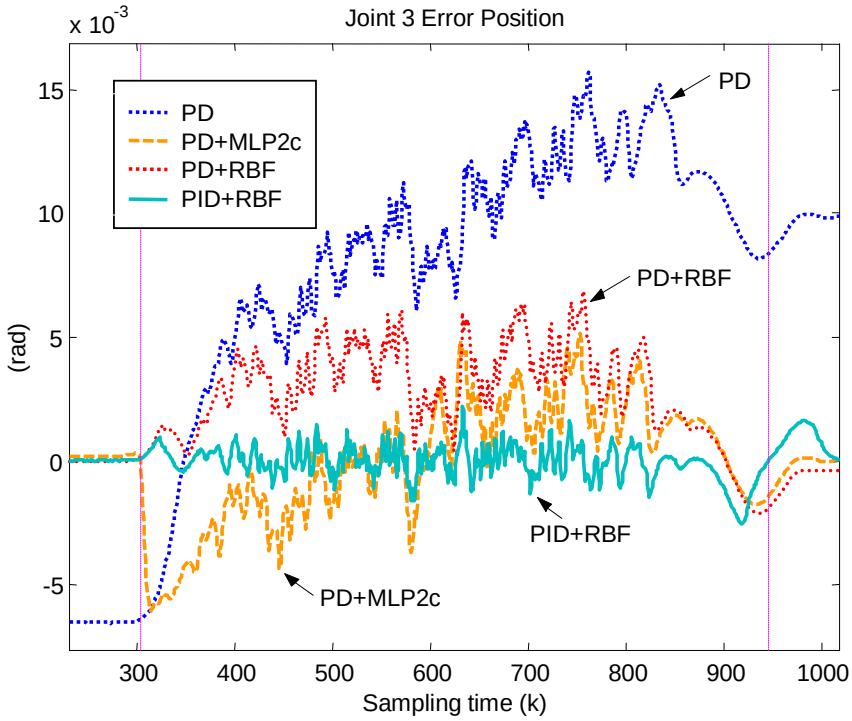


Figure 6: Joint 3 some tracking error.

but RBF does it better and faster than a MLP, mainly if it works in parallel with a PID controller. An additional and unexpected advantage could be achieved with the PID+RBF controller: robot motion with lowest noise levels (quite silent). On the other hand, there is also a drawback: NN requires more processing power to work in parallel with the conventional controller – see table 4 and fig 7 that summarize the results achieved by the different controllers.

Table 4: Power computer resources required.

Controller	Processing time	
	Min	Max
PD	104	104
PD + MLP2c	194	385
PD + RBF(5)	333	579
PD + RBF(7)	425	679
PD + RBF(9)	104	809

Note: values expressed in microseconds (µs).

Even if computer resources are short, a simple PD+MLP with one hidden layer allows better results than a PD controller. Otherwise, if higher processing power is available, a PID+RBF achieves the best results. Since the neural controllers proposed here have performed very successfully, future directions of this works intend to establish an integrated position/force control over a hybrid control architecture to deal with robot applications that imply some contact with the environment.

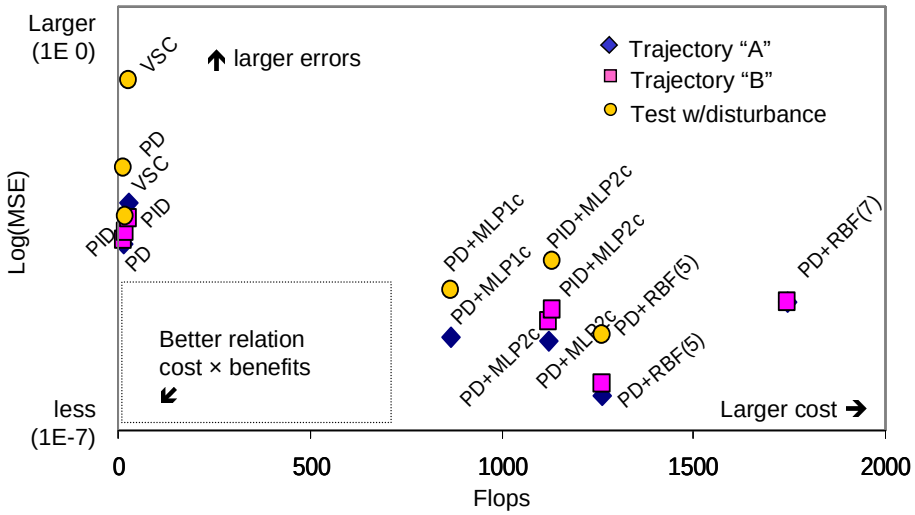


Figure 7: Relationship within computational cost × benefit of the controllers tested here.

Bibliography

[1] Sciavicco, L. and Siciliano, B., *Modeling and Control of Robot Manipulators*, McGraw-Hill, 1996.

[2] Narendra, K.S., Neural networks for real-time control. In *36th IEEE Conference on Decision and Control - CDC'97*, 1026–1031, San Diego, California, USA, 1997.

[3] Katič, D. and Vukobratovič, M., Connectionist based robot control: an overview. In *13th IFAC*, volume 1b-05 6, 169–174. San Francisco, USA, 1996.

[4] Morris, A.S. and Khemaissia, S., Artificial neural network based intelligent robot dynamic control. In A.M.S. Zalzala and A.S. Morris (eds.), *Neural Networks for Robotic Control – Theory and Applications*, chapter 2, 26–63, Ellis Horwood, Great Britain. 1996.

[5] Kim, Y.H. and Lewis, F.L., Neural network output feedback control of robot manipulators. *IEEE Transaction on Robotics and Automation*, 15(2), 301–309, 1999.

[6] Haykin, S., *Neural Networks A Comprehensive Foundation*, Prentice Hall, New Jersey, USA, 2nd edition, 1999.

[7] Gabrijel, I. and Dobnikar, A., Adaptative RBF neural network. In *SOCO'97 Conference*, 164–170. Nimes, France. URL `http://cherry.fer.uni-lj.si:80/~gabriel/soco97/soco97.zip`, 1997.

[8] Girosi, F. and Poggio, T., Networks and the best approximation property. In M.M. Gupta and D.H. Rao (eds.), *Neuro-Control Systems, Theory and Applications*, 257–264. IEEE Pres, Piscataway, New Jersey, USA, 1993.

[9] Fritzke, B., Incremental neuro-fuzzy systems. In *Applications of soft computing, SPIE International Symposium on Optical Science, Engineering and Instrumentation*, San Diego, 1997.

[10] Kiguchi, K. and Fukuda, T., Intelligent position/force controller for industrial robot manipulators – application of fuzzy neural networks, *IEEE Transactions on Industrial Electronics*, 44(6), 753–761, 1997.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 158-166

Task Resource Allocation in Grid using Swift Scheduler

K. Somasundaram, S. Radhakrishnan

Arulmigul Kalasalingam College of Engineering
Department of Computer Science and Engineering
Krishnankoil-626190, Tamilnadu, India
E-mail: soms72@yahoo.com

Abstract:
In nature, Grid computing is the combination of parallel and distributed computing where running computationally intensive applications like sequence alignment, weather forecasting, etc are needed a proficient scheduler to solve the problems awfully fast. Most of the Grid tasks are scheduled based on the First come first served (FCFS) or FCFS with advanced reservation, Shortest Job First (SJF) and etc. But these traditional algorithms seize more computational time due to soar waiting time of jobs in job queue. In Grid scheduling algorithm, the resources selection is NP-complete. To triumph over the above problem, we proposed a new dynamic scheduling algorithm which is the combination of heuristic search algorithm and traditional SJF algorithm called swift scheduler. The proposed algorithm takes care of Job's memory and CPU requirements along with the priority of jobs and resources. Our experimental results shows that our scheduler reduces the average waiting time in the job queue and reduces the over all computational time.
Keywords: Grid Computing, Swift Scheduler, Dynamic Scheduling Algorithm, First Come First Serve, Shortest Job First.

1 Introduction:

A computational Grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities. According to the function, Grid is classified into three types: Computing Grid, Data Grid, and Service Grid. Computing Grid is used to connect varied computing resource on the network to construct a virtual high performance computer, which could offer high performance computer [1]. The traditional computing Grid systems involve many technologies such as certification, task scheduling, communication protocols, fault tolerance and so on. The task of Grid resource broker and scheduler is to dynamically identify and characterize the available resources and to select and allocate the most appropriate resources for a given job [2]. The resources are typically heterogeneous locally administered and accessible under different local policies. Advance reservation [3] is currently being added to Portable Batch System (PBS).

In a Grid Scheduler, the mapping of Grid resources and an independent job in optimized manner is so hard where we couldn't predict optimized mapping. So the combination of uninformed search and informed search will provide the good optimal solution for mapping a resources and jobs, to provide minimal turn around time with minimal cost and minimize the average waiting time of the jobs in the queue. A heuristic algorithm is an algorithm that ignores whether the solution to the problem can be proven to be correct, but which usually produces a good solution. Heuristics are typically used when there is no known way to find an optimal solution, or when it is desirable to give up finding the optimal solution for an improvement in run time.

The primary objective of this research is to investigate effective resource allocation techniques based on computational economy through simulation. We like to simulate millions of resources and thousands of users with varied requirements and study scalability of systems, algorithms, efficiency of resource allocation policies and satisfaction of users. In our simulation we would like to model applications in the areas of biotechnology, astrophysics, network design, and high-energy physics in order to study usefulness of our resource allocation techniques. The results of our work will have significant impact on the way resource allocation is performed for solving problems on grid computing systems.

The organization of this paper is as follows. In Section 2, the related works are discussed. In section 3, we introduce our scheduling algorithm model. In section 4 we present and discuss the experimental results. We conclude this study in section 5.

2 Related Work:

Job scheduling in parallel system has been extensively researched in the past. Typically this research has focused on allocating a single resource type (e.g., CPU usage) to jobs in the ready queue. The use of many of these scheduling algorithms has been limited due to restriction in application designs, runtime system, or the job management system itself. Therefore simple allocation scheme such as first come First Serve (FCFS) or FCFS with first fit back fill (FCFS/FF) are used in practice [4].

Current job scheduling practices typically support variable resource allocation to a job, and run to completion scheduling. Scheduling policies are also heavily based on First-come-First-serve (FCFS) methods [5]. A FCFS scheduling algorithm allocates resources to jobs in the order that they arrive. The FCFS algorithm schedules the next job in ready queue as soon as sufficient system resources become available to meet all of the job requirements. The advantage is that this provides level of determinism on the waiting time of each job[6]. The disadvantage of FCFS shows up when the jobs at the head of the ready queue cannot be scheduled immediately due to insufficient system resources, but jobs further down the queue would be able to execute given the currently available system resources. These latter jobs are essentially blocked from executing while the system resource remains idle.

Fidanova [7] compared the simulated annealing approach with the ant algorithm for scheduling jobs in Grid. David Beasley, Marek Mika and Grzegorz Waligora [8] formulated the scheduling problem as a linear programming problem and proposed local search meta-heuristic to schedule workflow jobs on a Grid. Fair Share scheduling [12] is compared with simple fare task order scheduling, adjusted fair task order scheduling and Max-min fair share scheduling algorithm are developed and tested with existing scheduling algorithms.

Rafael A. Moreno [9] addresses the issues that the resource broker has to tackle like resource discovery and selection, job scheduling, job monitoring and migration etc. Resource Management System [10, 11] was discussed and models of grid RMS availability by considering both the failures of Resource Management (RM) Servers and the length limitation of request queues. The resource management systems (RMS) can divide service tasks into execution blocks (EB), and send these blocks to different resources. To provide a desired level of service reliability, the RMS assigns the same EB to several independent resources for parallel (redundant) execution.

3 Swift Scheduler (SS) Model:

Let N be the number of jobs in Job Queue 'J_q' which is indicated as,

$$J_q = \{J_1, J_2, J_3, ..., J_N\} \tag{1}$$

Jobs are allotted to M number of resources in Resource Queue'R_q'which is indicated as,

$$R_q = \{R_1, R_2, R_3, ..., R_M\} \tag{2}$$

Let $F(J_i, R_j)$ be the overall job completion time for the i^{th} job in j^{th} resources can be calculated as,

$$\sum_{i=0}^{N}\sum_{j=0}^{M} F(J_i, R_j) = \sum_{i=0}^{N}\sum_{j=0}^{M} G(J_i, R_j) + \sum_{i=0}^{N}\sum_{j=0}^{M} H(J_i, R_j) \tag{3}$$

Let $G(J_i, R_j)$ be the expected job completion time of the i^{th} job in j^{th} resources which can be calculated as,

$$\sum_{i=0}^{N}\sum_{j=0}^{M} G(J_i, R_j) = \sum_{i=0}^{N}\sum_{j=0}^{M} (JL_i/RC_j) \tag{4}$$

JL_i be the Job length of i^{th} Jobs and RC_j be the capacity of the j^{th} resources.
Let $H(J_i, R_j)$ be the heuristic function of the i^{th} job in j^{th} resources which can be calculated as

$$\sum_{i=0}^{N}\sum_{j=0}^{M} H(J_i, R_j) = \sum_{i=0}^{N}\sum_{j=0}^{M} (JL_i/RC_j) + Communicationoverhead \tag{5}$$

3.1 Working Principle of Swift Scheduler (SS):

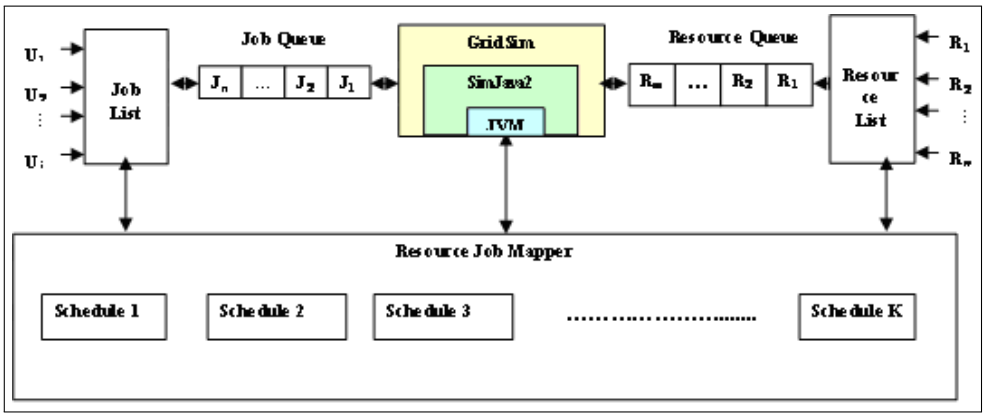


Figure 1: Architecture of SS

Figure 1 shows working principles and architecture of our proposed algorithm SS as follows: incoming jobs from different users are collected and stored in job list and available Resources are stored in resource list. Jobs are randomly arrived to job queue as well as resources are selected based on the availability. The swift scheduler in GridSim [13] maps the jobs from job queue and resources from resource queue where the resources are selected using the heuristics function. The function will select the optimized resource for the particular job to complete it with minimum time.

3.2 Pseudo Code for Swift Scheduler (SS):

```
  procedure SwiftScheduler()
begin:
   initialize job_queue, resource_queue;
   Loop exec for 'N' jobs
   begin:
        initialize job_servicetime;
        addJobs(N) to job_queue;
    end;
    Loop exec for 'M' Resources
    begin:
        addResources(R) to resource_queue;
    end;
    arrange jobs in ascending order based on job length and maintained in
    jobsQueue;
    Loop exec for 'N' jobs
    begin:
      Loop exec for 'M' resources
      begin:
       calculate the processing time of 'N' jobs in 'M' resources;
      end;
     end;
     Loop exec for 'N' jobs
 begin:
    Loop exec for 'M' resources
    begin:
      select the lowest processing time resource for the each jobs using
      heuristic function;
      allocate(n,m);
    end;
 end;
end;
```

4 Performance Analysis:

In this section, we analyze the performance of Swift Scheduler with existing Simple Fair Task Order Scheduling (SFTO) against large set of independent jobs with varying size and large number of heterogeneous resources. Assume, the arrival rates of jobs are based on the Poisson distribution. The following Fig.(2) and Fig.(3) shows the job allocation methods used in SFTO and SS respectively and the following Table (1) and Table (2) shows the arrival order of the particular jobs and at what time, the jobs will start its execution in the particular resource and its service time of the jobs in the particular resource where the selection of resources are based on the algorithms.

