ITHEA

International Journal INFORMATION THEORIES & APPLICATIONS

2011 Volume 18 Number 3

International Journal
INFORMATION THEORIES & APPLICATIONS
Volume 18 / 2011, Number 3

International Journal "INFORMATION THEORIES & APPLICATIONS" (IJ ITA)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITA welcomes scientific papers connected with any information theory or its application.
IJ ITA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITA as well as the **subscription fees** are given on *www.foibg.com/ijita*.

Responsibility for papers published in IJ ITA belongs to authors.

General Sponsor of IJ ITA is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol. 18, Number 3, 2011

Printed in Bulgaria

Edited by the **Institute of Information Theories and Applications** FOI ITHEA, Bulgaria, in collaboration with:
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Institute of Mathematics and Informatics, BAS, Bulgaria,
Universidad Politécnica de Madrid, Spain.

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

ISSN 1310-0513 (printed) ISSN 1313-0463 (online) ISSN 1313-0498 (CD/DVD)

CROSS INTERSECTION SEQUEL OF DISCRETE ISOPERIMETRY[1]

Levon Aslanyan, Vilik Karakhanyan

***Abstract**: This work inspired by a specifically constrained communication model. Given collections of communicating objects, and communication is by means of several relay centres. The complete cross connectivity of elements of different collections is the target, supposing that communicating objects differ by their connections to the relay centres. Such models exist only for proper object groups – when they have specific sizes and there is a corresponding number of relay points. We consider optimization problems studying the validity boundaries. Terms are combinatorial – geometry of binary cube, lexicographical orders, shadowing and isoperimetry. The main interest is methodological and aims at extending the consequences that can be delivered from the solution of the well known discrete isoperimetry problem.*

***Keywords**: communication, optimization, isoperimetry.*

***ACM Classification Keywords**: G.2.1 Discrete mathematics: Combinatorics*

Introduction

Mass communication models with resource limitations require a proper design and analysis stage. Practical examples are populations with sophisticated communication means, wireless sensor models with requirements of connectivity, coverage and energy efficiency, and other ad hoc networks with different additional requirements. Resource limitations which appear everywhere need to be checked against the existence of a valid network, and when it is, - optimization that brings the resource minimization and the quality enhancement.

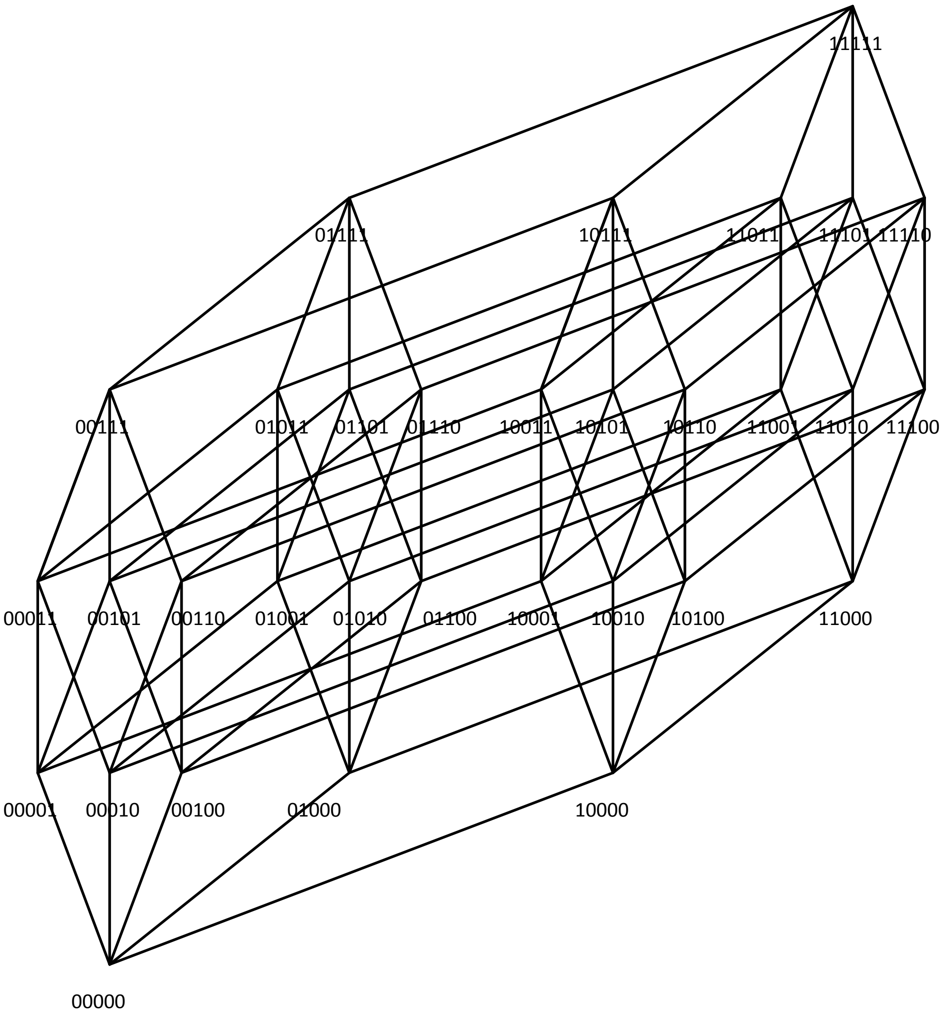
Communication model which we consider consists of several independent (none intersecting) societies $\Xi_1, \Xi_2, ..., \Xi_m$ whose elements are to be cross connected totally. This means that each pair $s_{i_1 p} \in \Xi_{i_1}$ and $s_{i_2 q} \in \Xi_{i_2}$, $i_1 \neq i_2$ is connected. Connection is through the set of n relay points $X_1, X_2, ..., X_n$. If object s is connected to the relay points $X_{j_1}, X_{j_2}, ..., X_{j_k}$ then we code it as the binary n-vector $\tilde{\alpha}(s) = (\alpha_1, \alpha_2, ..., \alpha_n)$ with coordinates $j_1, j_2, ..., j_k$ equal to 1, and all other coordinates - to 0. Now for objects $\tilde{\alpha}$ and $\tilde{\beta}$ connectivity means that they (their code vectors) intersect by their sets of 1 coordinates. Connectivity that we described requires all connections between the societies $\Xi_1, \Xi_2, ..., \Xi_m$. Inside the society we require that the binary codes defined for objects are all different. One more condition may require that objects are connected to a fixed number of relay points and this condition is also applied to our model.

[1] *Partially supported by the Grant **11-1b193** “Approximation algorithmic research on novel, open and hard combinatorial optimization problems” of State Committee of Science of Republic of Armenia.*

Given above - is a particular communication model for societies. There can be a broad diversity of models varying in requirements. For instance, one can require a balanced use of relay points. Diversity of object binary codes can be applied on the total integrated society, etc. But our goal is to see the optimization framework of models of this type. It is shown that combinatorial optimization appears as the instrument of design of such networks. The particular technique that appears is the binary n-cube geometry being linked to the fundamental results of combinatorial optimization in that area.

The Formal Model

Boolean domain: First formal model that we consider to interpret the cross intersection property is in terms of n-dimensional binary cube, Boolean functions and systems of Boolean functions. Let $E=\{0,1\}$. Cartesian degree E^n represents the set of all n- dimensional binary vectors $\tilde{e}=(e_1,e_2,...,e_n)$. As usual we define the weight of vector and the Hamming distance of a pair of vectors in E^n. Weight of $\tilde{e}$ is the number of its 1 coordinates. The Hamming distance $h(\tilde{e}_1,\tilde{e}_2)$ between $\tilde{e}_1$ and $\tilde{e}_2$ is the number of coordinates where these vectors are different. Consider the Hasse diagram of E^n. An example for E^5 is given below.



The diagram consists of $n+1$ layers $0,1,...,n$ placed vertically. Each layer is composed of the same weight vertices placed on that layer horizontally. k-th layer consists of C_n^k vertices denoted by E_k^n. Edges connect 2 neighbour vertices – those that have distance 1 (differing exactly in 1 coordinate). Vertices $\tilde{e}_1$ and $\tilde{e}_2$ are comparable, in particular $\tilde{e}_1 \prec \tilde{e}_2$, if all corresponding coordinate vice similar inequalities hold. In a special case when $\tilde{e}_1$ and $\tilde{e}_2$ are comparable and their distance equals 1, then they are connected by an edge in the diagram.

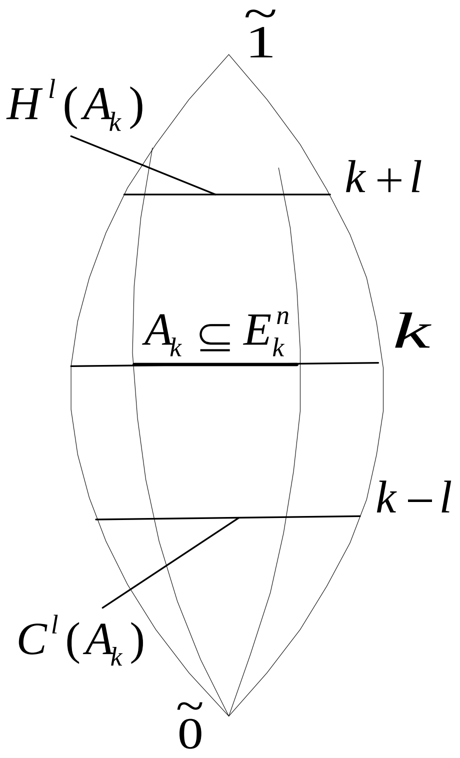
Consider the list of vertices of layer 2 in the order we see them on the diagram:

$$00011,00101,00110,01001,01010,01100,10001,10010,10100,11000.$$

It is easy to check that this is lexicographic sequence of all vertices of layer 2. In a similar way we may compose the lexicographic sequence of all n-dimensional words of weight k over the alphabet $\{0,1\}$. Here we suppose the usual precedence $0 \prec 1$. We denote this sequence as L_k^n, and let $L_k^n(\delta)$ is the initial δ-segment of L_k^n. And let T_k^n and $T_k^n(\delta)$ denote the reverse sequence to L_k^n and its initial δ-segment. In area of discrete isoperimetry it is common for T_k^n the term <u>standard placement</u>.