For example, In SFTO, the jobs J0 and J3 are allotted in resource R1. The residing time (T_r) of jobs J0 is the combination of jobs J0 waiting time (T_w) in queue and service time (T_s) of J0 (i.e $T_r = T_w + T_s$ = 2820.03ms + 61.35ms = 2881.38ms). Similarly, the residing time of job J3 is 1490 ms. The job J2 is allotted in resource R2 where the residing time of J2 is 1257.19 ms and jobs J4 and J1 are chosen

JobID	Resource Name	Start time in ms	Residing time in ms
4	R3	1851.12	1886.86
1	R3	3337.99	3391.73
2	R2	1214.79	1257.19
3	R1	1330.66	1490.66
0	R1	2820.03	2881.38

Table 1: Job, Resource, start and residing time for SFTO

JobID	Resource Name	Start time in ms	Residing time in ms
2	R1	1210.02	1316.02
0	R1	2805.39	2980.39
3	R2	1093.04	1178.27
1	R2	2688.17	2772.17
4	R3	1453.51	1489.25

Table 2: Job, Resource, start and residing time for SS

resources R3 where the residing time of J4 is 1886.86 ms and J1 is 3391.73 ms. The average waiting time of all jobs in grid system is 2181.56ms.

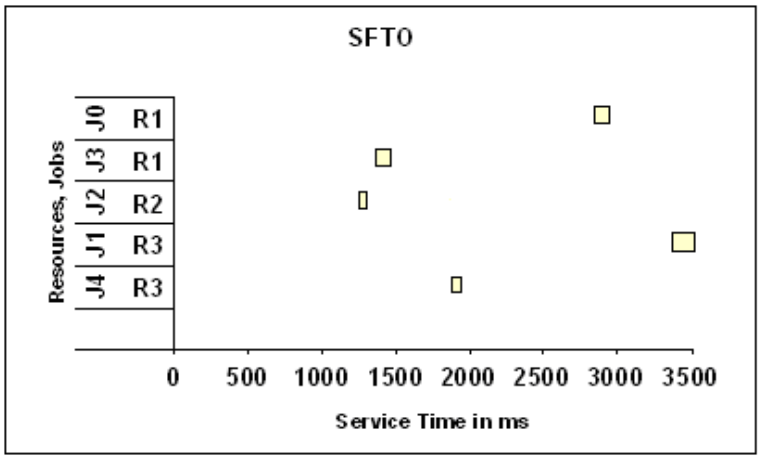


Figure 2: Job and Resource Allocation for SFTO

In SS, resource selection and jobs allocation are based on heuristic searching algorithm on SJF, which reduces the average waiting time of the jobs in queue. So, overall turn around time is reduced and resource utilization is increased. For example, in SS, the jobs J2 and J0 are allotted to resource R1 where its residing time is 1316.02 ms and 2980.39 ms respectively. Similarly, Jobs J3, J1 and J4 are chosen by resources R2, R2 and R3 respectively where residing time of Jobs J3,J1 and J4 are 1178.29 ms, 2772.17 ms and 1489.25 ms respectively. The average waiting time of all jobs in grid system is 1947.22 ms which is less than SFTO. The statistical data presented here is acquired by averaging the scheduling performance over different runs. The following figs and tables shows the cost based, total processing time and resource utilization based comparison test results of FCFS, SJF, SFTO and SS against varying number of jobs and resources.

Our proposed algorithm Swift Scheduler (SS) is compared with FCFS, SJF and Simple Fare Task Order (SFTO) scheduler. We have tested SS in GridSim by varying number of resources, no. of jobs

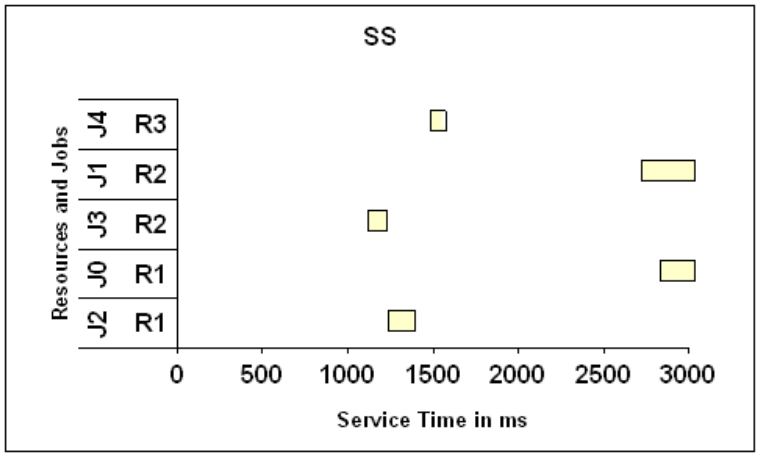


Figure 3: : Job and Resource Allocation for SS

against total processing time, cost, and resource utilization. We can vary the number of resources like 5, 10, 15, 20,... etc. For experimental purposes, two sample simulation results are shown in Figure 4, 5, 6, 7, 8 & 9. By analyzing the obtained results from the simulator, Swift Scheduler completed all jobs with minimum time and cost by utilizing maximum amount of resources compares with other scheduler like FCFS, SJF and SFTO.

Total Processing Time Analysis

Figure 4: : No of Jobs Vs Total Processing Time

Figure 5: : No of Jobs Vs Total Processing Time

Cost Analysis

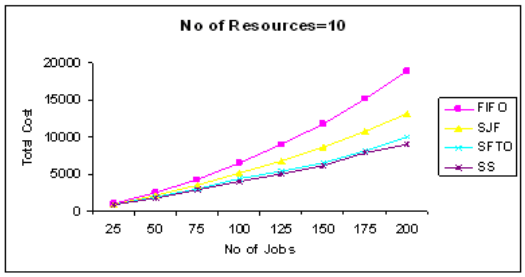


Figure 6: : No of jobs Vs Total cost

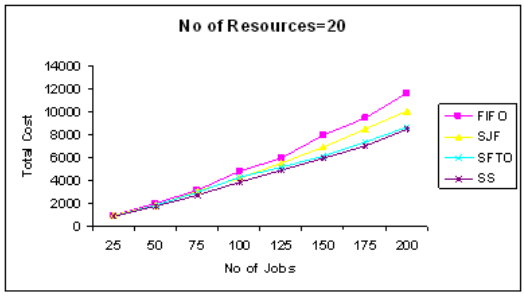


Figure 7: : No of jobs Vs Total cost

Resource Utilization

Figure 8: : No. of jobs Vs Resource Utilization

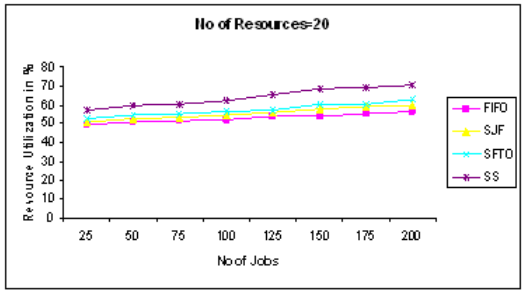


Figure 9: : No. of jobs Vs Resource Utilization

5 Conclusion and Future Work:

In this paper, we have presented the design and analyze the new scheduling algorithm Swift Scheduler. Our proposed Swift scheduler completed a task by using highly utilized low cost resources with minimum computational time. Our scheduling algorithm uses the heuristic function to select the best resources to achieve a higher throughput while maintaining the desired success rate of the job completion. This algorithm is performing better for real time job parameters and suitable for different job sizes in real environment. However, in all conditions, the proposed algorithms outperform the traditional ones. The SS policy is more effective than the FCFS, SJF and SFTO in the extent of computational complexity with lower cost but higher resources utilization. In future, we can hybrid the Swift Scheduler with any evolutionary scheduling algorithm like Genetic algorithm, Particle Swarm Optimization technique to achieve a high throughput and high resource utilization.

Bibliography

[1] Ranganathan,K, and I.Foster, "Decoupling Computation and Data Scheduling in Data Intensive Applications", *11th International Symposium on High Perfromance Distributed Computing*, Edinburgh, Scotland, Condor Project, Condor-G, 2002.

[2] Mitrani I, Palmer J," Dynamic Server Allocation Heterogenous Clusters ", *1st International working conference on Heterogeneous Networks*, Ilkley,UK, 2003.

[3] Foster,I, et al, "The Grid 2003 Production Grid : Principles and Practice", *13th International Symposium on High Performance Distributed Computing*, 2004.

[4] Vijay Subramanian, Rajkumar Kettimuthu, et al, "Distributed Job Scheduling on Computational Grids using Multiple Simultaneous Requests", *Proceedings 11th IEEE International Symposium on High Performance Distributed Computing*, 2002. HPDC-11 2002. Pages: 359- 366

[5] C.Bitten, J. Gehring, et. Al, "The NRW - Meta Computer : building block for a worldwide computational Grid", *proceeding of the 9th Heterogeneous Computing workshop*, pp.31-40, 2000.

[6] C.Ememann, V. Hamscher, et. al, "On Advantageous of Grid Computing for parallel job scheduling", *proceeding 2nd IEEE/ACM Int'l Symp. On cluster- computing and the Grid (CCGRID 2002)*, Berlin, 2002, IEEE press.

[7] Stefka Fidanova, "Simulated annealing for Grid Scheduling problem", *IEEE International Symposium on International Symposium on Modern Computing*, 2006. JVA apos;06 3-6 Oct. 2006 Page(s):41 - 45

[8] Marek Mika, Grzegorz Waligora and Jan Weglarz, A Meta-Heuristic Approach to scheduling Workflow jobs on a Grid, Grid resource management: *state of the art and future trends*, ISBN : 1-4020-7575-8 , Kluwer Academic Publishers, Norwell, MA, USA; pp : 295 - 318, 2004.

[9] Moreno, R., Alonso-Conde A.B, Job Scheduling and Resource Management Techniques in Dynamic Grid Environments In et al., F.F.R., ed.: *Across Grids 2003, Volume 2970 of Lecture Notes in computer science*, Springer, pp : 25 - 32, 2004.

[10] Yuan-Shun Dai, Min Xie and Kim-Leng Poh "Availability Modeling and Cost Optimization for the Grid Resource Management System" *IEEE Transactions on Systems, Man and Cybernetics*, Part A Volume 38, Issue 1, Jan. 2008 Page(s):170 - 179.

[11] Yuan-Shun Dai and Gregory Levitin "Optimal Resource Allocation for Maximizing Performance and Reliability in Tree-Structured Grid Services" *IEEE Transactions on Reliability* Volume 56, Issue 3, Sept. 2007 Page(s):444 - 453.

[12] Doulamis, N.D.; Doulamis, A.D.; Varvarigos, E.A.; Varvarigou, T.A "Fair Scheduling Algorithms in Grids" *IEEE Transactions on Parallel and Distributed Systems*, Volume 18, Issue 11, Nov. 2007 Page(s):1630 - 1648

[13] Anthony Sulistio, Uros Cibej, Srikumar Venugopal, Borut Robic and Rajkumar Buyya "A Toolkit for Modelling and Simulating Data Grids: An Extension to GridSim", *Concurrency and Computation: Practice and Experience (CCPE)*, Online ISSN: 1532-0634, Printed ISSN: 1532-0626, 20(13): 1591-1609, Wiley Press, New York, USA, Sep. 2008.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 167-177

Hierarchical Distributed Reasoning System for Geometric Image Generation

Nicolae Ţăndăreanu, Mihaela Verona Ghindeanu, Sergiu Andrei Nicolescu

University of Craiova, Romania
Department of Mathematics and Computer Science
A.I. Cuza St, No. 13, 200585
E-mail: ntand@rdslink.ro, mghindeanu@yahoo.com

Abstract: The concept of hierarchical reasoning system was introduced in [5], where an intuitive method to build such systems based on their inputs is given. In this paper we formalize several concepts which open a possible research line concerning the use of these structures. A hierarchical reasoning system H is a directed graph organized on several levels such that each node of the level j is a hyper-schema of order j. As a mathematical structure, H is an abstract one and a special kind of *formal computation* is introduced. As a result of this computation we obtain a set $\mathcal{F}(H)$ of formulas. We explain what we understand by an *interpretation* of H and define its corresponding *semantical computation*. By means of an interpretation $\mathcal{I}(H)$ for H and applying the rules of the semantical computation, each element of $w \in \mathcal{F}(H)$ becomes some object $\mathcal{I}(w)$ of a given space. We exemplify these concepts and we show that for two distinct interpretations $\mathcal{I}_1(H)$ and $\mathcal{I}_2(H)$ for the same system H, a given formula $w \in \mathcal{F}(H)$ is transformed into a sentence $\mathcal{I}_1(w)$ of a natural language whereas $\mathcal{I}_2(w)$ is a geometric image. A short description of a Java implementation of a hierarchical system generating images is also given in a separate section. By examples we show that the mechanism introduced in this paper allows us to model the distributed knowledge. Finally several open problems are specified.
Keywords: semantic schema, interpretation, hyper-schema, distributed reasoning system, geometrical image generation

1 Introduction

Various kinds of mechanisms for image synthesis were presented and implemented on computer. The panel of the mathematical models for this subject includes the *rewriting systems* and *graph-based models*. Picture-processing grammars ([2]), picture grammars ([3]), stochastic grammars ([14]) and L-systems are some of the rewriting systems used to process images. The L-systems are a class of string rewriting mechanism originally developed by a biologist, A. Lindenmayer, in 1968 ([7]). The original emphases were on plant topology - spatial relations between cells or larger plant modules. The L-systems are a practical tool for generating fractal forms. Today these models are applied in architecture, physiology ([1]) and music. In order to interpret the L-system as music, LMUSe system ([9]) maps any of the turtle's 3D movement, orientation directions (forward, up, and left), its drawing line length, and thickness into musical pitches, note durations and volume.

A great number of research works and practical implementations have confirmed the interest of mathematicians and computer scientists in developing and applying the methods of graph theory. These methods were applied to obtain new knowledge representation models and to process images. A very productive notion with large applications in knowledge representation is that of *conceptual graph*, a

notion introduced in literature by J.F.Sowa ([8],[10]). We can find several applications of the graph-based methods in [6] (low-level processing of digital images, learning algorithms for high-level computer vision and pattern recognition).

The concept of semantic schema was introduced in [11] as an extension of semantic networks. This structure is obtained by means of a labeled graph and a Peano algebra built over the edge labels. Since then many applications of this structure were presented (new semantics in logic programming, knowledge representation for intelligent dialog systems etc).

In [12] we defined a new mechanism for generating images similar with the edge rewriting in the way that both approaches can be used to define complex images based on some simple other images. In the mentioned paper the concept of Hierarchical Distributed Reasoning System was introduced. Each leaf of the system is given by a semantic schema. The other nodes are hyper-schemas ([12]). We presented an *intuitive method* to obtain geometrical images. The leaves represent the input of the system in semantic schemas and, by appending proper interpretations, they obtain the graphical illustrations of the received inputs. In this manner the leaves obtains the initial images. Then, at the upper levels, these images are combined by hyper-schemas to obtain complex images. We obtained a *bottom-up method* to obtain images from initiators.

In this paper we obtain the following results:

- Starting with the concept of Hierarchical Distributed and Reasoning System (HGR system) introduced in [12] in Section 3 we define a formal computation in such a structure. As a result of this computation we obtain a set $\mathcal{F}(H)$ of formulas for an arbitrary HDR system H. This is the *formal computation* in an HDR system.

- An HDR system H is an abstract structure. In Section 4 we introduce the concept of *interpretation* for H. By means of an interpretation $\mathcal{I}(H)$ for H each element of $\mathcal{F}(H)$ becomes some object of a given space. This gives the *semantical computation*. Both the formal and semantical computations are exemplified. We show that for two distinct interpretations $\mathcal{I}_1(H)$ and $\mathcal{I}_2(H)$ for the same system H we can generate sentences in a natural language giving the reasoning conclusions and geometrical images respectively.

- A short description of a Java implementation of an HDR system is also given in Section 5.

- By examples we show that the mechanism introduced in this paper allows us to model the distributed knowledge.

- The last section contains the conclusions and future works. Several open problems are specified in this section.

2 Basic concepts

Consider a symbol θ of arity 2. A θ-semantic schema ([11]) or shortly, θ-schema is a system $\mathcal{S} = (X, A_0, A, R)$, where:

- X is a finite non-empty set of symbols named object symbols;

- A_0 is a finite non-empty set of elements named label symbols and $A_0 \subseteq A \subseteq \overline{A}_0$, where $\overline{A}_0$ is the Peano θ-algebra generated by A_0;

- $R \subseteq X \times A \times X$ is a non-empty set of relations which fulfills the following conditions:

  1. $(x, \theta(u, v), y) \in R \Rightarrow \exists z \in X : (x, u, z) \in R, (z, v, y) \in R$
  2. $\theta(u, v) \in A, (x, u, z) \in R, (z, v, y) \in R \Rightarrow (x, \theta(u, v), y) \in R$

3. $\{\alpha \mid \exists(x, \alpha, y) \in R\} = A$

An element from $R \cap (X \times A_0 \times X)$ is a *regular arc* of $\mathcal{S}$.

We denote by $Ded(\mathcal{S})$ the least set satisfying the following properties ([13]):
• If $(x, a, y) \in R_0$ then $([x, y], a) \in Ded(\mathcal{S})$
• If $([x_i, \ldots, x_k], u) \in Ded(\mathcal{S})$ and $([x_k, \ldots, x_r], v) \in Ded(\mathcal{S})$, $i < k < r$ and $\theta(u, v) \in A$ then $([x_i, \ldots, x_r], \theta(u, v)) \in Ded(\mathcal{S})$.

An element of $Ded(\mathcal{S})$ is a **deductive path** of $\mathcal{S}$.

Let us consider the schemas $\mathcal{S}_1 = (X_1, A_{01}, A_1, R_1)$ and $\mathcal{S}_2 = (X_2, A_{02}, A_2, R_2)$. In the remainder of this section we describe a new structure which relieves a special kind of cooperation between $\mathcal{S}_1$ and $\mathcal{S}_2$.

If $d_1 = ([x, \ldots, y], u) \in Ded(\mathcal{S}_i)$ and $d_2 = ([y, \ldots, z], v) \in Ded(\mathcal{S}_{3-i})$, where $i \in \{1, 2\}$, then we say that d_1 is **connected to right** by d_2 or d_2 is **connected to left** by d_1. We say that d_1 is **connected** by d_2 if d_1 is connected to right or to left by d_2.

We consider the sets of deductive paths $L_1 \subseteq Ded(\mathcal{S}_1)$ and $L_2 \subseteq Ded(\mathcal{S}_2)$. We say that $L_1 \cup L_2$ is a **pairwise connected set of deductive paths** if every deductive path of L_i is connected by some deductive path of L_{3-i}.

For each $i \in \{1, 2\}$ we consider a set V_i of symbols such that $V_i \cap (A_1 \cup A_2) = \emptyset$. We consider also a set E_i such that $E_i \subseteq X_i \times V_i \times X_i$, $Card(E_i) = Card(L_i)$ and $E_1 \cap E_2 = \emptyset$. Consider also a bijective mapping $g_i : L_i \longrightarrow E_i$ such that $g_i(d) = (x, e, y)$, where $d = ([x, \ldots, y], \theta(u, v)) \in L_i$. This mapping transforms each deductive path $([x, \ldots, y], \theta(u, v))$ from L_i into a regular arc (x, e, y). Shortly, we say that the path $([x, \ldots, y], \theta(u, v))$ **is designated** by (x, e, y). We can define now a cooperating structure of hyper-schemas.

A **hyper-schema of order zero** is a semantic schema. Consider the hyper-schemas $\mathcal{S}_1$ and $\mathcal{S}_2$ of order zero. A **hyper-schema of order one** over $\mathcal{S}_1$ and $\mathcal{S}_2$ obtained by means of L_1 and L_2 is a θ-schema $\mathcal{S}$ which includes the regular arcs obtained from L_1 and L_2 ([12]). We denote by $Hyp_1(\{\mathcal{S}_1, \mathcal{S}_2\})$ the set of all hyper-schemas of first order over $\mathcal{S}_1$ and $\mathcal{S}_2$. In general we write $S \in Hyp_k(\{S_1, \mathcal{S}_2\})$ and we name S a **hyper-schema of order** k if S_1 and S_2 are hyper-schemas of order $j \leq k-1$ and at least one of them has the order $k-1$.

An HDR system ([12]) is the tuple $H = (Q_1, Q_2, \ldots, Q_k)$ where $k \geq 2$ and

- $Q_1 = \{\mathcal{S}_1, \ldots, \mathcal{S}_{n_1}\}$, $n_1 > 1$, constitutes the first level of the system. The entities $\{\mathcal{S}_1, \ldots, \mathcal{S}_{n_1}\}$ are hyper-schemas of order zero. The set Q_1 gives the leaves of H.

- $Q_2 = \{\mathcal{S}_{n_1+1}, \ldots, \mathcal{S}_{n_2}\}$, $n_2 \geq n_1 + 1$, gives the second level of the system and $\mathcal{S}_{n_1+1}, \ldots, \mathcal{S}_{n_2}$ are hyper-schemas of order 1. More precisely, for every $m \in \{n_1+1, \ldots, n_2\}$ there are $m_1, m_2 \in \{1, \ldots, n_1\}$, $m_1 \neq m_2$ such that $\mathcal{S}_m \in Hyp_1(\{\mathcal{S}_{m_1}, \mathcal{S}_{m_2}\})$.

- For $j \in \{3, \ldots, k\}$, $Q_j = \{\mathcal{S}_{n_{j-1}+1}, \ldots, \mathcal{S}_{n_j}\}$ represents the j-th level of the system, where $n_j \geq n_{j-1} + 1$. For every $m \in \{n_{j-1} + 1, \ldots, n_j\}$ there is $m_1 \in \{n_{j-2}, \ldots, n_{j-1}\}$ and there is $m_2 \in \{1, \ldots, n_{j-1}\}$ such that $\mathcal{S}_m \in Hyp_{j-1}(\{\mathcal{S}_{m_1}, \mathcal{S}_{m_2}\})$.

3 Formal computations in HDR Systems

Suppose that $H = (Q_1, Q_2, \ldots, Q_k)$ is an HDR system. The components of H are the hyper-schemas $\mathcal{S}_1, \ldots, \mathcal{S}_{n_k}$. We can visualize H as a graph structure. In order to obtain this structure we represent each hyper-schema by a node and we draw two directed arcs from $\mathcal{S}_r$ to $\mathcal{S}_j$ and to $\mathcal{S}_m$ if $\mathcal{S}_r \in Hyp_p(\{S_j, S_m\})$ for some p. The structure obtained in this manner is not a tree. This can be observed in Figure 1: there are two distinct paths from $\mathcal{S}_7$ to $\mathcal{S}_2$ and there is no root of this structure.

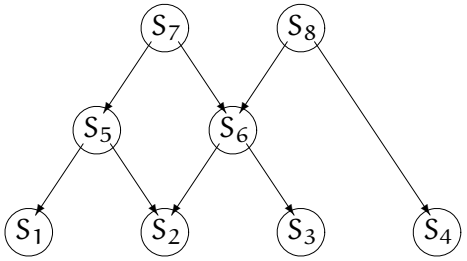


Figure 1: The graph structure of H

For each $i \in \{1, \ldots, n_k\}$ we consider that $\mathcal{S}_i$ is given by the tuple $\mathcal{S}_i = (X_i, A_{0i}, A_i, R_i)$ and we denote $R_{0i} = R_i \cap (X_i \times A_{0i} \times X_i)$. For each $r \in \{n_1 + 1, \ldots, n_k\}$ such that $\mathcal{S}_r$ is a hyper-schema over $\mathcal{S}_j$ and $\mathcal{S}_m$ in H we consider:

• the connected sets $L_{j,r} \subseteq Ded(\mathcal{S}_j)$ and $L_{m,r} \subseteq Ded(\mathcal{S}_m)$;

• the sets $E_{j,r}$, $E_{m,r}$ and the transformational mappings $g_{j,r} : L_{j,r} \longrightarrow E_{j,r}$, $g_{m,r} : L_{m,r} \longrightarrow E_{m,r}$.

By the assumptions of the previous section we have $R_{0r} \supseteq E_{j,r} \cup E_{m,r}$. We denote $N_{0r} = E_{j,r} \cup E_{m,r}$. Obviously we have the following property:

Proposition 1. $N_{0i} = \emptyset$ if and only if $\mathcal{S}_i$ is a leaf of H.

For a symbol h of arity 1 we consider the set:

$$M = \bigcup_{i=1}^{n_k} \{ h([x, y], a) \mid (x, a, y) \in R_{0i} \setminus N_{0i} \}$$

where we used the notation $h([x, y], a)$ instead of $h(([x, y], a))$.

We consider the symbols $\sigma_1, \ldots, \sigma_{n_k}$ of arity 2 and denote by $\mathcal{H}_H$ *the Peano* $\{\sigma_1, \ldots, \sigma_{n_k}\}$*-algebra generated by* M.

We consider the alphabet Z including the symbols σ_i, the elements of X_i, the elements of A_i, the left and right parentheses, the square brackets [and], the symbol h and comma. As in the theory of formal languages, the set Z^* defines all the words over Z. Because a hyper-schema is a semantic schema we have the following property:

Proposition 2. If $\mathcal{S}_i$ is a hyper-schema of H and $([x_1, \ldots, x_{k+1}], \theta(u, v)) \in Ded(\mathcal{S}_i)$ then there is r uniquely determined such that $([x_1, \ldots, x_{r+1}], u) \in Ded(\mathcal{S}_i)$ and $([x_{r+1}, \ldots, x_{k+1}], v) \in Ded(\mathcal{S}_i)$.

Definition 1. Let be $w_1, w_2 \in Z^*$. We define the following binary relation on Z^*, denoted by $\Rightarrow_H$:

• For $i \in \{1, \ldots, n_k\}$, if $(x, e, y) \in R_{0i} \setminus N_{0i}$ then $w_1([x, y], e)w_2 \Rightarrow_H w_1 h([x, y], e)w_2$;

• For $i \in \{1, \ldots, n_k\}$, if $(x, e, y) \in N_{0i}$ then $w_1([x, y], e)w_2 \Rightarrow_H w_1 d w_2$, where d is the deductive path designated by (x, e, y);

• Suppose that $([x_1, \ldots, x_{k+1}], \theta(u, v)) \in Ded(\mathcal{S}_i)$, $i \in \{1, \ldots, n(H)\}$, $([x_1, \ldots, x_{r+1}], u) \in Ded(\mathcal{S}_i)$ and $([x_{r+1}, \ldots, x_{k+1}], v) \in Ded(\mathcal{S}_i)$ then:

$w_1([x_1, \ldots, x_{k+1}], \theta(u, v))w_2 \Rightarrow_H w_1 \sigma_i(([x_1, \ldots, x_{r+1}], u), ([x_{r+1}, \ldots, x_{k+1}], v))w_2$

The reflexive and transitive closure of $\Rightarrow_H$ is denoted by $\Rightarrow_H^*$. We denote $\mathcal{F}(\mathcal{S}_i) = \{w \in \mathcal{H}(H) \mid \exists d \in Ded(\mathcal{S}_i) : d \Rightarrow_H^* w\}$ and $\mathcal{F}(H) = \bigcup_{i=1}^{n_k} \mathcal{F}(\mathcal{S}_i)$.

Let us exemplify this computation. We consider the hyper-schemas $\mathcal{S}_1$ and $\mathcal{S}_2$ of order zero from Figure 2 and the hyper-schema of order 1 from Figure 3.

If we take

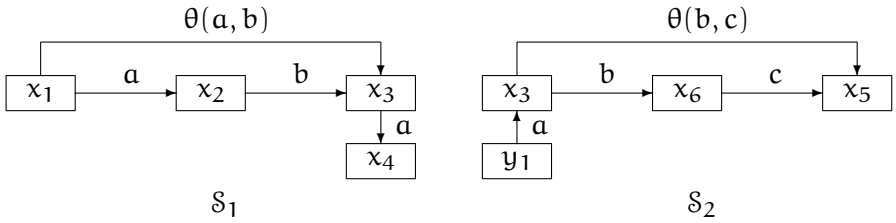


Figure 2: Semantic schemas $\mathcal{S}_1$ and $\mathcal{S}_2$ of order zero

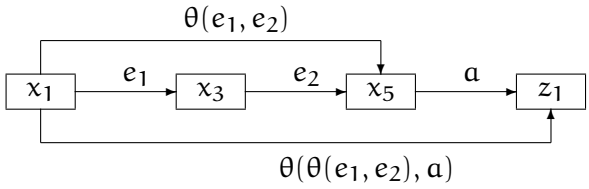


Figure 3: Hyper-schema $\mathcal{S}_3 \in Hyp_1(\{\mathcal{S}_1, \mathcal{S}_2\})$

- $L_{1,3} = \{([x_1, x_2, x_3], \theta(a, b))\}$, $L_{2,3} = \{([x_3, x_6, x_5], \theta(b, c))\}$
- $E_{1,3} = \{(x_1, e_1, x_3)\}$, $E_{2,3} = \{(x_3, e_2, x_5)\}$
- $g_{1,3}([x_1, x_2, x_3], \theta(a, b)) = (x_1, e_1, x_3)$, $g_{2,3}([x_3, x_6, x_5], \theta(b, c)) = (x_3, e_2, x_5)$

then we obtain the following computations:

- $([x_1, x_3, x_5], \theta(e_1, e_2)) \Rightarrow_H \sigma_3(([x_1, x_3], e_1), ([x_3, x_5], e_2))$
- $([x_1, x_3], e_1) \Rightarrow_H ([x_1, x_2, x_3], \theta(a, b)) \Rightarrow_H \sigma_1(([x_1, x_2], a), ([x_2, x_3], b)) \Rightarrow_H^*$ $\sigma_1(h([x_1, x_2], a), h([x_2, x_3], b)) \in \mathcal{F}(\mathcal{S}_1)$
- $([x_3, x_5], e_2) \Rightarrow_H ([x_3, x_6, x_5], \theta(b, c)) \Rightarrow_H \sigma_2(([x_3, x_6], b), ([x_6, x_5], c)) \Rightarrow_H^*$ $\sigma_2(h([x_3, x_6], b), h([x_6, x_5], c)) \in \mathcal{F}(\mathcal{S}_2)$

In conclusion,
$([x_1, x_3, x_5], \theta(e_1, e_2)) \Rightarrow_H^* \sigma_3(\sigma_1(h([x_1, x_2], a), h([x_2, x_3], b)), \sigma_2(h([x_3, x_6], b), h([x_6, x_5], c)))$
and the last formula is an element of $\mathcal{F}(H)$, where $H = (Q_1, Q_2)$, $Q_1 = \{\mathcal{S}_1, \mathcal{S}_2\}$ and $Q_2 = \{\mathcal{S}_3\}$.

4 Semantical computations in HDR systems

The semantical computation in an HDR system H transforms every formula of $\mathcal{F}(H)$ into an object of some space. In this section we describe this transformational process.
Let us consider the HDR system $H = (Q_1, Q_2, \ldots, Q_k)$ and an element $w \in \mathcal{F}(H)$. If $d = ([x_1, \ldots, x_k], \theta(u, v)) \in Ded(S_i$ and $d \Rightarrow_H^* w$ then we write $sort(w) = \theta(u, v)$.

Definition 2. An **interpretation** for H is a system $\mathcal{I} = (Ob, ob, ALG)$:
- Ob is a set of objects;
- $ob : X \longrightarrow Ob$, where $X = \bigcup_{i=1}^{n_k} X_i$, is a mapping that "interprets" each node as an object;
- $ALG = \bigcup_{i=1}^{n_k} \{Alg_u^i\}_{u \in A_i}$, where Alg_u^i is an algorithm with two input arguments and one output argument such that if $g_{j,k}([x, \ldots, y], \theta(u, v)) = (x, e, y)$ then $Alg_e^k = Alg_{\theta(u,v)}^j$.

Definition 3. The **valuation mapping** Val_H of the HDR system H is defined as follows:

- If $w = h([x, y], a) \in \mathcal{F}(H)$ then $Val_H(w) = \bigcup_{i=1}^{n_k} \{Alg_a^i(ob(x), ob(y))\}$.

- If $w = \sigma_j(w_1, w_2) \in \mathcal{F}(H)$, $w_1 \in \mathcal{F}(H)$, $w_2 \in \mathcal{F}(H)$ and $sort(w) = \alpha$ then

$$Val_H(w) = \{Alg^j_\alpha(o_1, o_2) \mid o_k \in Val_H(w_k), k = 1, 2\}$$

In order to exemplify the computations we consider again the HDR system H from Section 3. We define an interpretation of H by means of some *sentential forms*. Such a structure is a sentence containing two variables. If we substitute each variable by an object then a sentential form becomes a sentence in a natural language. We shall consider the following sentential forms:

$p_1(x, y)$="x is the father of y"; $p_2(x, y)$="x is the mother of y";

$p_3(x, y)$="x is the brother of a y"; $p_4(x, y)$="x likes to eat y";

$q_1(x, y)$="x is the grandmother of y"; $q_2(x, y)$="a brother of x likes to eat y";

$r(x, y)$="A nephew of x likes to eat y";

We consider the following algorithms:

Algorithm $Alg^1_a(o_1, o_2)$ { return $p_1(o_1, o_2)$}; **Algorithm** $Alg^1_b(o_1, o_2)$ { return $p_2(o_1, o_2)$};

Algorithm $Alg^2_b(o_1, o_2)$ { return $p_3(o_1, o_2)$}; **Algorithm** $Alg^2_c(o_1, o_2)$ { return $p_4(o_1, o_2)$};

Algorithm $Alg^1_{\theta(a,b)}(o_1, o_2)$ { if $o_1 = p_1(t_1, t_2)$, $o_2 = p_2(t_2, t_3)$ then return $q_1(t_1, t_3)$}

Algorithm $Alg^2_{\theta(b,c)}(o_1, o_2)$ { if $o_1 = p_3(t_1, t_2)$, $o_2 = p_4(t_2, t_3)$ then return $q_2(t_1, t_3)$}

Algorithm $Alg^1_{e_1}(o_1, o_2)$ { return $q_1(o_1, o_2)$};

Algorithm $Alg^1_{e_2}(o_1, o_2)$ { return $q_2(o_1, o_2)$};

Algorithm $Alg^1_b(o_1, o_2)$ { return $p_2(o_1, o_2)$};

Algorithm $Alg^3_{\theta(e_1,e_2)}(o_1, o_2)$ { if $o_1 = q_1(t_1, t_2)$, $o_2 = q_2(t_2, t_3)$ then return $r(t_1, t_3)$}

Consider the interpretation $\mathcal{I}_1 = (Ob_1, ob_1, ALG_1)$ of the system H, where we specify only the useful entities allowing to exemplify the computation:

- $Ob_1 = \{Peter, Helen, John, Sorin, pizza\}$
- $ob_1(x_1) = Peter, ob_1(x_2) = Helen, ob_1(x_3) = John, ob_1(x_6) = Sorin, ob_1(x_5) = pizza$
- $ALG_1 = \{Alg^1_a, Alg^1_b, Alg^2_b, Alg^2_c, Alg^1_{\theta(a,b)}, Alg^2_{\theta(b,c)}, Alg^3_{e_1}, Alg^3_{e_2}, Alg^3_{\theta(e_1,e_2)}\}$

where $Alg^3_{e_1} = Alg^1_{\theta(a,b)}, Alg^3_{e_2} = Alg^2_{\theta(b,c)}$.