It is well known the specific (and unique in this form) decomposition of set $T_k^n(\delta)$ and the number δ itself in the following way:

$$\delta = C_{n-m_1}^{k-m_1} + C_{n-m_2}^{k-m_2+1} + ... + C_{n-m_r}^{k-m_r+r-1}, \text{ where } 1 \le m_1 < m_2 < ... < m_r < n. \quad (1)$$

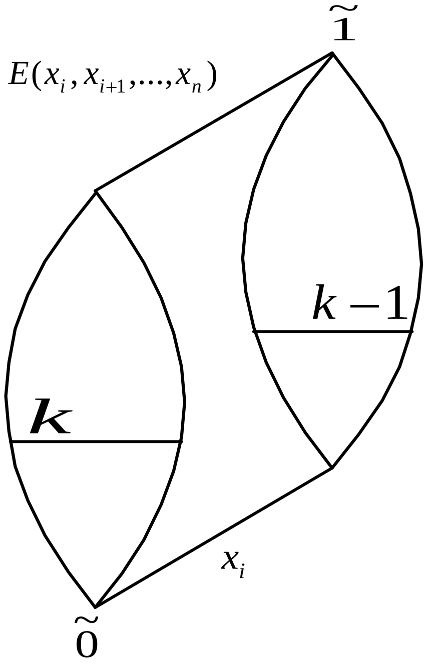


Consider arbitrary vertex subsets $A \subseteq E^n$. A_k denote the intersection $A \cap E_k^n$. Two type of concepts are introduced – internal (blocked) area $C^l(A_k)$, and bordering (shadow) area $H^l(A_k)$. $\underline{C^l(A_k)}$ is the set of all those vertices of layer $\underline{k-l}$ that are internal by the set of vertices of A_k. In other terms $\tilde{\alpha} \in C^l(A_k)$ iff <u>all</u> vertices of layer k comparable with $\tilde{\alpha}$ belong to A_k. $\underline{H^l(A_k)}$ belongs to layer $\underline{k+l}$ and consists of vertices $\tilde{\alpha}$ that <u>at least one</u> of elements of layer k comparable with $\tilde{\alpha}$ belongs to A_k. Below we suppose that $k \le n/2$. When after some transformations we receive subsets above the layer $n/2$, then $C^l(A_k)$ and $H^l(A_k)$ are in some sense bottom up constructions. In this case $C^l(A_k)$ belongs to layer $k+l$ and $H^l(A_k)$ is from layer $k-l$ (we refer to this by notion).

Consider the set $A_k = T_k^n(\delta)$. How this is related to the formula (1)? Consider series of splits in E^n. Coordinates $x_1, x_2, ..., x_n$ are applied consequently. Initially we are given the cube size n and δ vertices to be on layer k. Split of E^n by $x_1 = 0$ and $x_1 = 1$ brings two $n-1-$ subcubes. A_k (similarly any other subset of vertices of E^n) is split to one part on the layer k of subcube with $x_1 = 0$, a the reminder vertices of that are on layer k of (right) subcube with $x_1 = 1$. Depending on whether one of these parts completes the layer we choose the left or the right subcube for continuation.

Let after several splits that use coordinates $x_1, x_2, ..., x_i$ we face the situation that the $k-i$-th layer of right subcube is filled by elements of A_k. Here appears the first term C_{n-i}^{k-i} in formula (1). After that we continue splitting in left subcube, where we have the cube size $n-i$ and $\delta - C_{n-i}^{k-i}$ vertices to be on layer $k-i+1$.

Continuation of splitting that finally concludes the formula (1) is a series of splits similar to the case we considered.



Decomposition of δ given by formula (1) have two major properties. Explain them in terms of a particular step of split process. In a situation, when the right subcube become completed by δ reminder vertices, this part creates several new internal vertices on layer $k-i-l$, and several new bordering vertices on layer $k-i+l$ (in global cube E^n, not the splitted ones, these are layers $k-l$ and $k+l$). Internal a bordering vertices created during this step are non intersecting with the ones created during the previous splitting steps.

So the following formulas are consequences of (1).

First we suppose the case $l = 1$. Introduce the formulas:

$$c_{k-1}(\delta) = c_{k-1}^1(\delta) = C_{n-m_1}^{k-m_1-1} + C_{n-m_2}^{k-m_2} + ... + C_{n-m_r}^{k-m_r+r-2}$$

$$h_{k+1}(\delta) = h_{k+1}^1(\delta) = C_{n-m_1}^{k-m_1+1} + C_{n-m_2}^{k-m_2+2} + ... + C_{n-m_r}^{k-m_r+r} \qquad (2)$$

For l general we have:

$$c_{k-l}^l(\delta) = c_{k-l}(c_{k-l+1}(...(c_{k-1}(\delta)))) = C_{n-m_1}^{k-m_1-l} + C_{n-m_2}^{k-m_2-l+1} + ... + C_{n-m_r}^{k-m_r-l+r-1}$$

$$h_{k+l}^l(\delta) = h_{k+l}(h_{k+l-1}(...(h_{k+1}(\delta)))) = C_{n-m_1}^{k-m_1+l} + C_{n-m_2}^{k-m_2+l-1} + ... + C_{n-m_r}^{k-m_r+l-r+1} \qquad (3)$$

Here we define the base isoperimetry problem, give its one, very easy solution, and mention some consequences that we will use. Vertex $\tilde{\alpha} \in A \subseteq E^n$ is interior, if all 1 distance vertices from $\tilde{\alpha}$ belong to A. In general, $S_r^n(\tilde{\alpha})$ denotes the sphere of radius r with center $\tilde{\alpha}$. So $\tilde{\alpha}$ is interior in A, if $S_1^n(\tilde{\alpha}) \subseteq A$. $\mathrm{I}nt(A)$ will denote the set of all points interior in A. The reminder vertices $A \setminus \mathrm{I}nt(A)$ of A we call boundary vertices. The base discrete isoperimetry problem (DIP) by a given size $a,\ 0 \le a \le 2^n$ is seeking for subsets A so that

$$\left|\mathrm{I}nt(A)\right| \ge \left|\mathrm{I}nt(B)\right|_{B\subseteq E^n, |B|=a} .$$

Theorem 1. $S_{k-1}^n(\tilde{0}) \cup T_k^n(\delta)$ is a DIP solution for $a = \sum_{i=0}^{k-1} C_n^i + \delta,\ \delta < C_n^k$.

Consequence 1. If $A \subseteq E_k^n$, $|A| = \delta,\ \delta < C_n^k$ then

$$\left|C^l(A)\right| \le \left|C^l(T_k^n(\delta))\right| = c_{k-l}^l(\delta) .$$

Consequence 2 (Kruscal-Katona theorem). If $A \subseteq E_k^n$, $|A| = \delta,\ \delta < C_n^k$ then

$$\left|H^l(A)\right| \ge \left|H^l(T_k^n(\delta))\right| = h_{k+l}^l(\delta) .$$

Given is the basic knowledge that we need from the discrete isoperimetry area. We may use several extensions of results but we prefer to stay on basic postulations to be more or less transparent and understandable.

We may also use an equivalent terminology given in terms of Boolean functions. n-dimensional Boolean function f accepts value 1 (true) in some subset $\Xi_f \subseteq E^n$. Denote $E^n \setminus \Xi_f$ by $\overline{\Xi}_f$ which is now the set of all 0 values of function f (and 1 vertices of the inversion of function f). Introduce spectral characteristics for function f as the sequence $t_0, t_1, ..., t_n$ of sizes of sets $\Xi_f[k] = E_k^n \cap \Xi_f$, $k = 0,1,...,n$.

In terms of cross connected network design we associate one Boolean function to one society. Ξ_f consists of all codes of object in one society f. These codes are all different. An object code belongs to the layer k means that it is connected to the k relay points. In a simplest case we suppose that only the sets $\Xi_f[k_0]$ for one fixed values of k are nontrivial (non zero). In cross societal connectivity problem we posted above a system $f_1, f_2, ..., f_m$ of Boolean functions is considered.

Set-theoretical domain: The set-theoretical framework related to the applied communication model defined above was introduced and studied firstly in [HIL,1977]. Here interpretation is as follows. Let $k, m \geq 1$. k is the number of links from objects to the set of n relays. m is the number of societies. Let

$$\Xi_1 = \{S_{11}, S_{12}, ..., S_{1t_1}\}$$

...

$$\Xi_m = \{S_{m1}, S_{m2}, ..., S_{mt_m}\}$$

be a list of collections of subsets (societies) with subsets of set $\{1,2,...,n\}$. Suppose that

Condition 1.

$|S_{ir}| = k \leq n/2$ for $1 \leq i \leq m$ and $1 \leq r \leq t_i$ /objects are linked to the same number k of relays/

$S_{ir_1} \neq S_{ir_2}$, $1 \leq i \leq m, 1 \leq r_1 < r_2 \leq t_i$ /objects in one society are different by their connections to relays/ (4)

$S_{i_1 r'} \cap S_{i_2 r''} \neq 0$ $1 \leq i_1 < i_2 \leq m, 1 \leq r' \leq t_{i_1}, 1 \leq r'' \leq t_{i_2}$ /objects from different societies are intersecting/.

Theorem 2. [HIL,1977] proves that under the Condition 1.

$$t_1 + t_2 + ... + t_m \leq \begin{cases} C_n^k \ if \ n/k \geq m, \\ mC_{n-1}^{k-1} \ if \ n/k \leq m. \end{cases} \tag{5}$$

Here $C_n^k = n/k \cdot C_{n-1}^{k-1}$ so that the maximum between the C_n^k and mC_{n-1}^{k-1} is correlated with maximum between the n/k and m.

The proof of this postulation is hard in [HIL,1977], with intensive formula manipulations. After the result was achieved, a series of publications appeared bringing more simple and transparent results. We aim to demonstrate that the most suitable technique for this research area is the discrete isoperimetry technique [ASL,1979]. This not only gives the numerical estimates but also explains the structural properties of cross connectivity collections.

Intersection – Isoperimetry Relations

Theorem 3. If a collection of sets $\{\Xi_1, \Xi_2, ..., \Xi_m\}$ of characteristics $\{t_1, t_2, ..., t_m\}$ exists under the Condition 1. then the same Condition 1. properties obeyed by the set $\{T_k^n(t_1), T_k^n(t_2), ..., T_k^n(t_m)\}$.

To prove this consider an induction on m. If $m=1$, it is simply evident that as the set Ξ_1 we can take initial fragment $T_k^n(t_1)$. This choice is valid due to $t_1 \le C_n^k$ by **Condition 1.**

Suppose that the theorem postulation is correct for arbitrary collections $\{\Xi_1, \Xi_2, ..., \Xi_{m'}\}$ of characteristics $\{t_1, t_2, ..., t_{m'}\}$ under the **Condition 1.**, when $m' < m$. Consider the proof for values $m, m \ge 2$.

Let $\{\Xi_1, \Xi_2, ..., \Xi_m\}$ be an arbitrary collection of subsets of connections of objects that satisfies **Condition 1.** Consider the sub-collection $\{\Xi_1, \Xi_2, ..., \Xi_{m-1}\}$ and construct in accord to this collection the collection of compliments/negations $\{\overline{\Xi}_1, \overline{\Xi}_2, ..., \overline{\Xi}_{m-1}\}$ of initial sub-collections, where $\overline{\Xi}_i = \{\overline{S}_{i1}, \overline{S}_{i2}, ..., \overline{S}_{it_i}\}$, $1 \le i \le m-1$. It is clear *that the relation* $\left|\overline{S}_{ij}\right| = n-k \ge n/2 \ge k$ holds so that during this negations all points are transferring from layer k to the layer $n-k$.

Consider the set $\breve{S} = \bigcup_{i=1}^{m-1} \Xi_i$ and let $\widehat{S} = \bigcup_{i=1}^{m-1} \overline{\Xi}_i$. Compose by this the set $H^l\left(\bigcup_{i=1}^{m-1} \overline{\Xi}_i\right)$ for the value $l = n-2k$. Recall that if $\widehat{S} \subseteq E_{n-k}^n$, $k \le n/2$ (**notion**), then $H^l(\widehat{S})$ consists of some elements of E_k^n, those that are covered by elements of set $\widehat{S}$. Hence we received that $H^l\left(\bigcup_{s \in \breve{S}} \overline{s}\right) \subseteq E_k^n$, with the general requirement that

$$C_n^k - \left| H^l\left(\bigcup_{s \in \breve{S}} \overline{s}\right) \right| \ge t_m . \tag{6}$$

The last requirement takes into account that sets $S_{mj_1}, 1 \le j_1 \le t_m$ doesn't belong (are not covered) to any $H^l\left(\overline{S}_{ij_2}\right), 1 \le i \le m-1, 1 \le j_2 \le t_i$ and so, they doesn't belong to the union of these sets. If an inclusion $S_{mj_1} \in H^l\left(\overline{S}_{ij_2}\right)$ holds for some $1 \le i \le m-1$, then we receive that the initial vectors are not intersecting, $S_{mj_1} \cap S_{ij_2} = 0$, which contradicts to the theorem conditions. This means that the set Ξ_m is to be out of $H^l\left(\bigcup_{s \in \breve{S}} \overline{s}\right)$ and the corresponding relation of sizes of these sets is introduced in formula (6).

By the induction suppositions, there exists a collection $\{\Xi'_1, \Xi'_2, ..., \Xi'_{m-1}\}$ of characteristics $\{t_1, t_2, ..., t_{m-1}\}$ that satisfy **Condition 1.** and that $\Xi'_i = T_k^n(t_i)$. It is easy to check that $\bigcup_{i=1}^{m-1} \Xi'_i = T_k^n(t)$ for some value $t = \max(t_1, t_2, ..., t_{m-1})$.

Now the set $\widehat{S} = \bigcup_{i=1}^{m-1} \overline{\Xi}_i$ is represented as a finite sequence $L_{(n-k)}^n(t)$ of the $n-k$ -th layer of E^n.

According to Consequence 2.

$$C_n^k - \left|H^l\left(T_{n-k}^n(t)\right)\right| \geq C_n^k - \left|H^l\left(\bigcup_{s\in\bar{S}} \bar{s}\right)\right| \geq t_m . \tag{7}$$

Moreover, each subset $S \in E_k^n$, not belonging to the set $H^l\left(\bigcup_{s\in\bar{S}} \bar{s}\right)$ intersects with some subset S_{ij}, $1 \leq i \leq m-1$, $1 \leq j \leq t_i$.

In addition, $H^l\left(\bigcup_{s\in\bar{S}} \bar{s}\right) = H^l\left(\bigcup_{i=1}^{m-1} L_{n-k}^n(t_i)\right) = L_{n-k}^n(t)$, - and the compliment of $L_k^n(t)$ in the layer E_k^n is some $T_k^n(t')$. Now, constructing the proper $\{\Xi_1', \Xi_2', ..., \Xi_m'\}$ it is enough to take the Ξ_m' as the set $T_k^n(t_m)$ taking into account the proven relation $t' \geq t_m$. This proves the Theorem 3.

Denote by $R(m)$ the number of all those vectors $t_1 t_2, ..., t_m$ which correspond to some sets $\{\Xi_1, \Xi_2, ..., \Xi_m\}$ as the characteristics and obey the Condition 1.

Theorem 4. A necessary and sufficient condition for existence of a collection $\{\Xi_1, \Xi_2, ..., \Xi_m\}$ of characteristics $t_1 \geq t_2 \geq ... \geq t_m$ with Condition 1. is the existence of a collection $\{\Omega_1, \Omega_2\}$ of characteristics (t_1, t_2) that accords Condition 1.

The necessity point of theorem postulation is evident. To prove the sufficiency suppose that we are give collections $\{\Omega_1, \Omega_2\}$ of k-subsets, and collections have sizes t_1 and t_2 with Condition 1, satisfied. Theorem 3. implies, that without loss of generality we may suppose that $\Omega_i = T_k^n(t_i), i = 1,2.$ Take $\Xi_i = T_k^n(t_i), i = 3, ..., m$ and prove that the resulting system $\{\Omega_1, \Omega_2, \Xi_3, ..., \Xi_m\}$ obeys Condition 1. According to construction of collections $\{\Omega_1, \Omega_2, \Xi_3, ..., \Xi_m\}$ we have that first 2 points of Condition 1. are satisfied. Then, each pair of elements from $\{\Omega_1, \Omega_2\}$ are proven intersecting. Similarly, elements of Ω_1 and Ω_2 are intersecting with elements of other sets because of sets Ω_i, $j \geq 3$ are subsets of Ω_2. Last to prove is that subsets from Ξ_{i_1} intersect with subsets from Ξ_{i_2}, when $3 \leq i_1 < i_2 \leq m$. This happens because of $\Xi_{i_1} \subseteq \Omega_1$ and $\Xi_{i_2} \subseteq \Omega_2$.

Now we combine (1), (3) and (7) to achieve a quantitative condition for cross intersections. Consider again values $t_1 \geq t_2 \geq ... \geq t_m$ and the formula (1) for value t_1:

$t_1 = C_{n-m_1}^{k-m_1} + C_{n-m_2}^{k-m_2+1} + ... + C_{n-m_r}^{k-m_r+r-1}$, where $1 \leq m_1 < m_2 < ... < m_r < n$.

Apply a simple transformation of parameters. Replace k by $n-k$ taking into account that $0 \le n-k \le n$ and that $C_n^{n-k} - C_n^k$. In a similar way replace r values m_i by $\mu_i = n - m_i$. Based on (1) and (3) the modified formulas appear as:

$$t_1 = C_{\mu_1}^{k} + C_{\mu_2}^{k-1} + \ldots + C_{\mu_r}^{k-r+1} \text{ and } h_{k+l}^{l}(t_1) = C_{\mu_1}^{k+l} + C_{\mu_2}^{k+l-1} + \ldots + C_{\mu_r}^{k+l-(r-1)}.$$

When t_1 points T_1 belong to the layer $n-k$ and we consider $H^l(T_1)$ downward by the layers (note), then without major changes of parameters we receive that $h_{n-k-l}^{l}(t_1) = C_{\mu_1}^{k+l} + C_{\mu_2}^{k+l-1} + \ldots + C_{\mu_r}^{k+l-(r-1)}$ for the value $l = n-2k$ applied.

This representation of t_1 above and the one in (3) is used to formulate a necessary and sufficient condition for existence of set collections under the Condition 1.

Theorem 5. *A necessary and sufficient condition of existence of collection* $\{\Xi_1, \Xi_2, \ldots, \Xi_m\}$ of an n-element set, with characteristics $t_1 \ge t_2 \ge \ldots \ge t_m$ and with Condition 1., is the relation

$t_2 \le C_n^k - C_{\mu_1}^{k+l} - C_{\mu_2}^{k+l-1} + \ldots + C_{\mu_p}^{k+l-(p-1)}$.

The theorem, initially, can be given in terms of a two set collection, $m = 2$, by the Theorem 4. To prove the postulation it is essential to know the real volume of points projected from the t_1 set $T_{n-k}^{n}(t_1)$ onto the layer k.

Suppose, that the desired pair $\{\Omega_1, \Omega_2\}$ with (t_1, t_2) exists. Then, by Theorem 1. $C_n^k - \left|H^l(\overline{\Omega}_1)\right| \ge t_2$.

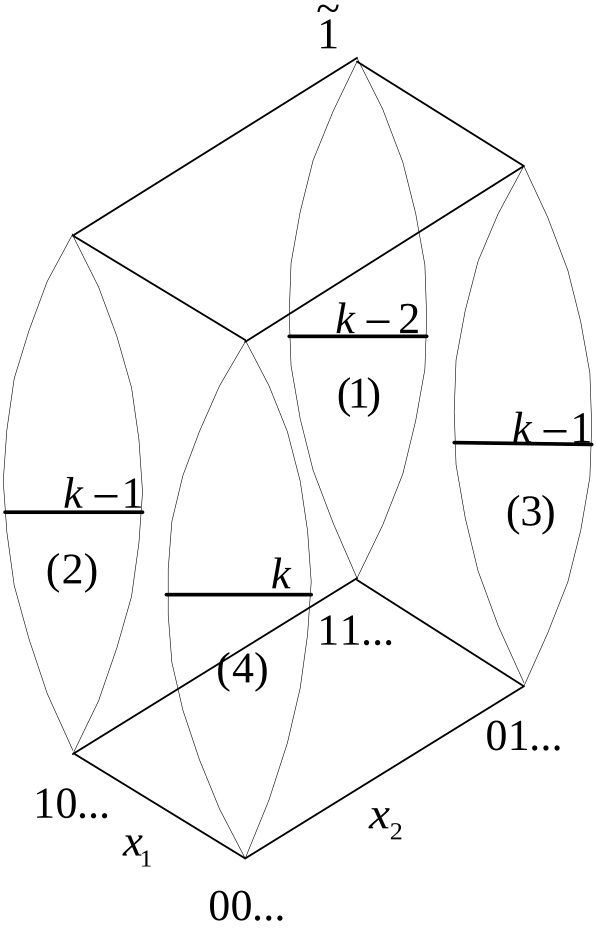
Write the Consequence 2. inequality for this case:

$$\left|H^l(\overline{\Omega}_1)\right| \ge C_n^k - C_{\mu_1}^{k+l} - C_{\mu_2}^{k+l-1} + \ldots + C_{\mu_p}^{k+l-(p-1)}$$

.

Further we come with this to the necessity point of theorem. Concluding the inequalities we receive the requirements

$t_2 \le C_n^k - C_{\mu_1}^{k+l} - C_{\mu_2}^{k+l-1} + \ldots + C_{\mu_p}^{k+l-(p-1)}$.



To consider the sufficiency part, and suppose that the last inequality is valid. Take $\Omega_1 = T_{n-k}^n(t_1)$. Now the possibility to choose a set Ω_2 that consists of t_2 elements and obeys the Condition 1. follows from the given inequality, taking into accounts considerations with complementary subsets, which was used regularly in the given descriptions.