It is not difficult to observe that for the formula

$w = \sigma_3(\sigma_1(h([x_1, x_2], a), h([x_2, x_3], b)), \sigma_2(h([x_3, x_6], b), h([x_6, x_5], c))) = \sigma_3(\alpha, \beta)$

from the last part of the previous section we obtain the following computations:

$Val_H(\alpha) = \{Alg^1_{e_1}(o_3, o_4) \mid o_3 \in Val_H(h([x_1, x_2], a)), o_4 \in Val_H(h([x_2, x_3], b))\}$

$Val_H(h([x_1, x_2], a)) = \{Alg^1_a(Peter, Helen)\} = \{p_1(Peter, Helen)\}$

$Val_H(h([x_2, x_3], b)) = \{Alg^1_b(Helen, John), Alg^2_b(Helen, John)\} = \{p_2(Helen, John), p_3(Helen, John)\}$

therefore $Val_H(\alpha) = \{Alg^1_{e_1}(p_1(Peter, Helen), p_2(Helen, John)), Alg^1_{e_1}(p_1(Peter, Helen), p_3(Helen, John))\} = \{q_1(Peter, John)\}$

$Val_H(\beta) = \{Alg^2_{e_2}(o_5, o_6) \mid o_5 \in Val_H(h([x_3, x_6], b)), o_6 \in Val_H(h([x_6, x_5], c))\}$

$Val_H(h([x_3, x_6], b)) = \{Alg^1_b(John, Sorin), Alg^2_b(John, Sorin)\} = \{p_2(John, Sorin), p_3(John, Sorin)\}$

$Val_H(h([x_6, x_5], c)) = Alg^2_c\{Sorin, pizza)\} = \{p_4(Sorin, pizza)\}$

therefore $Val_H(\beta) = \{Alg^2_{e_2}(p_2(John, Sorin), p_4(Sorin, pizza)), Alg^2_{e_2}(p_3(John, Sorin), p_4(Sorin, pizza))\} = \{q_2(John, pizza)\}$

Finally, from $Val_H(\alpha)$ and $Val_H(\beta)$ we deduce

$Val_H(w) = \{Alg^3_{\theta(e_1,e_2)}(q_1(Peter, John), q_2(John, pizza))\} = \{A\ nephew\ of\ Peter\ likes\ to\ eat\ pizza\}$

We observe that the conclusion obtained by H can not be obtained neither by $\mathcal{S}_1$, neither by $\mathcal{S}_2$. This explains why H is named a *distributed system*.

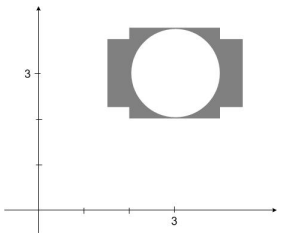


Figure 4: The image generated by $\mathfrak{I}_2$

We give now a short description of another interpretation $\mathfrak{I}_2$ for the same system H. As a result we obtain geometrical images.

• $Ob_2 = \{1, (3,3), (3,1.5)\}$

• $ob_2(x_1) = 1, ob_2(x_2) = ob_2(x_6) = (3,3), ob_2(x_3) = 1, ob_2(x_5) = (3,1.5)$

• $Alg^1_a(p,q)$*{Return the interior of circle with radius p and center q}*

• $Alg^1_b(p,q)$*{Return the interior of the square centered in p and the sides of length 2*q parallel with coordinate axes }*

• $Alg^1_{\theta(a,b)}(\alpha,\beta)$*{* If $\alpha = Alg^1_a(p,q)$ and $\beta = Alg^1_b(q,r)$ then return $\beta \setminus \alpha$ *}*

• $Alg^2_b(p,q)$*{Return the exterior of circle with radius p and center q}*

• $Alg^2_c(p,q)$*{Return the interior of the rectangle centered in p and the sides of lengths specified by q, parallel with coordinate axes }*

• $Alg^2_{\theta(b,c)}(\alpha,\beta)$*{* If $\alpha = Alg^1_{\theta(a,b)}(p,q)$ and $\beta = Alg^2_{\theta(b,c)}(q,r)$ then return $\beta \cap \alpha$ *}*

• $Alg^3_{\theta(e_1,e_2)}(\alpha,\beta)$*{* If $\alpha = Alg^2_a(p,q)$ and $\beta = Alg^2_c(q,r)$ then return $\beta \cup \alpha$ *}*

For the same formula $w \in \mathcal{F}(H)$ as in the previous computation, the object $Val_H(w)$ given by $\mathfrak{I}_2$ is shown in Figure 4.

5 A Java implementation

If we note by $\mathcal{A}$ the set consisting of some geometrical objects names then each system's input is an word $w = a_1 \dots a_k$ over the alphabet $V = \mathcal{A} \cup \{+,-\}$ having the following properties:

• $\mathbf{a_i} = +/-$ means a left/right rotation with a specific angle, denoted by δ and to draw a line on the current direction

• $\mathbf{a_i} = \mathbf{O_j}$ means to draw the graphical illustration of the object O_j such that its entry direction is on the current direction. In our implementation, each geometrical object used in the generation method is an instance of the a class named *Object*. Graphically, it is a representation of a figure inside a square. Every instance of this class can have one of the following types: **circle**, **triangle**, **star** and **square** corresponding to the figure it consists of. Other members of this class are the **entry direction** and the **exit direction** related to some corner of the object. The corner corresponding to the entry direction becomes the **entry point** of the object. Similar for the **exit point**. The main routine of the Algorithm is **createHDRS** (Algorithm 2). The construction of the system starts by defining the schemas of the agents (steps $1 \div 4$). The hyper-schemas of order one corresponding to the managers of the second level are constructed using the steps $7 \div 14$. The condition for existing a hyper-schema over two schemas is that their maximal paths are connected deductive paths. This property is verified using the routine **connectedPaths** (Algorithm 3). If the second level of the system was successfully defined (**If** condition of step 15) then the process of creating new levels in HDRS continues using the **While** loop of step 17. The hyper-schemas of orders greater than 2 are created using the routine **createHypSchs** (Algorithm 4).

The geometrical objects that are used for the image generation process are introduced using the first

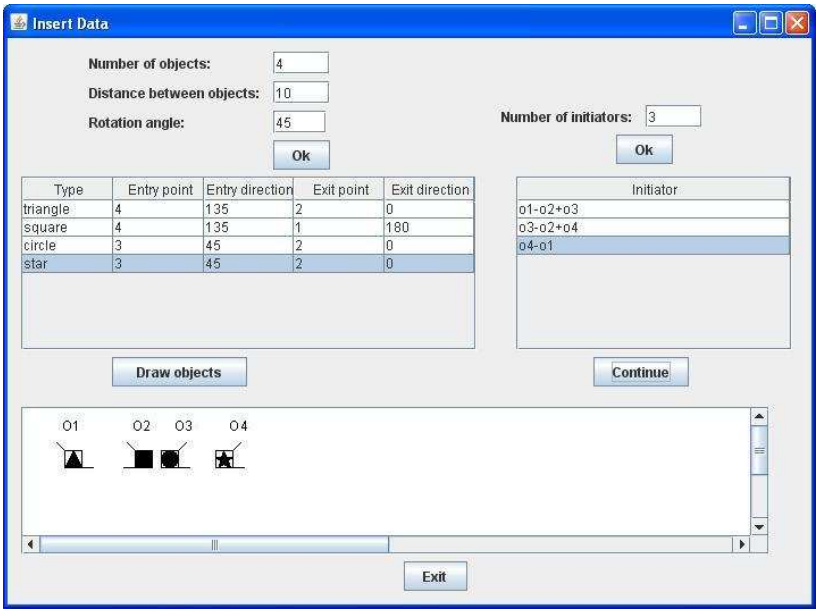
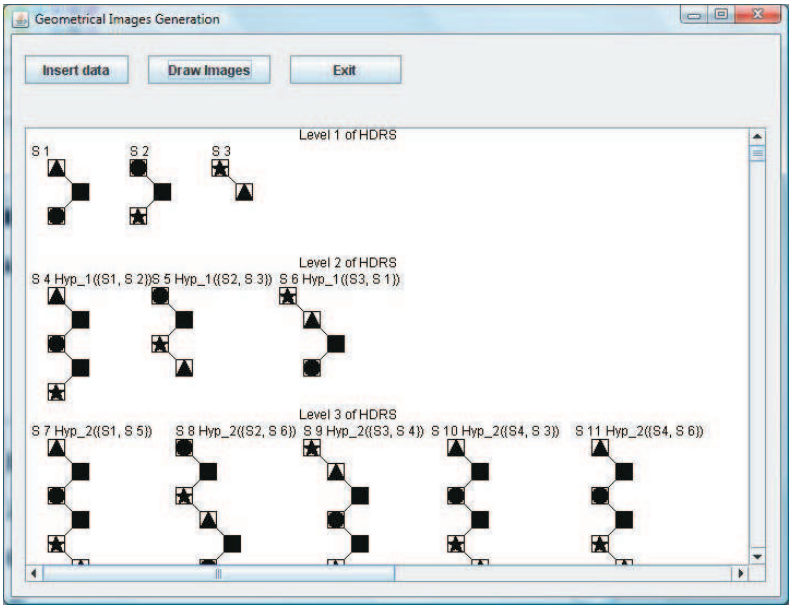


Figure 5: First window of the application

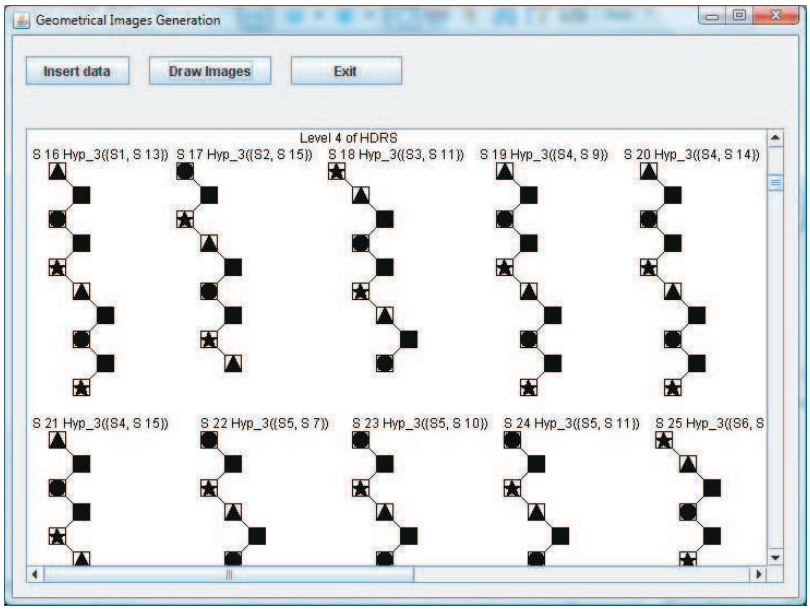
window of the application. For each object the user must specify the type, the entry and exit points (the corners are numbered starting from the down-left) and related to them the entry and the exit direction. Also, using the controls of the first window, the input descriptions can be edited(see Figure 5). The second window of the application gives the outputs provided by the system's reasoning components (see Figure 6). It consists of three buttons and a panel. The application can draw maximum 1000 images with maximum 50 geometrical objects per image.

6 Conclusions and future works

In this paper we formalized the syntactical and semantical computations in an HDR system. We exemplified these computations and for some HDR system H we gave two interpretations: one interpretation generates phrases and the other generates geometrical images. This examples give an idea concerning the generative power of our mechanism. We relieved also by these examples the fact that the distributed reasoning can be modeled by an HDR system. A short description of a Java implementation for an HDR system generating images is also given. We intend to develop the applications of an HDR system. First, we intend to use the mobile agents to process such systems ([4]). Second, we intend to use the HDR systems in e-learning. The basic idea comes from the fact that a link in an HTML document gives a reference to another document of the similar structure.



(a) The initiators and some images obtained by the managers of Q_2 and Q_3 levels



(b) Images obtained at the 4th level in the system

Figure 6: Second window of the application

Algorithm 2 Procedure createHDRS

```
Procedure createHDRS
1.    For i ← 1, noCmd
2.      call create_schema(commands[i], schema[i], agent[i])
3.      maximalPath[i] ← schema[i].getMaximalPath()
4.    EndFor
5.    noAg ← noComd
6.    noKM ← noAg + 1
7.    For i, j ← 1, noAg; j ≠ i
8.        If connectedPaths(maximalPath[i], maximalPath[j])
9.            call create_hyperSch(hypSch[noKM], schema[i], schema[j])
10.           hypSch[noKM].order ← 1
11.           maximalPath[noKM] ← hypSch[noKM].getMaximalPath()
12.           noKM ← noKM + 1
13.       EndIf
14.   EndFor
15.   If noKM > noAg + 1
16.       order ← 2
17.       While createHypSchs(order)
18.           order ← order + 1
19.       EndWhile
20.   EndIf
EndProcedure
```

Algorithm 3 Function connectedPaths

```
Function connectedPathsPath1, Path2
1.    If Path1.lastNode=Path2.firstNode
2.        return true
3.    EndIf
4.    If Path1.firstNode=Path2.lastNode
5.        return true
6.    EndIf
7.    return false
EndFunction
```

Algorithm 4 Function createHypSchs

```
Functioncreate HypSchsorder
1.    newHypSch ← false
2.    For i ← noKM − 1, noAg
3.        If hypSch[i].order ≠ order − 1
4.            continue
5.        EndIf
6.        For j ← 1, noKM − 1; j ≠ i
7.            If connectedPaths(maximalPath[i], maximalPath[j])
8.                newHypSch ← true
9.                If j ≤ noAg
10.                   call create_hyperSch(hypSch[noKM], hypSch[i], schema[j])
11.               Else
12.                   call create_hyperSch(hypSch[noKM], hypSch[i], hypSch[j])
13.               EndIf
14.               hypSch[noKM].order ← order
15.               maximalPath[noKM] ← hypSch[noKM].getMaximalPath()
16.               noKM ← noKM + 1
17.           EndIf
18.       EndFor
19.   EndFor
20.   return newHypSch
EndFunction
```

Bibliography

[1] Allen M., Prusinkiewicz P., DeJong T. (2004) *Systems for Modeling the Architecture and Physiology of Growing Trees: The L-PEACH Model*, Proceedings of the 4thInternational Workshop on Functional-Structural Plant Models, pp. 220-225

[2] Chang Shi-Kuo (1970) *The analysis of two-dimensional patterns using picture processing grammars*, Annual ACM Symposium on Theory of Computing archive, Proceedings of the second annual ACM symposium on Theory of computing, p. 206-216

[3] Drewes F., Ewert S., Klempien-Hinrichs R., Kreowsky H.J. (2003) *Computing raster images from grid picture grammars*, Journal of Automata, Languages and Combinatorics, Vol.8, Issue 3, p. 499-519

[4] Dzitac I., Bărbat B. E. (2009) Artificial Intelligence + Distributed Systems = Agents, *Int. J. of Computers, Communications & Control*, ISSN 1841-9836, E-ISSN 1841-9844, vol. IV, no. 1, pp. 17-26

[5] Ghindeanu M. (2008) Constructing Architectures for an Hierarchical Distributed Reasoning System Based on its Inputs, *International Multi-Conference on Engineering and Technological Innovation*, USA, p. 231-234

[6] Kandel A., Bunke H., Last M. (eds) (2007) Applied Graph Theory in Computer Vision and Pattern Recognition, Springer, Studies in Computational Intelligence 52

[7] Lindenmayer A. (1968) *Mathematical models for cellular interaction in development*, Parts I and II, Journal of Theoretical Biology (18), p. 280-315.