Theorem 5 gives a technique to check the cross intersection for given (t_1, t_2). We can consider the problem of maximising the $t_1 + t_2$. Increasing t_1 points that we have on the layer $n-k$, and composing the shadow of set $T_{n-k}^n(t_1)$ to the layer k we may take all the reminder part as the set $L_k^n(t_2)$.

Maximum is when these quantities are approximately equal. Practically there is a very simple construction explaining this convergence. Consider the two dimension split of the unite cube. Increase t_1 and consider the corresponding set $L_k^n(t_1)$. Find for this the corresponding maximal value of t_2. $L_k^n(t_1)$ starts by the point from (1) continued then by (2), (3), and (4). When it is in area of (1), then t_2 maximum equals t_1 but still this is not the total maximum. The same postulation is also true for area (1)+(2). Here intersection is provided by the value $x_1 = 1$.

Consider the next to the (1)+(2) vertex $\tilde{\alpha}$. $\tilde{\alpha}$ starts with 01 followed by $k-1$ entities of 1, and then 0's. $\tilde{\alpha} = 01 \quad \underbrace{11...1}_{k-1} \quad 00...0$. Due to condition $k < n/2$ number of right 0's are not less than $k-1$ so that $k-1$ 1's can be shifted right without an intersection by the initial set of 1 coordinates. Do this shift, and replace first 2 coordinates by 10. We receive a vertex which belongs to (2) $\tilde{\beta} = 10 \quad 00...0 \quad \underbrace{11...1}_{k-1}$. It is easy to check that $\tilde{\alpha}$ and $\tilde{\beta}$ are non-intersecting vertices which proves that $t_1 \le C_{n-1}^{k-1}$ when $m \ge 2$. Further increase of Ω_1 leads to a similar decrease of Ω_2. This construction appears in Lemma 2.2 of [5] proving the inequality indicated in (5).

Conclusion

The set theoretical issue of complete cross intersecting set systems is considered. This is one of cases of applied societies' connectivity model but variations of models are possible and their analysis come to similar set theoretical optimizations. The paper can be characterised as methodological and it continues the line of possible application of Discrete Isoperimetry Property started at [5]. [8] represents another important application in area of search engines. The cross intersection topic itself is not yet expired. Extensions to consider include nodes with different number of connections to relays. Instead of artificial requirement that nodes are different by their connections to relays, more realistic models are required. Optimization of use of relays and their balancing can be studied which can bring natural conditions of nodes to be different inside the societies. The proof technique will move from the flat layer k consideration to the space constructions in entire cube E^n.

Bibliography

1. Ahlswede R., Katona G.O.H. Contributions to the Geometry of Hamming Spaces, Discr. Math., 17, No.1(1977), 1–22.
2. Ahlswede R.,Khachatrian L.H., The complete intersection theorem for systems of finite sets, European J. Combin. 18 (1997) 125–136.
3. Ahlswede R., Khachatrian L.H., The complete nontrivial-intersection theorem for systems of finite sets, J. Combin. Theory Ser. A 76 (1996) 121–138.
4. Aslanyan L.A., Karakhanyan V.M., Torosyan B.E. A Solution of the Discrete Isoperimetric Problem (in Russian), Doklady AN Arm. SSR, LXV, No.5(1977), 257–262.
5. Aslanyan L.A. An Isoperimetric Problem and Related Extremal Problems for Discrete Spaces (in Russian), in Problemy Kibernetiki 36, Nauka, Moscow 1979, 85–127.
6. Aslanyan L.A. The Discrete Isoperimetric Problem - Asymptotic Case (in Russian), Doklady AN Arm. SSR, LXXIV, No.3(1982), 99–103.
7. Aslanyan L.A., Akopova I.A. On the Distribution of the Number of Interior Points in Subsets of the n-Dimensional Unit Cube, in 37. Finite and Infinite Sets, Eger (Hungary) (1981), 47–58.
8. Aslanyan L., Metric decompositions and the Discrete Isoperimetry, IFAC Symposium on Manufacturing, Modelling, Management and Control, July 12-14, Patras, Greece, pp. 433-438, 2000.
9. Aslanyan L. and Castellanos J., Logic based Pattern Recognition - Ontology content (1), Information Theories and Applications, ISSN 1310-0513, Sofia, Vol. 14, N. 3, pp. 206-210, 2007.
10. Aslanyan L. and Ryazanov V., Logic Based Pattern Recognition - Ontology Content (2), Information Theories and Applications, ISSN 1310-0513, Sofia, Vol. 15, N. 4, pp. 314-318, 2008.
11. P. Erd˝os, C. Ko, R. Rado, Intersection theorems for systems of finite sets, Q. J. Math. 12 (1961) 313–320.
12. Garey, M. R. and Johnson, D. S. (1979). Computers and Intractibility: A Guide to the Theory of NP-Completeness. W. H. Freeman and Company, New York.
13. Harper L.H. Optimal Assignment of Numbers to Vertices, J. Sos. Ind. Appl. Math, 12, No.1(1964), 131–135.
14. Harper L.H. Optimal Numberings and Isoperimetric Problems on Graphs, J. Comb. Theory, 1, No.3(1966), 385–393.
15. Hilton A.J.W. An intersection theorem for a collection of families of subsets of a finite set, J. London Math. Soc. (2), 15 (1977), 369-376.
16. Jablonski S.V. and Lupanov O.B. (editors), Discrete Mathematics and Mathematical Problems of Cybernetics, vol. 1, (in Russian), Nauka, Moscow, 1974.
17. Katona G.O.H. The Hamming-Sphere Has Minimum Boundary, Studia Scient. Math. Hungarica, 10(1975), 131–140.
18. Kleitman D.J. Extremal Hypergraph Problems, in London Math. Soc. Lect. Note Ser., Cambridge Univ. Press Cambridge, No.38(1979), 44–65.
19. Korte B., Jens V., Combinatorial optimization: Theory & algorithms (21 Algorithms and Combinatorics), 4th Ed., Springer, 2008, 627p.
20. Kruskal J.B. The Number of Simplices in a Complex, in Mathematical Optimization Techniques, Univ. of California Press, Berkeley, Calif. 1963, 251–278.
21. Nigmatullin R.G. Variational Principle in Boolean Algebra (in Russian), in Diskretny Analiz 10, Novosibirsk 1967, 69–89.

22. Papadimitriou C.H., Steiglitz K., Combinatorial optimisation: algorithms and complexity, Dover Publications Inc., 1998. New York, 498p.
23. Zhuravlev Yu., Elected Research Works, Magister, Moscow, 1998, 420p.
24. Wang D-.L., Wang P. Discrete Isoperimetric Problems, SIAM J. Appl. Math., 32, No.4(1977), 860–870.

Authors' Information

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

Vilik Karakhanyan – Senior Researcher, Institute for Informatics and Automation Problems, NAS RA, P.Sevak St. 1, Yerevan 14, Armenia, e-mail: kvilik@yahoo.com

ADAPTIVE NEURO-FUZZY KOHONEN NETWORK WITH VARIABLE FUZZIFIER

Yevgeniy Bodyanskiy, Bogdan Kolchygin, Iryna Pliss

Abstract*: The problem of neuro-fuzzy Kohonen network self-learning with fuzzy inference in tasks of clustering in conditions of overlapped classes is considered. The basis of the approach are probabilistic and possibilistic methods of fuzzy clustering. The main distinction of the introduced neuro-fuzzy network is the ability to adjust the values of fuzzifier and synaptic weights in on-line mode, as well as the presence except convenience Kohonen layer an additional layer to calculate the current values of the membership levels. The network characterized of computational simplicity, and is able to adapt to data uncertainty and detect new clusters appearance in real time. The experimental results confirm effectiveness of the approach developed.*

Keywords*: clustering, neuro-fuzzy network, self-learning algorithm, self-organizing Kohonen map.*

ACM Classification Keywords*: I.2.6 Artificial Intelligence - Learning - Connectionism and neural nets*

Introduction

The problem of multidimensional data clusterization is an important part of exploratory data analysis [Tukey, 1977; Höppner, Klawonn, Kruse, Runkler, 1999], with its goal of retrieval in the analyzed data sets of observations some groups (classes, clusters) that are homogeneous in some sense. Traditionally, the approach to this problem assumes that each observation may belong to only one cluster, although more natural is the situation where the processed vector of features could refer to several classes with different levels of membership (probability, possibility). This situation is the subject of fuzzy cluster analysis [Bezdek, 1981; Gath, Geva, 1989; Höppner, Klawonn, Kruse, Runkler, 1999], which is based on the assumption that the classes of homogeneous data are not separated, but overlap, and each observation can be attributed to a certain level of membership to each cluster, which lies in the range of zero to one [Höppner, Klawonn, Kruse, Runkler, 1999].

Initial information for this task is a sample of observations, formed from N-dimensional feature vectors $x(1), x(2), \dots, x(k), \dots, x(N)$. The result of clustering is segmentation of the original data set into m classes with some level of membership $u_j(k)$ of k-th feature vector $x(k)$ to j-th cluster, $j = 1, 2, \dots, m$.

In this paper we propose computationally simple adaptive procedure for recurrent fuzzy clustering of data processing in real time mode for operating in conditions of a priori uncertainty about the boundaries between classes, and their using for learning of fuzzy self-organizing Kohonen network.

Probabilistic fuzzy clustering

In the class of fuzzy clustering procedures the most mathematically strict are algorithms based on objective functions [Bezdek, 1981], that solve the problem of their optimization under various a priori assumptions. The most common is a probabilistic approach based on the minimization of the goal function

$$E\left(u_j, c_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)x\left(k\right)-c_j^{\,2} \tag{1}$$

under constraints

$$\sum_{j=1}^{m}u_j(k)=1, \tag{2}$$

$$0\le\sum_{k=1}^{N}u_j\left(k\right)\le N, \tag{3}$$

where $u_j\left(k\right)\in[0,1]$ is the level of membership of vector $x(k)$ to k -th class c_j-, centroid (prototype) of j -th cluster, $\beta-$ a non-negative parameter of fuzzification (fuzzifier) defining boundaries blur between clusters, $k=1,2,\ldots,N$. The result of clustering is $\left(N\times m\right)-$ matrix $U=\left\{u_j(k)\right\}$, called the matrix of the fuzzy partitioning.

Let’s note that because of constraints (2), elements of the matrix U can be considered as the probabilities of hypotheses of data vectors membership to defined clusters, because of what the procedures generated by minimizing (1) are called probabilistic algorithms of fuzzy clustering. The number of clusters m is given a priori and cannot be changed during the computation.