[8] Priss U., Corbett D., Angelova G. (Eds.) (2002) Conceptual Structures: Integration and Interfaces, *10th Int. Conf. on Conceptual Structures, ICCS 2002*

[9] Sharp D. (1998) *LMUSe version 0.7b,* http://www.geocities.com/Athens/Academy/8764/ lmuse/lmusetxt.html

[10] Sowa J.F. (1984) Conceptual structures- Information Processing in Mind and Machine, Addison-Wesley

[11] Ţăndăreanu N. (2004). *Semantic schemas and applications in logical representation of knowledge*, Proceedings of the 10th International Conference on Cybernetics and Information Technologies, Systems and Applications, USA, Vol.III, p. 82-87

[12] Ţăndăreanu N., Ghindeanu M. (2008) Hierarchical Semantic Structures Applied in Automatic Image Generation, *Proceedings of 11th IASTED International Conference on Intelligent Systems and Control*, ISBN: 978-0-88986-777-2

[13] Ţăndăreanu N., Ghindeanu M. (2008) Path-based Reasoning in Semantic Schemas, *Annals of University of Craiova, Mathematics and Computer Science Series*, Vol.35, p.171-181

[14] Zu Song-Chun , Mumford D. (2006) *A stochastic grammar of images*, Foundations and Trends in Computer Graphics and Vision, Vol. 2, Issue 4, p. 259-362

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 178-184

Optimization for Date Redistributed System with Applications

Mădălina Văleanu, Smaranda Cosma, Dan Cosma, Grigor Moldovan, Dana Vasilescu

Mădălina Văleanu
University of Medicine and Pharmacy "Iuliu Hatieganu"
Medical Informatics and Biostatistics Department
Cluj-Napoca
E-mail: mvaleanu@umfcluj.ro

Smaranda Cosma
Babes-Bolyai University
Business Department, Faculty of Business
Cluj-Napoca
E-mail: smaranda.cosma@tbs.ubbcluj.ro

Dan Cosma
University of Medicine and Pharmacy "Iuliu Hatieganu"
Department of Pediatric Surgery and Orthopedics
Cluj-Napoca
E-mail: dcosma@umfcluj.ro

Grigor Moldovan
Babes-Bolyai University
Computer Systems Department
Cluj-Napoca
E-mail: moldovan@cs.ubbcluj.ro

Dana Vasilescu
University of Medicine and Pharmacy "Iuliu Hatieganu"
Department of Pediatric Surgery and Orthopedics
Cluj-Napoca
E-mail: dana.vasilescu@umfcluj.ro

Abstract: In this paper we intend to define a strategy for managing databases with mobile structures, taking into account their redistribution in the nodes of a computer network. The minimal cost of the redistribution is highlighted and some applications for medical and business databases are presented.
Keywords: distributed database, costs, medical databases, economical databases, wireless network

1 Introduction

The paper presents a generalization and extension for mobile environments of the context of the problem expressed in paper [1]. It also presents some applications of the problem. Let us consider the mobile databases (tables) $B_i, i = \overline{1,n}$ distributed in r nodes of a network of computer stations with own memories $S_i, i = \overline{1,r}$. Hence, we have:

$B = \{B_1, B_2, \ldots, B_n\}$ and $S = \{S_1, S_2, \ldots, S_r\}$,

where B is a distributed database.

Now we identify the nodes, that is the stations of the computer network considered with the same symbols as for the memory supports S_i in S.

The architectures of the fixed computer networks with nodes $S = \{S_1, S_2, \ldots, S_r\}$ can differ, but on the horizontal level, in general, they are modeled through a relative graph. Particular computer network architectures can exist, such as the hypercube or generalizations of the hypercube [2] [3], which provide simple and efficient routing means but whose complexity in number of nodes and restrictions make them inefficient in the end. Modeling a computer network in a graph, in general, will give a static network with fixed and well defined geographical locations.

In our previous paper [1] we have supposed that subbases B_i of the database B have a well defined location and that this location is maintained fixed during the running of a distributed application.
At present, mobile databases characterized by allocation in permanent change are known. Consequently, dynamism characterizes all their aspects. They are searched for (selected), accessed and processed from a mobile environment made up of laptops, mobile phones etc. They become stronger and stronger elements of data processing. They are connected by means of Wireless stations in special fixed points (nodes) belonging to a computer network. The traditional model of transactions migrates towards a mobile transaction model.
The following presents a strategy for this kind of databases processing.

Databases $B_i, i = \overline{1,n}$ and other soft resources (programs) shall be stored in fixed hosts, noted FH and identified through S_i (fixed host computers!), in a well established network, $S = \{S_1, S_2, \ldots, S_r\}$. The mobile environment will inherit the properties of the distributed environment. The mobile environment with ever increasing storing spaces will be able to take over the mobile databases for processing data, by duplicating procedures, and the results will turn back to a fixed host.

The fixed hosts will be those that permanently preserve the data subbases B_i. Databases B_i could be preserved in mobile hosts, but any update (modification) should end with their replication to a fixed host mentioned with a new version number. In the fixed network, it is necessary to update all duplicated subbases B_i, using the most recent version numbers allotted in order to maintain the consistency of the distributed database B. The update of the databases should occur instantly after the alteration of their version number. The update of duplicated data subbases B_i can also be achieved with the help of a jeton function periodically passing through the nodes S of the network in consideration.

It is obvious that between the mobile hosts noted MH and fixed hosts noted FH belonging to a computer network, there is a fixed interface, called mobile support station, noted MSS (mobile support station) or base station. The connection between a mobile support station (MSS) and a mobile host (MH) is wireless. Each MSS allotted to a node S_i controls a cell of mobile hosts identified by $\{i_1, i_2, \ldots, i_k\}$.
A mobile host can disconnect from a MSS and possibly reconnect to another MSS just while running a distributed application. The disconnection and connection of mobile stations occurs frequently. Disconnection establishes a new distribution of subbases in the network nodes.

The figure below presents the general architecture of a mobile platform.

In the design of a mobile database, in each node S_i we shall highlight - for B_i - an often modifying and dynamic component that frequently is accessed from other nodes of the network, noted b_i and other components, noted $m_{i(s)}$ associated to mobile units, quite often accessed locally to mobile units but where the modification of data is less frequent.

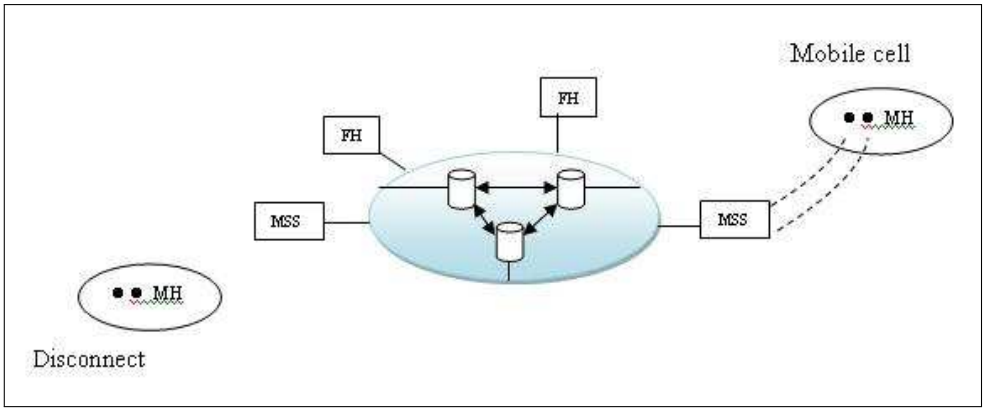


Figure 1: Mobile platform

We have

$B_i = (b_i; m_{i(1)}, m_{i(2)}, \ldots, m_{i(k)})$

where we used for indices i_p notation $i(p)$ and $p = \overline{1,k}$. We can consider that we made a decomposition of the subbases B_i by selection and/or projection operations so that by further union operations we get updated subbase B_i.

At a given moment in the computer network stations we witness a certain distribution, respectively a grouping of databases B_i, $i = \overline{1,n}$. In general, if $n \geq r$, then, more subbases B_i will exist in the memory S_i. For reasons of simplification, in our study, we will suppose that consecutively, in each station S_i of the r station, we will have d subbases S_i , hence $n = d.r$. We also suppose that we have removed, by means of a certain conveniently selected strategy, the duplications of subbases B_i, therefore B_i are distinct. A distributed application supposing programs running in the network under consideration leads to the access, from the nodes S_i, of the subbases B_j in a defined succession, until the result needed is obtained. We note the successively accessed data subbases (some of them more often accessed in a distributed application) in a vectorial form, as follows:

$B_L = (b_{m_1}, b_{m_2}, \ldots, b_{m_s})$

then

$L = (m_1, m_2, \ldots m_s); m_k \in \{1, 2, \ldots, n\}$

Remember that m_k identifies the place in the succession of accesses of the subbases in B performed.

In general, in the case of an access from S_i network node to a subbase found in S_j, a so-called penalty should also be considered (for instance: time, cost) noted with p_{ij} for all $i, j = \overline{1,r}$ which will be composed of a fixed penalty pf_{ij} established in a fixed computer network and possibly from a penalty due to the specific working manner with mobile databases noted pm_{ij}. We will obtain $p_{ij} = pf_{ij} + pm_{ij}$.

2 Grouping data subbases on the network fixed nodes

Let distributed database be $B = \{B_1, B_2, \ldots, B_n\}$. The indices of subbases build the set $I_n = \{1, \ldots, n\}$. The reorganisation of these data is defined by permutation indices

$$\sigma = \begin{pmatrix} 1\ 2\ \ldots k \ldots n \\ i_1 i_2 \ldots i_k \ldots i_n \end{pmatrix}$$

where $i_k \in I_n; k = \overline{1,n}$ are distinct.

The permutation σ is also written as $\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(n))$. If we have two permutations σ and τ ,

their produce $\sigma\tau$ is obtained by the composition of the two functions, so that

$\sigma\tau = (\sigma(\tau(1)), \sigma(\tau(2)), \ldots, \sigma(\tau(n)))$.

Let us mark with $Supp(\sigma)$, the permutation support σ, i.e. the set of the elements $i \in \{1, 2, \ldots, n\}$ having the property $\sigma(i) \neq j$.

A permutation σ is called cyclic of length $m, m \geq 2$ if elements $i_1, i_2, \ldots, i_m \in Supp(\sigma)$ exist so as $\sigma(i_1) = i_2, \sigma(i_2) = i_3, \ldots \sigma(i_{m-1}) = i_m, \sigma(i_m) = i_1$.

It is known that any permutation that is different from the identical one can be written as a produce of distinct cycles. [4]

Let us consider that $n = d.r$, i.e. on any S_i station in the r computer network we have d data subbases. Let us take the following reorganisation (distribution) of the n data subbases on the r workstations, successively:

$B_{\sigma^{-1}(1)}$	$B_{\sigma^{-1}(2)}$	$B_{\sigma^{-1}(d)}$	in S_1
$\ldots \ldots \ldots \ldots$			
$B_{\sigma^{-1}(1+(i-1)d)}$	$B_{\sigma^{-1}(2+(i-1)d)}$	$B_{\sigma^{-1}(id)}$	in S_i
$\ldots \ldots \ldots \ldots$			
$B_{\sigma^{-1}(1+(r-1)d)}$	$B_{\sigma^{-1}(2+(r-1)d)}$	$B_{\sigma^{-1}(rd)}$	in S_r

Note: σ is a bijective application, $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ where $\sigma(k) = i_k$,

respectively $\sigma^{-1}(i_k) = k$; $k = \overline{1, n}$.

Consider the sequence of successive accesses $L_D = (m_1, m_2, \ldots, m_s)$ of the subbases in B and the indices of these stations are. With these notations, we define the cost of a distributed application with the relationship

$$C(S, B, L_D, \sigma) = \sum_{k=1}^{s-1} (p_{R_\sigma(m_k), R_\sigma(m_{k+1})} + q_{R_\sigma(m_k)})$$

where $q_{R_\sigma(m_k)}$ represents the cost of the activities in station $S_{R_\sigma(m_k)}$.

Let us note with a_{ij} the number of the times $(m_k, m_{k+1}) = (i, j)$, $k = \overline{1, s-1}$;
$i, j \in \{1, 2, \ldots, n\}$, and with c_i , how many times
$m_k = i, k = \overline{1, s}; i, j \in \{1, 2, \ldots, n\}$.
Then,

$$C(S, B, L_D, \sigma) = \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{ij} p_{R_\sigma(m_k), R_\sigma(m_{k+1})} + c_i q_{R_\sigma(m_k)})$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{\sigma^{-1}(i), \sigma^{-1}(j)} p_{R_I(i), R_I(j)} + c_i q_{R_I(i)})$$

The cost of a distributed application will be:

$$C(S, B, L_D, \sigma) = \sum_{i=1}^{n} \sum_{j=1}^{n} (a_{\sigma^{-1}(i), \sigma^{-1}(j)} p^*_{ij} + c_i q^*_i), \text{ where } p_{R_I(i), R_I(j)} = p^*_{ij}; q_{R_I(i)} = q^*_i$$

A way to find out this sums can see in [5].

Notes

1. In practice, we can consider that p_{ij} is symmetrical, hence p^*_{ij} will also be symmetrical, i.e. $p^*_{ij} = p^*_{ji}$. Penalties p_{ij} can be determined, for instance, with the help of statistical data after more run-

nings of the programs of the distributed application D.

2. If permutation σ is decomposed in a produce of cyclical permutations, the formula for the application cost can be simplified accordingly.

3 Applications

3.1 Medical Databases

The medical field is one of the most important fields of social concern. It is defined by high level dynamism.

Nowadays the healthcare system from every country have to faced to many challenges of the 21st century which influenced significant the financial aspects of the organizations' activities. In these new conditions the competitions increased and pressed on the costs to acquire and maintain a high quality of the technology and capital outlay.

Today, the data needed for research are registered in medical documents, in a written form, but not in a unique form and the work necessary to retrieve and process such data is enormous.

It is for this reason that it is required to design general interest databases, for doctors, patients, researchers and health units. Such databases are characterised by the fact that they will include an enormous volume of data distributed among the nodes of a computer network.

The medical field activities can be divided into primary level care, secondary level care activities and the division can go even further.

Primary care refers to family medicine that is to the first contact and consultation point for the patient. Secondary care is the service provided by a specialist, who does not have the first contact with the patient, in general. Usually, a doctor providing secondary care services treats patients previously consulted by a family doctor.

To establish a correct diagnosis and a proper treatment it is necessary to handle the databases in question that, often mean to return to their redistribution in the nodes of the computer network. A concrete examples are given in papers [6]. In such an application it is equally important to find an answer to a query to the databases in as short a time as possible.

A valuable component of a heath care computer-based system is its capacity to work with information on the patients stored in various locations: the national health authority, the social insurance system, providers of primary and secondary care services. In order to find solutions, a distributed approach should be taken into consideration. [7] [8]

Health care information, such as medical records, X-rays, lab tests results are more are more kept and processed by computers. That is why standards to send such data away in an unambiguous manner among computers are required. In this way, uniform health care-related information shall also be available.

A standard represents a conceptual lexis common for all the stakeholders in the healthcare network. Standards of this kind are already available for many medicine subfields.

3.2 Economical Databases

The economic applications of distributed databases are diverse and numerous and their importance is also valuable. It is enough to take as an example a system in the field of international trade involving the handling of very large volumes of data relative to the varied goods distributed among shop networks. The management information system of a company becomes decisive for international corporations of the Learning-Company type that extend more and more their activities using varied international market penetration strategies.

E-commerce develops and includes larger and larger scopes. That is why it is necessary to establish standards related to marketed products, in this field too. The queries put to distributed and increasing in size databases should be made more efficient in time. It is sure that in the near future the optimising of costs (query types) of the used databases will have to be considered.

Marketing research identifies, collects, processes, analyses, interprets and communicates information relevant for a specific marketing situation to make a decision. All the data are organised in distributed databases in more locations.

Marketing research mainly aims at reducing risk and uncertainty in the conception and grounding of marketing-related decisions and at implementing and controlling the practical putting into application of these decisions. [7]

4 Conclusions

The fundamental problem, with respect to the distributed application D under consideration, consists in the determination of a permutation σ in the set of possible permutations P having elements $\{1, 2, \ldots, n\}$, indices of the B data subbases so that the cost of the use of the distributed application D programs would be minimal, i.e.

$\min\{C(S, B, L_D, \sigma); \sigma \in P\}$.

It is obvious that the problem relates to combination, its solution is important when the distributed application D is used repeatedly. The problem is solved once and the advantage remains operational all along the use of the respective distributed application.