Introducing the Lagrange function

$$L\left(u_j\left(k\right),c_j,\lambda(k)\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}\left(k\right)x\left(k\right)-c_j^{\,2}+\sum_{k=1}^{N}\lambda\left(k\right)\left(\sum_{j=1}^{m}u_j\left(k\right)-1\right) \tag{4}$$

(there $\lambda\left(k\right)-$ undetermined Lagrange multiplier) and solving Karush-Kuhn-Tucker system of equations, we can easily obtain the desired solution in the form

$$\begin{cases} u_j\left(k\right)=\dfrac{\left(x\left(k\right)-c_j^{\,2}\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}\left(x\left(k\right)-c_l^{\,2}\right)^{\frac{1}{1-\beta}}}, \\ c_j=\dfrac{\sum_{k=1}^{N}u_j^{\beta}\left(k\right)x(k)}{\sum_{k=1}^{N}u_j^{\beta}\left(k\right)}, \\ \lambda\left(k\right)=-\left(\left(\sum_{l=1}^{m}\beta x\left(k\right)-c_l^{\,2}\right)^{\frac{1}{1-\beta}}\right)^{1-\beta}, \end{cases} \tag{5}$$

which at $\beta=2$ coincides with Bezdek’s Fuzzy C-means algorithm (FCM) [Bezdek, 1981], and at $\beta\to0$ is close to results obtained with the help of popular algorithm of ordinary (hard) the k-means clusterization (HCM) [McQueen, 1965].

Algorithm (5) cannot be in the fullest sense called fuzzy, because it contains crisp value of fuzzifier $1<\beta<\infty$, usually chosen from empirical considerations and essentially impact on the results. In this case variation of β from 1 to ∞ corresponds to transition from absolutely crisp boundaries ($\beta\to 1$) to total their blurring ($\beta\to\infty$) when all the observations belong to all clusters with the same level of membership.

Due to this Klawon and Höppner [Klawonn, Höppner, 2003] have proposed to use following expression for solving the problem of fuzzy clustering instead of the criterion (1) with a crisp fuzzifier:

$$E\left(u_j,c_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}(\alpha u_j^2(k)+(1-\alpha)u_j(k))x\left(k\right)-c_j{}^2 \tag{6}$$

with restrictions (2), (3), where $0<\alpha\le 1$ is a tunable parameter that determines character of solution obtained.

Introducing the Lagrange function

$$L\left(u_j\left(k\right),c_j,\lambda\left(k\right)\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2\left(k\right)+\left(1-\alpha\right)u_j\left(k\right)\right)x\left(k\right)-c_j{}^2+\sum_{k=1}^{N}\lambda\left(k\right)\left(\sum_{j=1}^{m}u_j\left(k\right)-1\right) \tag{7}$$

and solving the system of Karush-Kuhn-Tucker equations

$$\begin{cases}\dfrac{\partial l\left(u_j\left(k\right),c_j,\lambda\left(k\right)\right)}{\partial u_j\left(k\right)}=\left(2\alpha u_j\left(k\right)+1-\alpha\right)x\left(k\right)-c_j{}^2+\lambda\left(k\right)=0,\\ \nabla_{c_j}L\left(u_j\left(k\right),c_j,\lambda\left(k\right)\right)=-\sum_{k=1}^{N}2\left(\alpha u_j^2\left(k\right)+\left(1-\alpha\right)u_j\left(k\right)\right)\left(x\left(k\right)-c_j\right)=\vec{0},\\ \dfrac{\partial L\left(u_j\left(k\right),c_j,\lambda\left(k\right)\right)}{\partial\lambda\left(k\right)}=\sum_{j=1}^{m}u_j\left(k\right)-1=0,\end{cases} \tag{8}$$

we obtain the solution

$$\begin{cases}u_j\left(k\right)=-\dfrac{1-\alpha}{2\alpha}+\dfrac{1+m\dfrac{1-\alpha}{2\alpha}}{\sum_{l=1}^{m}\dfrac{x\left(k\right)-c_j{}^2}{x\left(k\right)-c_l{}^2}},\\ c_j=\dfrac{\sum_{k=1}^{N}(\alpha u_j^2\left(k\right)+(1-\alpha)u_j(k))x(k)}{\sum_{k=1}^{N}(\alpha u_j^2\left(k\right)+(1-\alpha)u_j(k))},\\ \lambda\left(k\right)=-\dfrac{1+m\dfrac{1-\alpha}{2\alpha}}{\sum_{l=1}^{m}\left(2\alpha x\left(k\right)-c_l{}^2\right)^{-1}}.\end{cases} \tag{9}$$

It is easy to see that when $\alpha = 1$ we obtain the Bezdek’s FCM algorithm

$$\begin{cases} u_j^{FCM}\left(k\right) = \dfrac{x\left(k\right) - c_j^{-2}}{\sum_{l=1}^{m} x\left(k\right) - c_l^{-2}}, \\ c_j^{FCM} \ = \dfrac{\sum_{k=1}^{N} (u_j^{\beta}\left(k\right))^2 x(k)}{\sum_{k=1}^{N} (u_j^{\beta}\left(k\right))^2}. \end{cases} \tag{10}$$

Taking that into account the first relation of (9) can be rewritten as

$$u_j\left(k\right) = -\frac{1-\alpha}{2\alpha} + \left(1 + m\frac{1-\alpha}{2\alpha}\right) u_j^{FCM}\left(k\right). \tag{11}$$

It is interesting to note that when $\alpha = \frac{1}{3}$ we obtain a compact expression

$$u_j\left(k\right) = (1+m) u_j^{FCM}\left(k\right) - 1. \tag{12}$$

Using all of above procedures assumes that the sample to be clustering containing N observations is given beforehand and cannot be changed during operation. In [Park, Dagher, 1984; Chung, 1994; Бодянский, Горшков, Кокшенев, Колодяжный, 2002; Bodyanskiy, Kolodyaznhiy, Stephan, 2002; Bodyanskiy, 2005] a group of recurrent adaptive procedures that allow to solve the problem of clustering in on-line mode was introduced.

Applying to (7) the Arrow-Hurwicz-Uzava nonlinear programming procedure, we come to an adaptive algorithm of fuzzy clustering according to the criterion (6):

$$\begin{cases} u_j\left(k+1\right) = -\dfrac{1-\alpha}{2\alpha} + \dfrac{1 + m\dfrac{1-\alpha}{2\alpha}}{\sum_{l=1}^{m} \dfrac{x\left(k+1\right) - c_j(k)^2}{x\left(k+1\right) - c_l(k)^2}}, \\ c_j\left(k+1\right) = c_j\left(k\right) + \eta(k)\left(\alpha u_j^2\left(k+1\right) + \left(1-\alpha\right) u_j\left(k+1\right)\right)\left(x\left(k+1\right) - c_j(k)\right), \end{cases} \tag{13}$$

where $\eta(k)$ – the learning parameter.

Possibilistic fuzzy clustering

The main disadvantages of the probabilistic approach associated with the constraint (2) requiring that the sum of memberships of each observation to all clusters would equal to one and the number of clusters m would set a priori. These disadvantages do not have methods of possibilistic fuzzy clustering, which were originally proposed by Krishnapuram and Keller [Krishnapuram, Keller, 1993].

In possibilistic clustering algorithms, the objective function has the form:

$$E\left(u_j,c_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}\left(k\right)x\left(k\right)-c_j^{\,2}+\sum_{j=1}^{m}\mu_j\sum_{k=1}^{N}\left(1-u_j(k)\right)^{\beta}, \tag{14}$$

where the scalar parameter $\mu_j>0$ determines the distance at which level of membership takes the value 0.5, i.e. if

$$x\left(k\right)-c_j^{\,2}=\mu_j, \tag{15}$$

then $u_j\left(k\right)=0,5$.

Direct minimization of (14) on $u_j\left(k\right)$ and c_j gives the known solution

$$\begin{cases} u_j\left(k\right)=\left(1+\left(\dfrac{x\left(k\right)-c_j^{\,2}}{\mu_j}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ c_j=\dfrac{\sum_{k=1}^{N}u_j^{\beta}\left(k\right)x(k)}{\sum_{k=1}^{N}u_j^{\beta}\left(k\right)}, \\ \mu_j\left(k\right)=\dfrac{\sum_{k=1}^{N}u_j^{\beta}(k)x\left(k\right)-c_j^{\,2}}{\sum_{k=1}^{N}u_j^{\beta}\left(k\right)}. \end{cases} \tag{16}$$

Using instead of (14) criterion

$$\begin{aligned} E\left(u_j,c_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2\left(k\right)+\left(1-\alpha\right)u_j\left(k\right)\right)x\left(k\right)-c_j^{\,2}+ \\ +\sum_{j=1}^{m}\mu_j\sum_{k=1}^{N}\left(\alpha(1-u_j\left(k\right))\right)^2+(1-\alpha)(1-u_j\left(k\right)) \end{aligned} \tag{17}$$

leads to a possibilistic procedure with variable fuzzifier

$$\begin{cases} u_j(k) = \dfrac{\mu_j(1+\alpha) - (1-\alpha)x(k) - c_j{}^2}{2\alpha\left(x(k) - c_j{}^2 + \mu_j\right)}, \\ c_j = \dfrac{\sum_{k=1}^{N}(\alpha u_j^2(k) + (1-\alpha)u_j(k))x(k)}{\sum_{k=1}^{N}(\alpha u_j^2(k) + (1-\alpha)u_j(k))}, \\ \mu_j = \dfrac{\sum_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)x(k) - c_j{}^2}{\sum_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)}. \end{cases} \tag{18}$$

In adaptive variant with processing sequentially receiving data we obtain at the recurrent procedure

$$\begin{cases} u_j(k+1) = \dfrac{\mu_j(k)(1+\alpha) - (1-\alpha)x(k+1) - c_j(k)^2}{2\alpha\left(x(k+1) - c_j(k)^2 + \mu_j(k)\right)}, \\ c_j(k+1) = c_j(k) + \eta(k)\left(\alpha u_j^2(k+1) + (1-\alpha)u_j(k+1)\right)\left(x(k+1) - c_j(k)\right), \\ \mu_j(k+1) = \dfrac{\sum_{p=1}^{k+1}\left(\alpha u_j^2(p) + (1-\alpha)u_j(p)\right)x(p) - c_j(k+1)^2}{\sum_{p=1}^{k+1}\left(\alpha u_j^2(p) + (1-\alpha)u_j(p)\right)} \end{cases} \tag{19}$$

Let’s note that in (13) and (19) algorithms for tuning the centroids of clusters are identical, and the difference consists in computation of membership levels to concrete cluster. It is also important that in contrast to probabilistic algorithms, possibilistic procedures allow during data processing to detect the appearance of new clusters. For example, if the level of membership of observation $x(k+1)$ for all clusters is below some preset threshold, we can speak about appearance of a $(m+1)$-th cluster with the initial coordinates of the centroid $c_{m+1} = x(k+1)$.