The applications for this are multiple. Here are describes some applications in medical and economical domain, that can use distributed databases and where the problem of cost minimizing (time need to transferred information from one point of the network to another) is very important.

Bibliography

[1] G. Moldovan, M. Văleanu, *The performance optimization for date redistributing system in computer network,* International Journal of Computer, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844, Vol III, Supl. issue, p. 116-118, 2008.

[2] G. Moldovan, M. Văleanu, *Redistributing databases in a computer network,* Analele Univ. Bucureşti, Ser. Math.-Info., 56, 2006.

[3] G. Moldovan, I. Dzitac, *Sisteme distribuite - Modele matematice,* Ed. Univ. Agora, 2006.

[4] L. Aspinall, *Data base. Re-organisation –Algorithms,* IBM, UKSC – 0029, 1972.

[5] A. Gog, H. Grebla, *Evolutionary Tuning for Distributed Database Performance,* The 4th International Symposium on Parallel and Distributed Computing (ISPDC), Lille, France, IEEE Computer Society, 2005, p 275-281.

[6] S. Cosma, D. Cosma , A. Negrusa, M. Văleanu, G. Moldovan, D. Vasilescu, *Implementation of the communication system for clubfoot,* WSEAS Transactions on Communications, ISSN:1109-2742, ISSUE 9, Vol 7, sept 2008, p.932-941

[7] D.E. Vasilescu, D. Cosma, M. Văleanu, I. Negreanu, D. Vasilescu, *The results of the early conservative orthopedic treatment in the congenital talipes equinovarus,* Applied Medical Informatics 2004, Vol 15, p 34-43.

[8] D. Cosma, S. Cosma, M. Văleanu, D.E. Vasilescu, G. Moldovan, *Web - based guideline for clubfoot: patient - orientated materials,* Journal of International Business and Economics, ISSN 1544-8037, 8-1, 2008.

Int. J. of Computers, Communications & Control, ISSN 1841-9836, E-ISSN 1841-9844
Vol. IV (2009), No. 2, pp. 185-197

Quality Control of Statistical Learning Environments and Prediction of Learning Outcomes through Reproducible Computing

Patrick Wessa

K.U.Leuven Association
Lessius, Dept. of Business Studies
Belgium
E-mail: patrick@wessa.net

Abstract: This article introduces a new approach to statistics education that allows us to accurately measure and control key aspects of the computations and communication processes that are involved in non-rote learning within the pedagogical paradigm of Constructivism. The solution that is presented relies on a newly developed technology (hosted at www.freestatistics.org) and computing framework (hosted at www.wessa.net) that supports reproducibility and reusability of statistical research results that are presented in a so-called Compendium. Reproducible computing leads to responsible learning behaviour, and a stream of high-quality communications that emerges when students are engaged in peer review activities. More importantly, the proposed solution provides a series of objective measurements of actual learning processes that are otherwise unobservable. A comparison between actual and reported data, demonstrates that reported learning process measurements are highly misleading in unexpected ways. However, reproducible computing and objective measurements of actual learning behaviour, reveal important guidelines that allow us to improve the effectiveness of learning and the e-learning system.
Keywords: Reproducible Computing, Learning Environment, Quality Control, Statistics Education, Psychometrics

1 Introduction

In education-related research it is common practice to investigate learning processes through measurements that are based on questionnaires. Reported measures often reveal interesting information about a wide variety of aspects of computing-assisted learning such as: computer attitudes [22]; computer emotions and knowledge [17]; learner experiences and satisfaction [34]; etc... The importance of such measurements has been highlighted by many authors from various perspectives ([7], [15], [12]) especially from the perspective of the constructivist pedagogical paradigm ([35], [30], [11], [24]).

These reported measures, while intrinsically interesting, may not always provide us with the information we need to assess and improve systems that support e-learning. Moreover, the implementation of new learning technologies and data analysis tools open up a wide array of measurement opportunities which lead to new areas of research. An excellent example is the use of data mining tools in the open source e-learning environment called Moodle [28].

Even though it seems to be very difficult to measure and empirically prove [25], there is no doubt in my mind that the introduction of computers in homes and classrooms has led to an improvement of overall learning productivity, educational communication mechanisms, social constructivism, and collaboration. However, the use of computers and software in statistics education may - unwillingly - result in several types of adverse effects because the complex processes that are required to learn and (truly)

understand statistical concepts are often mystified by technicalities and a variety of practical problems that have nothing to do with mathematics or statistics. It is within this context that I argue that a system for Quality Control should be embedded into the e-learning system, which is not limited to the Virtual Learning Environment but extends to the statistical software, databases, and learning repositories (Statistical Learning Environment).

There is an important, additional benefit for implementing such a monitoring and control system - it is directly related to the problem of irreproducible research which has received a great deal of attention within the statistical computing community ([9], [26], [29], [14], [13], [18], [10]). The most prominent citation about the problem of irreproducible research is called Claerbout's principle ([9]):

> *An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and that complete set of instructions that generated the figures...*

Several solutions have been proposed ([5], [10], [19]) but have not been adopted in statistics education because they require students to understand the technicalities of scientific word processing (LaTex) or statistical programming (R code). Based on a newly developed Statistical Learning Environment (SLE) I propose a solution that is feasible for educational purposes and allows us to monitor, research, and control the learning processes based on the dynamics of between-student communication and collaboration.

2 Reproducible Computing

2.1 R Framework

The R Framework allows educators and scientists to develop new, tailor-made statistical software (based on the R language) within the context of an open-access business model that allows us to create, disseminate, and maintain software modules efficiently and with a very low cost in terms of computing resources and maintenance efforts [36]. The so-called R modules empower students to perform statistical analysis through a web-based interface that does not require them to download or install anything on the client machine. This permits students to focus primarily on the interpretation of the analysis - however, the R Framework also allows advanced students and scientists to inspect and change the R code that was coded by the original author. This results in the creation of so-called derived R modules that may be better suited for particular purposes.

There are several important reasons why the R Framework helps in controlling the quality of the statistical learning processes that are supported by the computer:

- The R modules are web applications with an advanced session management which includes all aspects of the computations that are executed. In addition, the session manager uses attributes that identify the student and the course in which (s)he is enrolled. Therefore all computations that are performed within the context of a statistics course can be associated with an individual student - to implement this feature, the educator only needs to use certain HTML tags in the hyperlink that is inserted in the virtual learning environment.

- Every R module is uniquely described by an expandable set of meta data (incl. the actual statistical code) which can be stored and transmitted. This implies that every computation that is executed can be uniquely defined by the R module's meta data and additional information about the data and the parameters that have been specified by the user. As a consequence, every computation can be uniquely described and archived with meta data.

- The R Framework allows other servers (under certain conditions) to send meta data through an ordinary HTTP request which allows it to rebuild and execute the R module with the specified data and parameters in real time. Therefore it is possible to remotely store computational objects and send them back to the R Framework such that the original computation can be reproduced and reused.

- All the processes that are associated with the above items are automatically stored in a so-called process measurement database. This implies that all computer-assisted learning activities are objectively measured and stored for the purpose of analysis.

2.2 Compendium Platform

If a derived R module contains generic improvements or if a computation needs to be communicated to other students/scientists then it is necessary to have a simple, transparent mechanism that allows one to permanently store the computation in a repository of computational objects that can be easily retrieved, recomputed, and reused. Such a repository was recently created within the OOF 2007/13 project of the K.U.Leuven Association and is called the Compendium Platform. The main reason for creating the R Framework and the Compendium Platform, is that it allows anyone to create and use Compendia of reproducible research. A Compendium is defined as [37]: *a research document where each computation is referenced by a unique URL that points to an object that contains all the information that is necessary to recompute it*. Such documents can be easily created (even by students) and permit any reader to (exactly) recompute the statistical results that are presented therein. A few simple clicks are sufficient to have the R Framework reproduce the results and to reuse them in derived work [37]. The practical implications of this technology will become obvious in section 3 because the three figures that are presented can be recomputed and reused through the Compendium Platform.

2.3 Communication, Feedback, and Learning

The concept of Reproducible Computing was implemented in several undergraduate statistics courses in order to thoroughly test the new system and to measure key aspects of the educational activities and experiences. Two different student populations were investigated in detail: a group of (academic) bachelor students, and a group of so-called switching students. The second population is of particular interest because it consists of students who obtained a (professional) bachelor degree and decided to make the switch to an academic master which requires them to complete a preparatory year.

On the one hand, switching students are highly motivated and more mature than the bachelor students. A priori, one would expect them to prefer practical activities (such as communication and computing) above theory and critical reflection. On the other hand, one might expect the bachelor students to have a more critical (scientific) attitude and better mathematical background than the switching students.

Students from both populations took a similar statistics course which covered topics from introductory statistics, regression analysis, and introductory time series analysis. The main learning activities in both statistics courses were based on a weekly series of workshops where each student was required to investigate practical, empirical problems. At the end of each week, students submitted their papers electronically. During the lecture I proposed a series of solutions and illustrated commonly made mistakes. After the lecture, students had to work on the next assignment and complete a series of peer reviews (assessments) about the work that was submitted the week before. The assessment grades did not count towards the final score - however, each submitted peer review was accompanied by verbal feedback messages. I graded a (quasi random) sample of these messages in order to provide students with a strong incentive to take the review process seriously. There is strong empirical evidence that this approach had beneficial effects on non-rote learning of statistical concepts [38].

3 Objective Measurements versus Reported Data

In a recent paper [37] it is illustrated how the Compendium Platform's repository supports "technical" quality control of the statistical software and accompanying documentation for students. On the one hand, reproducible computing allows students to accurately communicate computational problems and questions without the need to understand the underlying technicalities. On the other hand, it allows the educator (and creator of the computational software) to analyze the reported problem (based on the detailed, raw output of the R engine that executed the request) and to transparently communicate the solutions to the students. Moreover, the measurement of learning activities and experiences is a conditio sine qua non for controlling the "overall" quality of the SLE. This will be illustrated, based on the data that have been collected from both student groups. At the same time, the importance of objective (as opposed to reported) measurements is illustrated based on a simple, comparative diagnostic tool.

The reported measurements were obtained through questionnaires on a 5-point Likert scale and should consequently be treated as ordinal data. The questions were based on well-known psychological surveys ([12], [8]) and the IBM computer system usability survey [20] which was adapted and extended [27]. Useful data was obtained from a total of 111 bachelor students and 129 switching students - the response ratio was very high (between 82.9% and 92% depending on the questionnaire). All observations of actual learning activities were measured on a ratio scale (the number of archived computations and the number of submitted feedback messages). A total number of 34438 meaningful, verbal feedback communications and 6587 archived computations were registered.

In order to compare the actual and reported data, all measurements were converted to ordinal rank orders. In addition, the Pearson's rho correlations and Kendall's tau rank correlations ([1], [2], [16]) that represent the degree of linear association between the properties under investigation, were computed (these can be consulted in the archived computations about the Figures). In electronic versions of this paper, one can simply (ctrl-)click the hyperlinks below Figures 1, 2, and 3 to view the archived computation in the repository. Readers of the printed version of this document, have to manually enter the respective URLs into their internet browser to view the statistical computations that have been stored (at www.freestatistics.org).

Figure 1 displays the bivariate kernel density [21] between the rank order of the number of feedback messages that have been submitted in peer reviews about the workshops (x-axis) and the rank order of the number of (reproducible) computations that have been archived in the repository (y-axis).

The rank orders have been computed within the Bachelor population for the top panels, and within the Switching population for the bottom panels. This implies that the ranks that are attributed to female and male students are expressed on the same axes and can be compared. Figure 1 clearly demonstrates that female bachelor students are much more involved in feedback and computing than their male colleagues. At the same time, female switching students are more computing-oriented whereas the male switching students seem to have a slight preference for feedback communication. This information has important repercussions for controlling the quality of the learning environment and it provides clear guidelines towards actions that should be taken (by me) to improve participatory incentives towards male bachelor students in future courses. Would I have been able to gain this insight based on reported measurements alone? The answer is clearly negative (as is illustrated in Figures 2 and 3).

It is quite obvious that male bachelor students highly over-estimate their performance in terms of feedback submissions (see Figure 2) because the rank orders of reported measures (x-axis) are higher than the ranks of actual feedback submissions (y-axis). Female bachelor students however, underestimate their involvement (relative to their male colleagues) because they are concentrated above the diagonal line. In the male switching student population several clusters of high density can be detected which leads us to conclude that we cannot treat them as one homogeneous group.

In Figure 3 the comparison between reported computing measures (x-axis) and actual computing (y-axis) leads to similar conclusions. Male bachelor students highly exaggerate their efforts, whereas

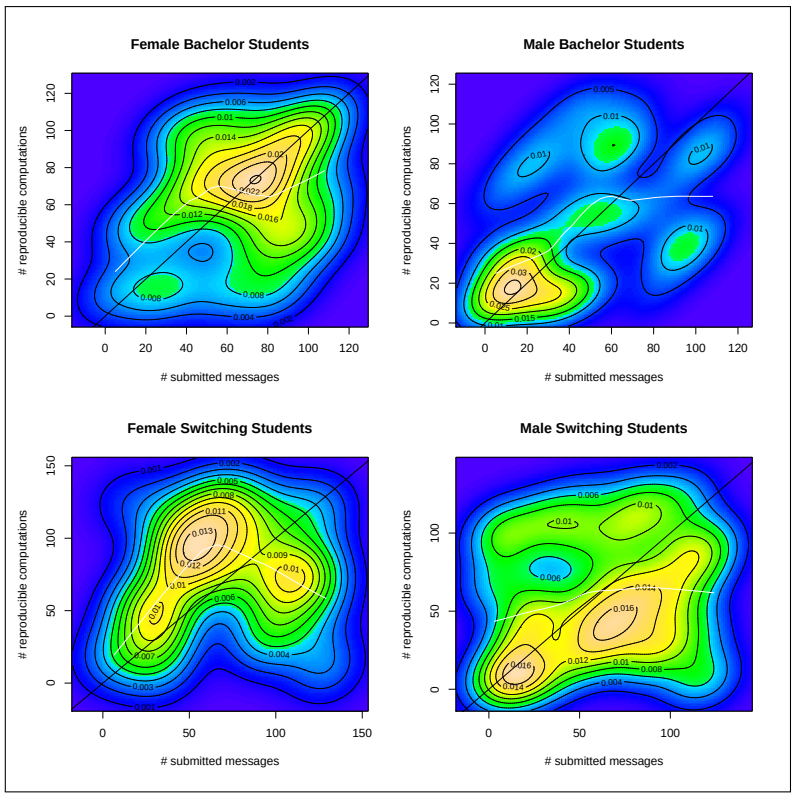


Figure 1: Submitted Feedback versus Reproducible Computations
www.freestatistics.org/blog/date/2008/Jun/30/t1214840420q0fyankop4x9ebf.htm

female bachelor and switching students underestimate themselves. The group of male switching students is heterogeneous.

Overall, the testimony of students is extremely misleading and poorly correlated with actual observations. If we would have recomputed Figure 1 with reported measures then the conclusions would have been the opposite of what is true. The reader can try out this experiment by simply reproducing the computation of Figure 1 with reported measures on both axes.

4 Quality Control

In order to be able to control (and improve) the quality of the SLE, it is necessary to estimate the impact of key-aspects of the learning processes that are associated with the SLE. The methodology that allows us to do this is based on a mathematical model which is described in [40] and relates the learning outcomes to objectively measured activities and reported experiences.

Typically, models that predict learning outcomes based on exogenous variables that are related to the learning (and computing) environment have an extremely low percentage of variance explained. In a recent and extensive study [25], six models were discussed that predicted the Statistics subtest scores of the Massachusetts Comprehensive Assessment System - the variance explained ranged between 4% and 7%.

It is obvious that any model that is used to control the quality of an SLE should perform much better. There are three important requirements to build high-quality models:

1. high-quality exogenous variables (preferably based on objective measurements) [39];

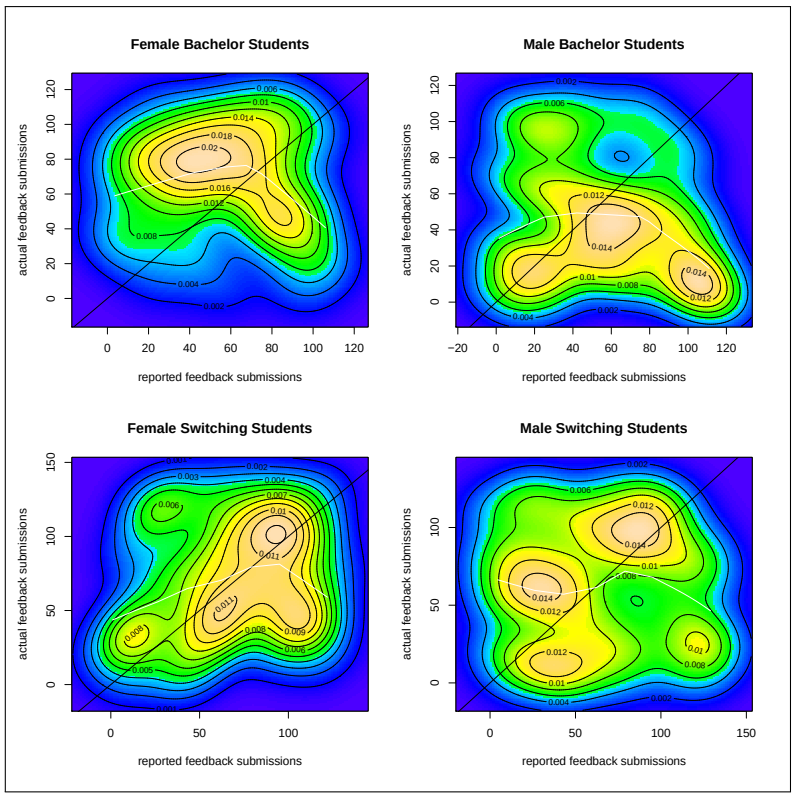


Figure 2: Reported versus Actually Submitted Feedback
http://www.freestatistics.org/blog/date/2008/Jun/30/t12148409608o0dnj2k4s04jil.htm

2. high-quality endogenous variable (c.q. test scores) based on optimal weights of the individual items (section 4.1, [40]);

3. homogeneous sample for which the model is computed.

The third condition refers to the fact that student populations may consist of different types of students with specific learning behaviors. In the formentioned statistics course there were 4 groups with distinct characteristics. This is clearly illustrated in section 3 and in Figures 1, 2, and 3.

Instead of computing separate models (for each of the sub populations) section 4.2 presents a comprehensive model with all combinations of interaction effects (male/female and Bachelor/Switching). This greatly improves the interpretation of the prediction model and allows us to perform differential quality control of the SLE.

4.1 Model

First, a classical regression approach is used to predict the learning outcomes (c.q. exam scores) as a linear function of $(K-1) \in \mathbb{N}_0$ exogenous variables of interest. Let $\vec{y}$ represent an $N \times 1$ vector for all $N \in \mathbb{N}$ students (with $N > K$), containing the weighted sum of G item scores (c.q. scores on individual exam questions): $\vec{y} \equiv \sum_{j=1}^{G} \omega_j \vec{y}_j$ with initial unit weights $\omega_j \equiv 1$. In addition, define an $N \times K$ matrix X that represents all exogenous variables (including a one-valued column which represents the constant), and a $K \times 1$ parameter vector $\vec{b}$ that represents the weights of the linear combination of all columns in X that is used to describe $\vec{y}$. The complete model is denoted M_1 and is defined by $\vec{y} = X\vec{b} + \vec{e}$ where $\vec{e} \leftarrow$ iid $N(\vec{0}, \sigma_e^2)$ represents the prediction error.

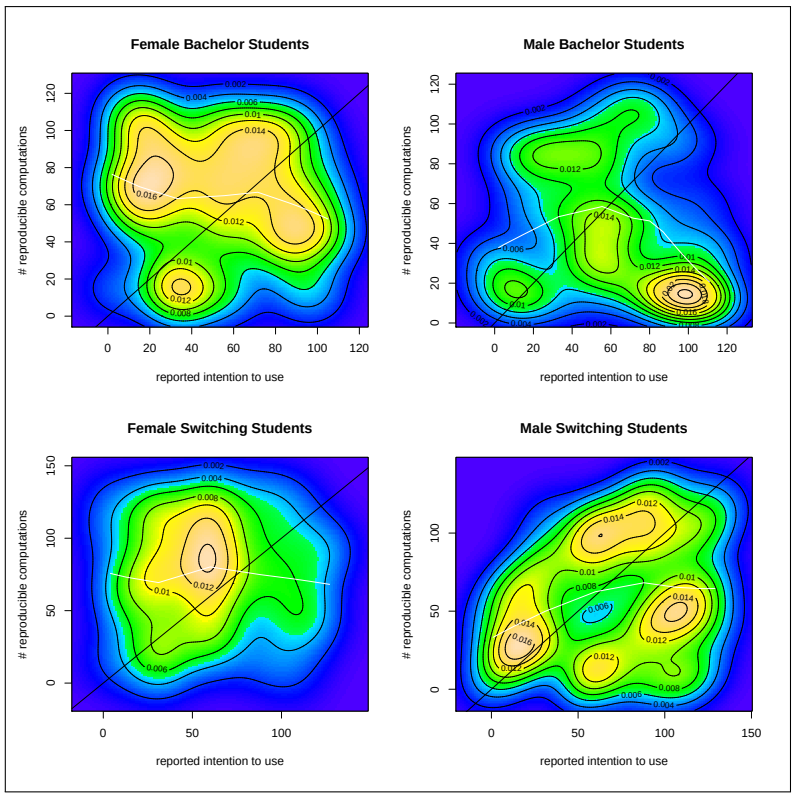


Figure 3: Reported versus Actual Reproducible Computing
http://www.freestatistics.org/blog/date/2008/Jun/30/t1214841152sn6jlyhgseclgqm.htm

In the second model M_2, the prediction of the first model is specified by a linear combination of the individual items (questions) that made up the total exam score. Let Y represent the $N \times G$ matrix that contains all G item scores, then it is possible to define the model $\hat{\vec{y}} = Y\vec{c} + \vec{a}$ where $\vec{a} \leftarrow$ iid $N(\vec{0}, \sigma_a^2)$. Note that there is no constant term in this model.

The third model (M_3) simply combines M_1 and M_2 by relating $\hat{\hat{\vec{y}}}$ to X in the regression model $\hat{\hat{\vec{y}}} = X\vec{f} + \vec{u}$. The estimator for $\vec{f}$ can be shown to be $\hat{\vec{f}} = (X'X)^{-1}X'\hat{\hat{\vec{y}}} = (X'X)^{-1}X'Y(Y'Y)^{-1}Y'X(X'X)^{-1}X'\vec{y}$ ([40]). M_3 is likely to yield different results from M_1 unless the estimated parameters M_2 are (nearly) equal to the original weights $\hat{\vec{c}} = (\hat{c}_1, \hat{c}_2, \hat{c}_3, ..., \hat{c}_G)' \simeq (\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3, ..., \hat{\omega}_G)'$.

From a statistical point of view it is not possible to test the improvement that is induced by the objective exam score transformations. The reason for this is that the traditional F-test assumes that the endogenous variables in two models (M_1 and M_3) to be compared are identical. Therefore it is necessary to use an auxiliary model (M_3^*) which is based on M_3 and includes $\vec{y}$ as an explanatory variable. This extended model $\hat{\hat{\vec{y}}} = X\vec{f} + \vec{y}g + \vec{u}$ can be shown to be equivalent to $\left(Y(Y'Y)^{-1}Y'X(X'X)^{-1}X' - gI_N\right)\vec{y} = X\vec{f} + \vec{u}$ such that it can be concluded that M_3^* is equal to M_1 with a transformed endogenous variable. The interesting aspect about this auxiliary regression is the limiting case when $g \rightarrow 0$ and $Y(Y'Y)^{-1}Y'X(X'X)^{-1}X' \rightarrow I_N$ because it leads to M_1 with $\vec{f} = \vec{b}$ and $\vec{u} = \vec{e}$. This result is important because it is now easy to test if it is necessary to apply the transformation to the endogenous variable. The null hypothesis is simply $H_0 : g = 0$ versus $H_1 : g \neq 0$ which can be tested with the conventional t-test. In other words, if the null hypothesis is rejected then the transformation is necessary and the estimated parameters $\hat{\vec{c}}$ and $\hat{\vec{f}}$ interpretable. The usefulness of this modeling approach is illustrated in the next subsection.

4.2 Empirical Evidence

The data that was collected from the implemented SLE (as described in section 2.3) contained the following exogenous variables:

- Bcount: actual computations
- Gender: 0 = female / 1 = male
- Future: intention to use
- Pop: 0 = Bachelor / 1 = Switching
- nnzfg: actually submitted feedback messages in peer review
- Reflection: reported feedback messages in peer review

Table 1 presents the empirical results of two models (M_1 and M_3). The endogenous variable in M_1 is the sum of all exam questions with unit weights whereas M_3 is based on objective exam score transformations (optimal weights of individual questions).

Table 1: Empirical results

Variable	Estimate M_1	sig.	Estimate M_3	sig.
(Intercept)	6.935987	*	6.333557	***
Bcount	0.033281		0.035939	***
Gender	-2.166419		-1.465320	
Pop	-4.616769		-0.494553	
nnzfg	0.027379	*	0.030161	***
Future	0.625812	.	0.639711	***
Reflection	-0.167591		-0.167980	**
Bcount:Gender	-0.027786		-0.036090	**
Bcount:Pop	0.018510		-0.007901	
Gender:Pop	3.699211		2.116220	
Gender:nnzfg	-0.001449		-0.013074	**
Pop:nnzfg	-0.020088		-0.024768	***
Gender:Future	-0.274359		-0.354713	*
Pop:Future	-0.038318		-0.143013	
Gender:Reflection	0.161042		0.225622	*
Pop:Reflection	0.289011		0.160574	.
Bcount:Gender:Pop	0.019236		0.021735	
Gender:Pop:nnzfg	-0.002538		0.009896	.
Gender:Pop:Future	-0.248325		-0.157158	
Gender:Pop:Reflection	-0.128991		-0.160408	
Residual standard error	3.446		0.9593	
Degrees of freedom	179		179	
Adj. R-squared	0.1607		0.6626	
F-statistic	2.995	***	21.47	***

Signif. codes:
0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

From the results in Table 1 it is clear that M_3 provides - unlike M_1 - a lot of interesting information about the relationship between optimally weighted exam scores and the exogenous variables which are

under the control of the educator. The percentage of variance explained (adjusted R^2) in M_3 is more than 66% which allows us to make much better predictions than what is usually reported in - otherwise excellent - academic articles [25]. As explained before, the traditional F-test cannot be used to test the significance of the improvement. However, the auxiliary regression's null hypothesis $H_0 : g = 0$ is rejected even if an extremely low type I error is chosen (the p-value is 3.23×10^{-11}). This implies that M_3 performs significantly better and the objective exam score transformations are necessary. In addition, several diagnostic tests about the final model (M_3) are shown in Figure 4 - they indicate no statistical inadequacies.

The most interesting aspects of this analysis are the estimated parameters of M_3. With regard to quality control of the SLE the following conclusions can be made:

- There is a positive effect of performing reproducible, statistical computations (Bcount). This effect is significant at the 0.1% type I error level and cannot be measured without optimal weights (M_1). However, this effect is only relevant for female students because the parameter that is associated with Bcount:Gender is also significant and has a negative sign.

- Submitting feedback messages (in peer review) is very beneficial and improves exam scores (p-value $< 0.01\%$). This effect is about twice as large for female students than for males (the Gender:nnzfg parameter partially offsets the effect for male students). In addition, students from the switching population benefit less from feedback submissions.

- The reported "intention to use" (as measured in the usability survey) positively affects exam scores. This effect is strongest for female students. Note that previous research has shown that intention is mainly related to student's perception about the comparative advantage (of the software system) to learn statistics as compared to other alternatives (such as textbooks) [27].

- Females who report a high number of submitted feedback messages have significantly lower exam scores. On the other hand, male students who exaggerate their efforts are not in danger of having lower exam scores. This implies that the female exaggeration bias is small but harmful - the male exaggeration bias is big and harmless.

Based on these empirical results it is now possible to control (improve) the quality of the SLE:

- Female students should be encouraged to generate more reproducible computations.

- Peer review (based on Reproducible Computing) is highly beneficial to learn statistics - especially when it requires students to engage in submitting feedback messages to their peers. Male students need to (at least) double their efforts (compared to females) in order to obtain the same effect. Students from the switching population also need more feedback submissions than bachelor students.

- It is important to explain the SLE to students - emphasizing the comparative advantages of the system and the potentially improved exam scores. However, male students need more (or better) arguments before they accept the new technology and exhibit an increased degree of "intention to use."

- Female students who exaggerate their reported efforts should receive accurate feedback about their real performance which is based on objective measurements. Self assessment and reflection about student's actual efforts (as compared to perceived efforts) should be an integral part of the SLE.

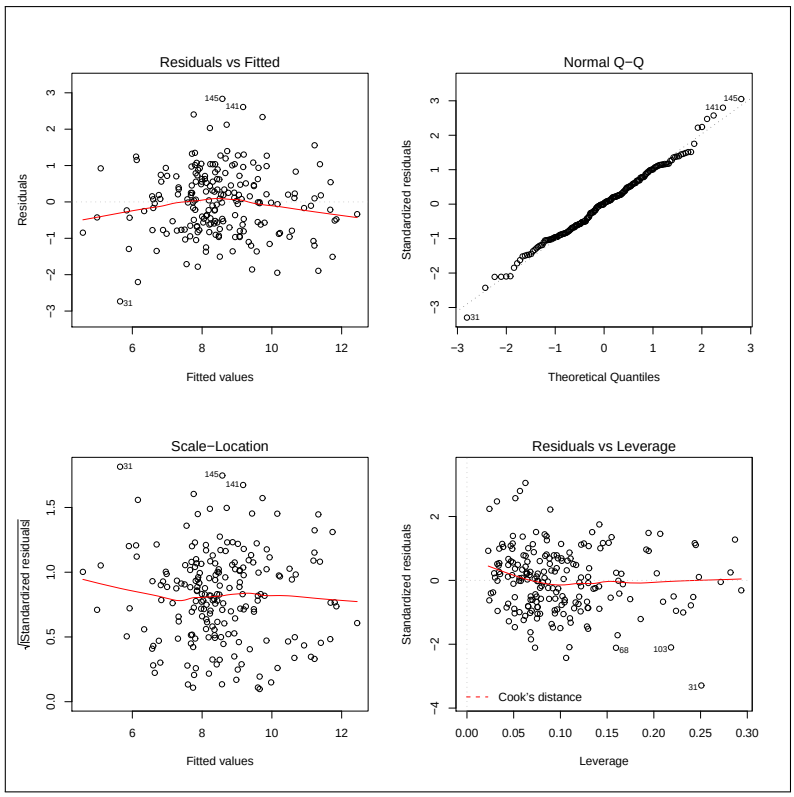


Figure 4: Diagnostics of M_3

5 Summary and Conclusions

The good news is that we now have a technology and methodology to assess actual and reported learning activities for any student population that makes use of the new compendium technology. Ultimately, this allows us to take control and improve the SLE which includes the e-learning environment, the statistical software, the course materials, and the overall learning experiences of all students.

Bibliography

[1] Arndt S., Turvey C., Andreasen N. (1999), Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's tau correlation, *Journal of Psychiatric Research*, 33, 97-104

[2] Arndt S., Magnotta V. (2001), Generating random series with known values of Kendall's tau, *Computer Methods and Programs in Biomedicine*, 65, 17-23

[3] Attitudes to Thinking and Learning Survey, (n.d.), Retrieved December 22, 2004, from www.moodle.org

[4] Benson J. (1989). Structural components of statistical test anxiety in adults: An exploratory study, *Journal of Experimental Education*, 57, 247-261.

[5] Buckheit J., and Donoho D. L. (1995), *Wavelets and Statistics*, Springer-Verlag, Editor: Antoniadis, A.

[6] Chambers J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole.

[7] Chen Z. (2008), Learning about Learners: System Learning in Virtual Learning Environment, *International Journal of Computers, Communications & Control*, Vol. III, No. 1, pp. 33-40

[8] Constructivist On-Line Learning Environment Survey, (n.d.), Retrieved December 22, 2004, from www.moodle.org

[9] de Leeuw J. (2001), Reproducible Research: the Bottom Line, *Department of Statistics Papers*, 2001031101, Department of Statistics, UCLA., URL http://repositories.cdlib.org/uclastat/papers/2001031101

[10] Donoho D. L., and Huo, X. (2005), BeamLab and Reproducible Research, *International Journal of Wavelets, Multiresolution and Information Processing*, 2(4), 391-414

[11] Eggen P., and Kauchak, D. (2001), *Educational Psychology: Windows on Classrooms (5th ed.)*, Upper Saddle River, NJ: Prentice Hall.

[12] Galotti K. M., Clinchy B. M., Ainsworth K., Lavin B. and Mansfield A. F. (1999), A new way of assessing ways of knowing: the attitudes towards thinking and learning survey (ATTLS), *Sex roles*, 40(9/10) p745-766

[13] Gentleman R. (2005), Applying Reproducible Research in Scientific Discovery, BioSilico, URL http://gentleman.fhcrc.org/Fld-talks/RGRepRes.pdf

[14] Green P. J. (2003), Diversities of gifts, but the same spirit, *The Statistician*, 52(4), 423-438

[15] Hilton S., Schau C., Olsen J. (2004), Survey of Attitudes Toward Statistics: Factor Structure Invariance by Gender and by Administration Time, *Structural Equation Modeling*, 11(1)

[16] Hollander M., and Wolfe D. A. (1973), *Nonparametric statistical inference*, New York: John Wiley & Sons., 185-194 (Kendall and Spearman tests).