Adaptive neuro-fuzzy Kohonen network

It is easy to see that the second relations in the systems of equations (13), (19), being rewritten in the form

$$c_j(k+1) = c_j(k) + \eta(k)\varphi_j(k+1)\left(x(k+1) - c_j(k)\right), \tag{20}$$

represent nothing else but setting rules of self-organizing Kohonen neural network based on the principle of "winner takes more" [Kohonen, 1995], where $\varphi_j(k+1)$ is a neighborhood function. For these reasons, it is convenient to solve the considered problem of fuzzy clusterization on a basis of adaptive neuro-fuzzy Kohonen network, which architecture is shown in Fig. 1 and is a generalization of the architectures introduced in [Bodyanskiy, Gorshkov, Kolodyaznhiy, Stephan, 2005; Gorshkov, Kolodyaznhiy, Bodyanskiy, 2009].

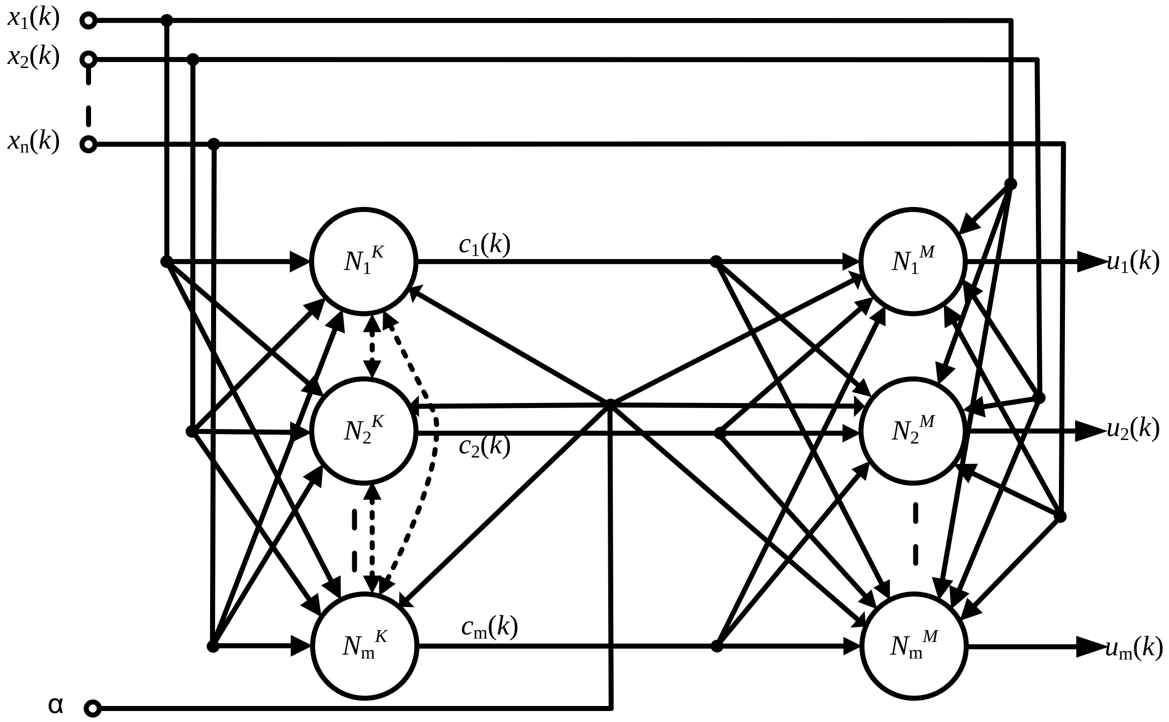


Figure 1 – Adaptive neuro-fuzzy Kohonen network with a variable fuzzifier

This network contains two layers: the Kohonen layer, which defines the prototypes (centroids) of clusters, and the layer of membership computing. Input vectors-pattern $x(k)=(x_1(k),\ x_2(k),\ldots,x_n(k))^T$ from reception (zero) layer are sequentially fed to the neurons of Kohonen layer N_j^K, that are tuned by the rule (20). Synaptic weights $c_{ji}(k),\ j=1,2,\ldots,m;i=1,2,\ldots,n$ define centroids of m overlapping clusters $c_j(k)=(c_{j1}(k),\ldots,c_{ji}(k),\ldots,c_{jn}(k))^T$. In the output layer formed by neurons N_j^M levels of membership $u(k)=(u_1(k),\ u_2(k),\ldots,u_n(k))^T$ of the presented vector $x(k)$ to j-th cluster according to the first relations of systems (13), (19) are calculated.

At lateral connections of Kohonen layer (shown dashed) processes of competition and cooperation are implemented, that are inherent to process of the Kohonen network training. To an additional input of the zero layer a customizable value of parameter α is fed.

Conclusion

The problem of fuzzy clustering of multidimensional observations with variable fuzzifier is considered and group of adaptive algorithms which enable to process data in real time as it received is proposed. Introduced algorithm is characterized by numerical simplicity and offers more flexibility when working in conditions of a priori uncertainty about the nature of data distribution in clusters. Proposed algorithms are used as learning rules of self-organizing neuro-fuzzy Kohonen network.

Acknowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Tukey, 1977] Tukey J. W. Exploratory Data Analysis. Reading, MA: Addison-Wesley Publ. Company, Inc., 1977.

[Höppner, Klawonn, Kruse, Runkler, 1999] Höppner F., Klawonn F., Kruse R., Runkler T. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester: John Willey & Sons., 1999,

[Bezdek, 1981] Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms, N.Y.:Plenum Press., 1981.

[Gath, Geva, 1989] Gath I., Geva A.B. Unsupervised optimal fuzzy clustering In: Pattern Analysis and Machine Intelligence., 1989., 2., 7., P. 773-787

[McQueen, 1965] McQueen J. On convergence of k-means and partitions with minimum average variance In: Ann. Math. Statist., 1965., 36., P. 1084.

[Klawonn, Höppner, 2003] Klawonn F., Höppner F. What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier. In: Lecture Notes in Computer Science., Springer Berlin-Heidelberg., 2003., Vol. 2811., P. 254-264.

[Park, Dagher, 1984] Park D. C., Dagher I. Gradient based fuzzy c-means (GBFCM) algorithm In: Proc. IEEE Int. Conf. on Neural Networks., 1984., P. 1626-1631.

[Chung, 1994] Chung F.L., Lee T. Fuzzy competitive learning In: Neural Networks., 1994., 7., №3., P. 539-552.

[Бодянский, Горшков, Кокшенев, Колодяжный, 2002] Бодянский Е.В., Горшков Е.В., Кокшенев И.В, Колодяжный В. В. Об адаптивном алгоритме нечёткой кластеризации данных In: Адаптивні системи автоматичного управління., Вип.5(25)., Дніпропетровськ: Системні технології., 2002., С. 108-117.

[Bodyanskiy, Kolodyaznhiy, Stephan, 2002] Bodyanskiy Ye., Kolodyaznhiy V., Stephan A. Recursive fuzzy clustering algorithms In: Proc. 10th East West Fuzzy Colloqium., Zittau, Germany., 2002., P. 276-283.

[Bodyanskiy, 2005] Bodyanskiy Ye. Computational intelligence techniques for data analysis In: Lecture Notes in Informatics., Bonn, Germany., 2005., Vol. P-72., P. 15-36.

[Krishnapuram, Keller, 1993] Krishnapuram R., Keller J. M. A possibilistic approach to clustering In: Fuzzy Systems., 1993., 1., №2., P. 98-110.

[Kohonen, 1995] Kohonen T. Self-Organizing Maps / Kohonen T., Berlin: Springer-Verlag., 1995.

[Bodyanskiy, Gorshkov, Kolodyaznhiy, Stephan, 2005] Bodyanskiy Ye. Combined learning algorithm for a self-orginizing map with fuzzy inference In: B. Reusch (Ed) “Computational intelligence: theory and applications”., Advanced in Soft Computing, Vol. 3: Berlin-Heidelberg: Springer-Verlag, 2005., P. 641-650.

[Gorshkov, Kolodyaznhiy, Bodyanskiy, 2009] Gorshkov Ye. New recursive learning algorithms for fuzzy Kohonen clustering network. In: Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems., Rapperswil, Switzerland, 2009., P. 58-61.

Authors' Information

***Yevgeniy Bodyanskiy** – Professor, Dr.-Ing. habil., Scientific Head of Control Systems Research Laboratory, Kharkiv National University of Radio Electronics; 14 Lenin Ave., Office 511, 61166 Kharkiv, Ukraine; e-mail: bodya@kture.kharkov.ua*

Major Field of Scientific Research: hybrid systems of computational intelligence.

***Bogdan Kolchygin** – Ph.D. student, Junior Researcher of Control Systems Research Laboratory, Kharkiv National University of Radio Electronics; 14 Lenin Ave., Office 511, 61166 Kharkiv, Ukraine; e-mail: quasimail@gmail.com*

Major Field of Scientific Research: adaptive fuzzy clustering.

***Iryna Pliss** – Ph.D., Leading Researcher of Control Systems Research Laboratory, Kharkiv National University of Radio Electronics; 14 Lenin Ave., Office 511, 61166 Kharkiv, Ukraine; e-mail: pliss@kture.kharkov.ua*

Major Field of Scientific Research: hybrid systems of computational intelligence.

CORRELATION MAXIMIZATION IN REGRESSION MODELS BASED ON CONVEX COMBINATIONS

Oleg Senko, Alexander Dokukin

Abstract: A new regression method based on convex correcting procedures over sets of predictors is developed. In contrast to previously developed approach based on minimization of generalized error, the proposed one utilies correcting procedures of maximal correlation with the target value. In the proposed approach a concept of a set of predictors irreducible against target functional is used where irreducibility is understood as lack of combinations of at least the same value of the functional after removing any of its predictors. Sets of combinations simultaniously irreducilbe and unexpandable are used during the construction of a prognostic rule. Results of some computational experiments described in the present article show an efficiency comparison between the two approaches.

Keywords: forecasting, bias-variance decomposition, convex combinations, variables selection.

ACM Classification Keywords: G.3 Probability and Statistics - Correlation and regression analysis, Statistical computing.

Introduction

Several statistical methods were developed last years that allow improving significantly prognostic ability of regression modeling in tasks of high dimension. Efficiency of these methods is associated with effective selecting of prognostic variables. Least angle regression or Lasso [Efron et al., 2004], [Tibshirani, 1996] methods may be mentioned thereupon. However a problem of low generalization ability of empirical models in high-dimensional tasks cannot be considered completely solved. Development of new alternative approaches may be useful for estimating of forecasting ability upper boundaries or for evaluating of selected variables optimal number. An approach in which optimal forecasting models are built by ensembles of preliminary trained predictors is discussed in this paper. It is supposed that initial predictors are simple. For example they may be one-variate or two-variate regression models. Suppose that we have set of L predictors $z_1,\dots,z_L$ that forecast some variable Y. Let $c=(c_1,\dots,c_L)$ be a vector of nonnegative coefficients satisfying condition $\sum_{i=1}^{L} c_i = 1$. Convex correcting procedure (CCP) calculates forecasted value as a weighted sum of prognoses that are calculated by single predictors:

$$Z_{ccp}(c)=\sum_{i=1}^{L} c_i z_i .$$

Convex combinations are widely used in pattern recognition. The bagging and boosting techniques [Breiman, 1999], [Kuncheva, 2004] may be mentioned as an example, as well as methods based on collective solutions by sets of regularities [Zhuravlev et al., 2008], [Zhuravlev et al., 2006], [Kuznetsov et al., 1996]. Convex correction is used in regression tasks also. Thus, neural networks ensembles are discussed in [Brown et al., 2005] that are based on optimal balance between individual forecasting ability of predictors and divergence between them. Efficiency of convex combinations of repressors' pairs was shown in [Senko, 2004]. Earlier it was shown that error of predictors' convex combination in any case is not greater than the same convex combination of single predictors' generalized errors [Krogh et al., 1995]. In previous works [Senko, 2009], [Senko et al., 2010] a method for CCP optimization has been studied that is based on minimization of general error estimates. Experiments with simulated data demonstrated that CCP error optimization also implements effective selection of informative prognostic variables. It is easy showing that the decrease of CCP variance comparing to the same combination of single predictors' variances is also a quality of convex combinations. Such a decrease deteriorates the CCP's prognostic ability. So, CCP predictions must be additionally adjusted, that may be done with the help of simple linear uni-variate regression. But forecasting ability of a linear regression model depends monotonically on correlation coefficients between Z_{ccp} and Y. In this paper we develop a new technique for constructing the CCP of maximal correlation with Y. This technique is based on the same concept of irreducible ensembles searching that was used in [Senko et al., 2010].

It is supposed further that predictors from initial set are additionally transformed with the help of optimal uni-dimensional regression models to achieve best forecasting ability. Such predictors will be further called reduced. In other words predictor z will be called reduced if for all α, β the inequality

$$E_\Omega\left(Y-\alpha z-\beta\right)^2 \le E_\Omega\left(Y-z\right)^2$$

is correct. Here $E_\Omega\left(X\right)$ is mathematical mean of X by space of admissible objects with defined σ-algebra and probability measure. It will be further denoted as $\hat{X}$. Variance of X will be denoted as $V(X)$. It is known that following equalities are true for a reduced predictor z:

$$\operatorname{cov}\left(Y,z\right)=E_\Omega\left[\left(Y-\hat{Y}\right)\left(z-\hat{z}\right)\right]=E_\Omega\left(z-\hat{z}\right)^2.$$

The use of the described conditions allows effectively searching ensembles with maximal prognostic ability, but the approach has its drawbacks. First of all, there are many ensembles with the prognostic ability close to the optimal one and it would be rational using them all. Secondly, CCP always decrease prognoses' variation and uni-dimensional correcting transformation becomes inevitable. Of all predictors the maximal quality is provided by the one most correlated with Y.

Irreducible ensembles relatively correlation coefficients

Standard Pearson correlation coefficient is defined as the ratio:

$$K\left(Y,Z_{ccp}\right)=\frac{\operatorname{cov}\left(Y,Z_{ccp}\right)}{\sqrt{V\left(Y\right)V\left(Z_{ccp}\right)}}.$$

On the other hand $\operatorname{cov}\left(Y,Z_{ccp}\right)=\sum_{i=1}^{L}c_i\operatorname{cov}\left(Y,z_i\right)$. But z_i is a reduced predictor. So, $\operatorname{cov}\left(Y,z_i\right)=V\left(z_i\right)$, $i=1,\dots,L$ and therefore

$$K\left[Y,Z_{ccp}\left(c\right)\right]=\frac{\sum_{i=1}^{L}c_iV\left(z_i\right)}{\sqrt{V\left(Y\right)}\sqrt{\sum_{i=1}^{L}c_iV\left(z_i\right)-\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho_{ij}}}.$$

Further discussions are based on irreducible ensemble concept. A set of predictors $\tilde{z}$ is called irreducible ensemble if removing of at least one predictor from it does not allow constructing CCP with the same prognostic ability as of $\tilde{z}$. The following is a strict definition of ensemble’s irreducibility.

Definition 1. Sets $\overline{D_L}$, D_L from $\mathbb{R}^L$ are defined as

$$\overline{D_L}=\left\{c\left|\sum_{i=1}^{L}c_i=1;\ c_i\geq 0, i=1,\ldots,L\right.\right\},$$

$$D_L=\left\{c\left|\sum_{i=1}^{L}c_i=1;\ c_i>0, i=1,\ldots,L\right.\right\}.$$

Definition 2. Set of predictors $z_1,\ldots,z_L$ is called irreducible ensemble relative to some functional $F\left(c\right)$, that characterize forecasting ability, if there is such vector $c^*\in D_L$, that $\forall c'\in\overline{D_L}\setminus D_L$, $F\left(c^*\right)>F\left(c'\right)$.

A set of points from $\mathbb{R}^L$ simultaneously satisfying constraints: $\sum_{i=1}^{L}c_i=1$ and $\sum_{i=1}^{L}c_iV\left(z_i\right)=\theta$ will be further referred to as $W\left(\theta\right)$.

Theorem 1. A necessary condition of irreducibility of predictors set $z_1,\ldots,z_L$ relative to $K\left(Y,Z_{ccp}\right)$ is existence of such real θ that quadratic functional

$$P_f\left(c\right)=\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho_{ij}^{v}.$$

achieves strict maximum at $W\left(\theta\right)$ in $c_1^*,\ldots,c_L^*$ that satisfies conditions $c_i^*>0$, $i=1,\ldots,L$.

The maximum necessary condition is existing of positive $\theta>0$, such that the following equation holds

$$\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho\left(z_i,z_j\right)\rightarrow\max \qquad (1)$$

with the next contingencies:

$$\sum_{i=1}^{L}c_iV\left(z_i\right)=\theta,$$

$$\sum_{i=1}^{L}c_i=1,$$

$$c_i\geq 0,\ i=1,\ldots,L. \qquad (2)$$

Lets write down a Lagrange functional for the task (1)

$$L=\sum_{i=1}^{L}\sum_{j=1}^{L}c_ic_j\rho\left(z_i,z_j\right)+\lambda\left(\sum_{i=1}^{L}c_iV\left(z_i\right)-\theta\right)+\mu\left(\sum_{i=1}^{L}c_i-1\right),$$

and equal its partial derivatives to zero

$$\frac{\partial L}{\partial c_k}=2\sum_{i=1}^{L}c_i\rho\left(z_i,z_k\right)+\lambda V\left(z_k\right)+\mu=0,$$

$$\frac{\partial L}{\partial\lambda}=\sum_{i=1}^{L}c_iV\left(z_i\right)-\theta=0,$$

$$\frac{\partial L}{\partial\mu}=\sum_{i=1}^{L}c_i-1=0.$$

Moving to a vectorial form we get

$$2DC+\lambda E+\mu I=O,$$

$$E^TC=\theta,$$

$$I^TC=1.$$

where $D=\left\|\rho\left(z_i,z_j\right)\right\|_{n\times n}$, $E=\left\|E\left(z_i^2\right)\right\|_{1\times n}=\left\|V\left(z_i\right)\right\|_{1\times n}$, $C=\left\|c_i\right\|_{1\times n}$, $I=\left\|1\right\|_{1\times n}$, $O=\left\|0\right\|_{1\times n}$.

Lets denote $\alpha=E^TD^{-1}E$, $\beta=I^TD^{-1}E$, $\gamma=I^TD^{-1}I$ for short. The received equation system gets the following form

$$2\theta+\lambda\alpha+\mu\beta=0,$$

$$2+\lambda\beta+\mu\gamma=0.$$

From these equations a dependence between c and θ can be derived

$$c_k=\frac{\theta\gamma-\beta}{\alpha\gamma-\beta^2}\sum_{i=1}^{L}d_{ki}V\left(z_i\right)+\frac{\theta\beta-\alpha}{\beta^2-\alpha\gamma}\sum_{i=1}^{L}d_{ki}>0,\ k=1,\dots,L, \tag{3}$$

where d_{ij} is an element of the D^{-1} matrix.

It must be noted also that the point c^* can be a point of strict maximum of P_f only if

$$\sum_{i=1}^{L}\sum_{j=1}^{L}\rho_{ij}\varepsilon_i\varepsilon_j>0 \tag{4}$$

for any $\left(\varepsilon_0,\dots,\varepsilon_L\right)$ satisfying conditions $\sum_{i=1}^{L}\varepsilon_i=0$. Let $\theta_{\min}$ is minimal and $\theta_{\max}$ is maximal value of θ for which one of inequalities (3) becomes equality. Let $R_k^v=\sum_{i=1}^{L}V\left(z_i\right)\rho_{ki}$, $P_k=\sum_{i=1}^{L}\rho_{ki}$,

$$\Gamma_i^1 = \frac{\gamma R_i^v + \beta P_k}{\alpha\gamma - \beta^2},$$

$$\Gamma_i^0 = \frac{\alpha R_i^v + \beta P_k}{-\alpha\gamma + \beta^2}.$$

then $P_f = B_0 + B_1\theta + B_2\theta^2$, where

$$B_0 = \sum_{i=1}^{L}\sum_{j=1}^{L}\Gamma_i^0\Gamma_j^0\rho_{ij},$$

$$B_1 = \sum_{i=1}^{L}\sum_{j=1}^{L}\left(\Gamma_i^0\Gamma_j^1 + \Gamma_i^1\Gamma_j^0\right)\rho_{ij},$$

$$B_2 = \sum_{i=1}^{L}\sum_{j=1}^{L}\Gamma_i^1\Gamma_j^1\rho_{ij}.$$

It is easy to show that

$$K\left(Y, Z_{ccp}\right) = \kappa\left(\theta\right) = \frac{1}{\sqrt{V(Y)}}\frac{\theta}{\sqrt{B_1\theta - B_2\theta^2 - B_0}}.$$

Theorem 2. Simultaneous correctness of inequalities $\theta_{\min} < \frac{2B_0}{B_1} < \theta_{\max}$, $\kappa\left(\frac{2B_0}{B_1}\right) > \kappa\left(\theta_{\min}\right)$ and negativity of the condition (4) is necessary condition of irreducibility of predictors set $z_1,\ldots,z_L$.