[17] Kay R. H. (2008), Exploring the relationship between emotions and the acquisition of computer knowledge, *Computers & Education*, 50, 1269-1283

[18] Koenker R., and Zeileis A.(2007), A., Reproducible Econometric Research (A Critical Review of the State of the Art), *Research Report Series*, Department of Statistics and Mathematics Wirtschaftsuniversit Wien

[19] Leisch F. (2003), Sweave and beyond: Computations on text documents, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*

[20] Lewis J. R. (1993), IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use, IBM Corporation, *Technical Report* 54.786

[21] Lucy D., Aykroyd R. G. and Pollard A. M.(2002), Non-parametric calibration for age estimation, *Applied Statistics* 51(2), 183-196

[22] Meelissen M. R. M., Drent M. (2008), Gender differences in computer attitudes: Does the school matter?, *Computers in Human Behavior*, 24, 969-985

[23] Miller J. B., (n.d.), Examining the interplay between constructivism and different learning styles, Retrieved October 20, 2005 from http://www.stat.auckland.ac.nz/ĩase/publications/1/8a4_mill.pdf

[24] Mvududu N. (2003), A Cross-Cultural Study of the Connection Between Students' Attitudes Toward Statistics and the Use of Constructivist Strategies in the Course, *Journal of Statistics Education*, 11(3)

[25] O'Dwyer L. M., Russell M., Bebell D., Seeley K. (2008), Examining the Relationship between Students Mathematics Test Scores and Computer Use at Home and at School, *Journal of Technology, Learning, and Assessment*, 6 (5)

[26] Peng R. D., Dominici F., and Zeger S. L. (2006), Reproducible Epidemiologic Research, *American Journal of Epidemiology*, 163(9), 783-789

[27] Poelmans S., Wessa P., Milis K., Bloemen E., and Doom C. (2008), Usability and Acceptance of E-Learning in Statistics Education, based on the Compendium Platform, *Proceedings of the International Conference of Education, Research and Innovation*, International Association of Technology, Education and Development

[28] Romero C., Ventura S., Garcia E. (2008), Data mining in course management systems: Moodle case study and tutorial, *Computers & Education*, 51, 368-384

[29] Schwab M., Karrenbach N., and Claerbout J. (2000), Making scientific computations reproducible, *Computing in Science & Engineering* 2(6), 61-67

[30] Smith E. (1999), Social Constructivism, Individual Constructivism and the Role of Computers in Mathematics Education, *Journal of mathematical behavior*, 17(4)

[31] Statistical Computations at FreeStatistics.org (2008a), Office for Research Development and Education, Retrieved Mon, 30 Jun 2008, URL http://www.freestatistics.org/blog/date/2008/Jun/30/t1214840420q0fyankop4x9ebf.htm

[32] Statistical Computations at FreeStatistics.org (2008b), Office for Research Development and Education, Retrieved Mon, 30 Jun 2008, URL http://www.freestatistics.org/blog/date/2008/Jun/30/t12148409608o0dnj2k4s04jil.htm

[33] Statistical Computations at FreeStatistics.org (2008c), Office for Research Development and Education, Retrieved Mon, 30 Jun 2008, URL http://www.freestatistics.org/blog/date/2008/Jun/30/t1214841152sn6jlyhgseclgqm.htm

[34] Sun P., Tsai R. J., Finger G., Chen Y., Yeh D. (2008), What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction, *Computers & Education*, 50, 1183-1202

[35] Von Glasersfeld E. (1987), Learning as a Constructive Activity, *Problems of Representation in the Teaching and Learning of Mathematics*, Hillsdale, NJ: Lawrence Erlbaum Associates, 3-17.

[36] Wessa P. (2008a), A framework for statistical software development, maintenance, and publishing within an open-access business model, *Computational Statistics*, www.springerlink.com (DOI 10.1007/s00180-008-0107-y)

[37] Wessa P. (2008b), Learning Statistics based on the Compendium and Reproducible Computing, *Proceedings of the International Conference on Education and Information Technology*, Berkeley, San Francisco, USA

[38] Wessa P. (2008c), How Reproducible Research Leads to Non-Rote Learning Within a Socially Constructivist E-Learning Environment, *Proceedings of the 7th European Conference on e-Learning*, Cyprus

[39] Wessa P. (2008d), Measurement and Control of Statistics Learning Processes based on Constructivist Feedback and Reproducible Computing, *Proceedings of the 3rd International Conference on Virtual Learning*, Constanta, Romania

[40] Wessa P. (2009a), Discovering Computer-Assisted Learning Processes based on Objective Exam Score Transformations, *Proceedings of the World Congress on Educational Sciences*, Cyprus

Author index

Description

International Journal of Computers, Communications & Control (IJCCC) is a quarterly peer-reviewed publication started in 2006 by Agora University Editing House - CCC Publications, Oradea, ROMANIA.

Beginning with 2007, EBSCO Publishing is a licensed partner of IJCCC Publisher.

Every issue is published in online format (ISSN 1841-9844) and print format (ISSN 1841-9836).

Now we offer free online access to the full text of all published papers.

The printed version of the journal should be ordered, by subscription, and will be delivered by regular mail.

IJCCC is directed to the international communities of scientific researchers from the 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

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).

The articles submitted to IJCCC must be original and previously unpublished in other journals. The submissions will be revised independently by minimum two reviewers and will be published only after end of the editorial workflow.

The peer-review process is single blinded: the reviewers know who the authors of the manuscript are, but the authors do not have access to the information of who the peer-reviewers are.

IJCCC also publishes:

- papers dedicated to the works and life of some remarkable personalities;
- reviews of some recent important published books

Also, IJCCC will publish as supplementary issues the proceedings of some international conferences or symposiums on Computers, Communications and Control, scientific events that have reviewers and program committee.

The authors are kindly asked to observe the rules for typesetting and submitting described in Instructions for Authors.

Thomson Reuters Subject Category of IJCCC:

AUTOMATION & CONTROL SYSTEMS

Category Description: Automation & Control Systems covers resources on the design and development of processes and systems that minimize the necessity of human intervention. Resources in this category cover control theory, control engineering, and laboratory and manufacturing automation.

COMPUTER SCIENCE, INFORMATION SYSTEMS

Category Description: Computer Science, Information Systems covers resources that focus on the acquisition, processing, storage, management, and dissemination of electronic information that can be read by humans, machines, or both. This category also includes resources for telecommunications systems and discipline-specific subjects such as medical informatics, chemical information processing systems, geographical information systems, and some library science.

Editorial Workflow

The editorial workflow is performed using the online Submission System.

The peer-review process is single blinded: the reviewers know who the authors of the manuscript are, but the authors do not have access to the information of who the peer-reviewers are.

The following is the editorial workflow that every manuscript submitted to the IJCCC during the course of the peer-review process.

Every IJCCC submitted manuscript is inspected by the Editor-in-Chief/Associate Editor-in-Chief. If the Editor-in-Chief/Associate Editor-in-Chief determines that the manuscript is not of sufficient quality to go through the normal review process or if the subject of the manuscript is not appropriate to the journal scope, Editor-in-Chief/Associate Editor-in-Chief *rejects the manuscript with no further processing*.

If the Editor-in-Chief/Associate Editor-in-Chief determines that the submitted manuscript is of sufficient quality and falls within the scope of the journal, he sends the manuscript to the IJCCC Executive Editor/Associate Executive Editor, who manages the peer-review process for the manuscript.

The Executive Editor/Associate Executive Editor can decide, after inspecting the submitted manuscript, that it should be rejected without further processing. Otherwise, the Executive Editor/Associate Executive Editor assigns the manuscript to the one of Associate Editors.

The Associate Editor can decide, after inspecting the submitted manuscript, that it should be rejected without further processing. Otherwise, the Associate Editor assigns the manuscript to minimum two external reviewers for peer-review. These external reviewers may or may not be from the list of potential reviewers of IJCCC database.

The reviewers submit their reports on the manuscripts along with their recommendation of one of the following actions to the Associate Editor: Publish Unaltered; *Publish after Minor Changes; Review Again after Major Changes; Reject* (Manuscript is flawed or not sufficiently novel).

When all reviewers have submitted their reports, the Associate Editor can make one of the following editorial recommendations to the Executive Editor: Publish Unaltered; Publish after Minor Changes; Review Again after Major Changes; Reject.

If the Associate Editor recommends *"Publish Unaltered"*, the Executive Editor/Associate Executive Editor is notified so he/she can inspect the manuscript and the review reports. The Executive Editor/Associate Executive Editor can either override the Associate Editor's recommendation in which case the manuscript is rejected or approve the Associate Editor's recommendation in which case the manuscript is accepted for publication.

If the Associate Editor recommends "Review Again after Minor Changes", the Executive Editor/Associate Executive Editor is notified of the recommendation so he/she can inspect the manuscript and the review reports.

If the Executive Editor/Associate Executive Editor overrides the Associate Editor's recommendation, the manuscript is rejected. If the Executive Editor approves the Associate Editor's recommendation, the authors are notified to prepare and submit a final copy of their manuscript with the required minor changes suggested by the reviewers. Only the Associate Editor, and not the external reviewers, reviews the revised manuscript after the minor changes have been made by the authors. Once the Associate Editor is satisfied with the final manuscript, the manuscript can be accepted.

If the Associate Editor recommends *"Review Again after Major Changes"*, the recommendation is communicated to the authors. The authors are expected to revise their manuscripts in accordance with the changes recommended by the reviewers and to submit their revised manuscript in a timely manner. Once the revised manuscript is submitted, the original reviewers are contacted with a request to review the revised version of the manuscript. Along with their review reports on the revised manuscript, the reviewers make a recommendation which can be "Publish Unaltered" or "Publish after Minor Changes" or "Reject". The Associate Editor can then make an editorial recommendation which can be "Publish Unaltered" or "Review Again after Minor Changes" or "Reject".

If the Associate Editor recommends rejecting the manuscript, either after the first or the second round of reviews, the rejection is immediate.

Only the Associate Editor-in-Chief can approve a manuscript for publication, where Executive Editor/Associate Executive Editor recommends manuscripts for acceptance to the Editor-in-Chief/Associate Editor-in-Chief.

Finally, recommendation of acceptance, proposed by the Associate Editor Chief, has to be approved by the Editor-in-Chief before publication.

Instructions for authors

Concurrent/Duplicate Submission

Submissions to IJCCC must represent original material.

Papers are accepted for review with the understanding that the same work has been neither submitted to, nor published in, another journal or conference. If it is determined that a paper has already appeared in anything more than a conference proceeding, or appears in or will appear in any other publication before the editorial process at IJCCC is completed, the paper will be automatically rejected.

Papers previously published in conference proceedings, digests, preprints, or records are eligible for consideration provided that the papers have undergone substantial revision, and that the author informs the IJCCC editor at the time of submission.

Concurrent submission to IJCCC and other publications is viewed as a serious breach of ethics and, if detected, will result in immediate rejection of the submission.

Preliminary/Conference Version(s)

If any portion of your submission has previously appeared in or will appear in a conference proceeding, you should notify us at the time of submitting, make sure that the submission references the conference publication, and supply a copy of the conference version(s) to our office. Please also provide a brief description of the differences between the submitted manuscript and the preliminary version(s).

Please be aware that editors and reviewers are required to check the submitted manuscript to determine whether a sufficient amount of new material has been added to warrant publication in IJCCC. If you have used your own previously published material as a basis for a new submission, then you are required to cite the previous work(s) and clearly indicate how the new submission offers substantively novel or different contributions beyond those of the previously published work(s). Any manuscript not meeting these criteria will be rejected. Copies of any previously published work affiliated with the new submission must also be included as supportive documentation upon submission.

Manuscript Preparing/Submission

The papers must be prepared using a LaTeX typesetting system. A template for preparing the papers is available on the journal website. In the template.tex file you will find instructions that will help you prepare the source file. Please, read carefully those instructions. (We are using MiKTeX 2.7).

Any graphics or pictures must be saved in Encapsulated PostScript (.eps) format.

Papers must be submitted electronically to the following email address: ccc.journal@gmail.com. You should send us the LaTeX source file (just one file - do not use bib files) and the graphics in a separate folder. You must send us also the pdf version of your paper.

The maximum number of pages of one article is 20. The publishing of a 12 page article is free of charge (including a bio-sketch). For each supplementary page there is a fee of 50 Euro/page that must

be paid after receiving the acceptance for publication. The authors do not receive a print copy of the journal/paper, but the authors receive by email a copy of published paper in pdf format.

The papers must be written in English. The first page of the paper must contain title of the paper, name of author(s), an abstract of about 300 words and 3-5 keywords. The name, affiliation (institution and department), regular mailing address and email of the author(s) should be filled in at the end of the paper. Manuscripts must be accompanied by a signed copyright transfer form. The copyright transfer form is available on the journal website.

Please note: We will appreciate if the authors writes their own final version of the paper in LaTeX. But if the authors have difficulties with LaTeX and wish to send us their manuscript in Microsoft Word, the technical secretariat can do the transcription of the document. In this case, the paper can be sent in MS Word format with the following specifications: paper A4, font TNR 12p, single column. The graphics will be placed in the document but it has to be also attached separately in jpeg format.

Checklist:

1. Completed copyright transfer form.
2. Source (input) files.
    - One LaTeX file for the text.
    - EPS files for figures in a separate folder.
3. Final PDF file (for reference).

Order

If you are interested in having a subscription to "Journal of Computers, Communications and Control", please fill in and send us the order form below:

ORDER FORM		
I wish to receive a subscription to "Journal of Computers, Communications and Control"		
NAME AND SURNAME:		
Company:		
Number of subscription:	Price Euro	for issues yearly (4 number/year)
ADDRESS: City: Zip code: Country: Fax: Telephone: E-mail:		
Notes for Editors (optional)		

1. Standard Subscription Rates for Romania (4 issues/year, more than 400 pages, including domestic postal cost): 200 EURO.

2. Standard Subscription Rates for EU member countries (4 issues/year, more than 400 pages, including international postal cost): 360 EURO.

3. Standard Subscription Rates for other countries (4 issues/year, more than 400 pages, including domestic postal cost): 450 US dollars

For payment subscription rates please use following data:

HOLDER: Fundatia Agora, CUI: 12613360
BANK: BANK LEUMI ORADEA
BANK ADDRESS: Piata Unirii nr. 2-4, Oradea, ROMANIA
IBAN ACCOUNT for EURO: RO02DAFB1041041A4767EU01
IBAN ACCOUNT for LEI/ RON: RO45DAFB1041041A4767RO01
SWIFT CODE (eq. BIC): DAFBRO22

Mention, please, on the payment form that the fee is "for IJCCC".

EDITORIAL ADDRESS:

CCC Publications
Piata Tineretului nr. 8
ORADEA, jud. BIHOR
ROMANIA
Zip Code 410526
Tel.: +40 259 427 398
Fax: +40 259 434 925
E-mail: ccc@univagora.ro, Website: www.journal.univagora.ro

Copyright Transfer Form

To The Publisher of the International Journal of Computers, Communications & Control

This form refers to the manuscript of the paper having the title and the authors as below:

The Title of Paper (hereinafter, "Paper"):

.....

The Author(s):

.....

.....

.....

.....

The undersigned Author(s) of the above mentioned Paper here by transfer any and all copyright-rights in and to The Paper to The Publisher. The Author(s) warrants that The Paper is based on their original work and that the undersigned has the power and authority to make and execute this assignment. It is the author's responsibility to obtain written permission to quote material that has been previously published in any form. The Publisher recognizes the retained rights noted below and grants to the above authors and employers for whom the work performed royalty-free permission to reuse their materials below. Authors may reuse all or portions of the above Paper in other works, excepting the publication of the paper in the same form. Authors may reproduce or authorize others to reproduce the above Paper for the Author's personal use or for internal company use, provided that the source and The Publisher copyright notice are mentioned, that the copies are not used in any way that implies The Publisher endorsement of a product or service of an employer, and that the copies are not offered for sale as such. Authors are permitted to grant third party requests for reprinting, republishing or other types of reuse. The Authors may make limited distribution of all or portions of the above Paper prior to publication if they inform The Publisher of the nature and extent of such limited distribution prior there to. Authors retain all proprietary rights in any process, procedure, or article of manufacture described in The Paper. This agreement becomes null and void if and only if the above paper is not accepted and published by The Publisher, or is withdrawn by the author(s) before acceptance by the Publisher.

Authorized Signature (or representative, for ALL AUTHORS): .....

Signature of the Employer for whom work was done, if any: .....

Date: .....

Third Party(ies) Signature(s) (if necessary): .....