Necessary conditions allows effectively evaluate irreducibility of predictors set. It is sufficient to calculate $\theta_{\min}$ and $\theta_{\max}$ to evaluate negativity conditions (4) and to evaluate inequalities $\theta_{\min} < \frac{2B_0}{B_1} < \theta_{\max}$. It is evident that in case when necessary conditions are satisfied and $\kappa\left(\frac{2B_0}{B_1}\right)$ for the evaluated ensemble is greater than maximal correlation coefficient for any irreducible ensemble with less predictors than the evaluated ensemble is irreducible. It is important that optimal coefficients c_k may be received from (3) when $\theta = \frac{2B_0}{B_1}$.

Regression models based on sets of unexpandable irreducible ensembles

At the first stage initial set of reduced predictors is formed with the help of standard uni-variate least squares technique. Let $\tilde{Z} = \left(z_1,\ldots,z_L\right)$ is initial set of L predictors. An irreducible ensemble $\tilde{z}'$ consisting of l' predictors will be called unexpandable irreducible ensemble (UIE) if there are no irreducible ensembles in $\tilde{Z}$ with number of predictors greater l' that contain all predictors from $\tilde{z}'$. Two ways of regression model construction by sets of UIE were considered that are based on enumerating of all possible UIE. The first method chooses single best UIE where correlation coefficient of optimal Z_{ccp} with Y is maximal. This optimal Z_{ccp} ($Z_{ccp}^{\max}$) is the final regression model of the first method. The second method selects set of UIE where correlation coefficient of optimal Z_{ccp} with Y is greater than $Tr * K\left(Y, Z_{ccp}^{\max}\right)$, $Tr \in (0,1)$. Thus threshold parameter Tr allows to select UIE with correlation coefficient of optimal Z_{ccp} with Y close to maximal value $K\left(Y, Z_{ccp}^{\max}\right)$. It is supposed that Tr must be close to 1. In the second method parameters of final regression models are calculated as average by all UIE with $K\left(Y, Z_{ccp}\right) > Tr * K\left(Y, Z_{ccp}^{\max}\right)$. Our experiments showed that second approach is more effective. Method of UIE enumerating is based on gradual raising of predicates set meeting irreducibility condition. First, a set of all possible predictor pairs P_2 is considered. A set of all irreducible pairs P_2^{irr} is then extracted using the Theorem 2 results. Subsequently, a set of triplets P_3^{irr} is formed using P_2^{irr}. The process is going on until step i in which P_i^{irr} becomes empty. UIE based method depends on squared variances of single predictors V and distances between predictors ρ. The parameters were evaluated from training data by standard formulae $V(z) = \frac{1}{M}\sum_{j=1}^{M}\left[z_j - \hat{z}\right]^2$, $\rho\left(z^1, z^2\right) = \frac{1}{M}\sum_{j=1}^{M}\left[z_j^1 - z_j^2\right]^2$, where M is training set size. Experiments showed that such type of estimates leads to selection of too many variables and so to decrease of prognostic ability. However effectiveness may be systematically improved by using additional penalty multiplier for ρ equal $\frac{1}{1+\frac{5}{M}}$. This effect demands mathematical explanation.

Experiments

In all studies dependent variable Y and regressor variables X are stochastic functions of 3 latent variables U_1, U_2, U_3. The vector levels of variables U are independently distributed multivariate normal with mean 0 and standard deviation 1. The value of dependent variable Y in j-th case is generated by formula $y_j = \sum_{k=1}^{3} u_{jk} + e_y^j$ where u_{jk} is a value of the latent variable U_k, e_y^j is a random error term distributed $N\left(0, d_y\right)$. At that 85% of cases were generated with $d_y = 1$, 15% of cases were generated with $d_y = 2$. Thus, main and noisy components of data were modelled. The values of relevant variable X_i were generated by binary vector $\beta^i = \left\{\beta_1^i, \beta_2^i, \beta_3^i\right\}$. In j-th case $x_{jk} = \sum_{k=1}^{3} u_{jk}\beta_k^i + e_{xi}^j$, where u_{jk} is a value of the latent variable U_k, e_{xi}^j is a random error term distributed $N\left(0, d_{xi}\right)$. At that for 5 relevant variables $d_{xi} = 0.2$ and rest relevant variables were generated according $d_{xi} = 0.5$. The levels of irrelevant variable X_i in j-th case are generated by formula $x_{jk} = e_{xi}^j$. In each experiment 100 pairs of data sets were calculated by the random numbers generator according to the same scenario. Each pair includes training set that was used for optimal regression model construction and control data set that was used to evaluate prognostic ability of this model. In all experiments relevant variables were generated at $\beta = \{1,1,0\}$, $\beta = \{1,0,1\}$, $\beta = \{0,1,1\}$. Results of experiments are given in the Table 1. For each pair of samples of size M the following characteristics of forecasting ability for LARS and multiple UIE regression with $Tr = 0.95$ are given: K – correlation coefficient between variable Y and calculated prognoses, N_c – average number of relevant variables that were correctly used in regression model, N_f – average number of irrelevant variables that were mistakenly used in regression model, R_f – ratio of $\frac{|\beta|}{|\beta_{\max}|}$ for irrelevant variables. Here $|\beta|$ is an absolute value of regression coefficient for some variable in regression model, $|\beta_{\max}|$ is the maximal absolute value of regression coefficient among variables of regression model. It was considered that variable v_i is used by regression model if corresponding ratio $\frac{|\beta|}{|\beta_{\max}|}$ is less than 0.001.

Table 1. Results of expiriments. Prognostic ability.

	CCP_{cor}				CCP_{error}			
M	K	N_c	N_f	R_f	K	N_c	N_f	R_f
20	0.729	15.13	6.03	0.132	0.694	4.84	0.23	0.092
30	0.752	16.81	5.76	0.054	0.716	6.57	0.07	0.006
40	0.772	17.36	7.35	0.066	0.744	8.34	0.03	0.012
50	0.776	17.21	5.69	0.03	0.742	9.27	0	0

Conclusion

The results shown in the table 1 clearly show the superiority of the described novel approach, i.e. correction based on correlation maximization, over previously described [Senko et al., 2010] error minimization based one. Namely the correlation K is about 0.03 higher in all tasks. It is achieved by correct selection of almost all informative variables of the samples and though the amount of falsely selected noise variables is also increased, their weights in resulting combinations are low.

The primary drawback of the proposed method is slow speed that is decreasing dramatically with the increase of a task dimension. It is planned that further research will be aimed at reduction of computational complexity it. Nevertheless, the method proved to be suitable for a wide range of forecasting applications, especially in tasks which require feature selection.

Bibliography

[Efron et al., 2004] B. Efron, T. Hastie, I. Jonnstone and R. Tibshirani. Least Angle Regression. Annals of Statistics. 2004, Vol. 32, No. 2, 407–499.

[Tibshirani, 1996] Tibshirani R., Regression shrinkage and selection via the lasso // J. Roy. Stat. Soc. 1996. Vol. 58, p. 267–288.

[Breiman, 1999] L. Breiman, Random forests - random features. Technical report 567. Statistics department. University of California, Berkley, September 1999 // www.boosting.org.

[Kuncheva, 2004] L.I. Kuncheva, Combining Pattern Classifiers. Methods and Algorithms. Wiley Interscience, New Jersey, 2004.

[Zhuravlev et al., 2008] Zhuravlev Yu.I., Kuznetsova A.V., Ryazanov V.V., Senko O.V., Botvin M.A., The Use of Pattern Recognition Methods in Tasks of Biomedical Diagnostics and Forecasting // Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2008, Vol. 18, No. 2, pp. 195–200.

[Zhuravlev et al., 2006] Zhuravlev Yi.I., Ryazanov V.V., Senko O.V., RECOGNITION. Mathematical methods. Program System. Applications. —Moscow: Phasiz, 2006, (in Russian).

[Kuznetsov et al., 1996] Kuznetsov V.A., Senko O.V. et all., Recognition of fuzzy systems by method of statistically weighed syndromes and its using for immunological and hematological norm and chronic pathology // Chemical Physics, 1996, v. 15, N 1, p. 81–100.

[Brown et al., 2005] Gavin Brown, Jeremy L. Wyatt, Peter Tino, Managing Diversity in Regression Ensembles. Journal of Machine Learning Research 6: 1621-1650. 2005.

[Krogh et al., 1995] A. Krogh and J. Vedelsby, Neural network ensembles, cross validation, and active learning. NIPS, 7:231–238, 1995.

[Senko, 2004] Senko Oleg V., The Use of Collective Method for Improvement of Regression Modeling Stability // InterStat. Statistics on the Internet http://statjournals.net/, June, 2004.

[Senko, 2009] O.V. Senko,. An Optimal Ensemble of Predictors in Convex Correcting Procedures // Pattern Recognition and Image Analysis, MAIK Nauka/Interperiodica. 2009, Vol. 19, No. 3, pp. 465–468.

[Senko et al., 2010] Senko O., Dokukin A. Optimal Forecasting Based on Convex Correcting Procedures.// New Trends in Classification and Data Mining -ITHEA, Sofia, Bulgaria, 2010, p. 62-72.

Acknowledgements

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Authors' Information

Oleg Senko *– CCAS, chief researcher, 119333, Vavilova Str. 40, Moscow, Russian Federation; e-mail: senkoov@mail.ru*

Major Fields of Scientific Research: Pattern Recognition, Data Mining

Alexander Dokukin *– CCAS, researcher, 119333, Vavilova Str. 40, Moscow, Russian Federation; e-mail: dalex@ccas.ru*

Major Fields of Scientific Research: Pattern Recognition, Data Mining

NEURAL NETWORK SEGMENTATION OF VIDEO VIA TIME SERIES ANALYSIS

Dmitry Kinoshenko, Sergey Mashtalir, Andreas Stephan, Vladimir Vinarski

Abstract: Semantic video retrieval which deals with unstructured information traditionally relies on shot boundary detection and key frames extraction. For content interpretation and for similarity matching between shots, video segmentation, i.e. detection of similarity-based events, are closely related with multidimensional time series representing video in a feature space. Since video has a high degree of frame-to-frame-correlation, semantic gap search is quite difficult as it requires high-level knowledge and often depends on a particular domain application. Based on principal components analysis a method of video disharmony authentication has been proposed. Regions features induced by traditional frame segmentations have been used to detect video shots. Results of experiments with endoscopic video are discussed.

Keywords: Video Data, Frames, Time series segmentation, Principal component

ACM Classification Keywords: I.2.10 Vision and Scene Understanding (Video analysis), G.3 Probability and Statistics (Time series analysis).

Introduction

Nowadays video retrieval methods are rapidly evolving as the modus operandi for information creation, exchange, storage and content search [Petkovic, 2004, Shanmugam, 2009]. There arises an access to a tremendous amount of video information so it is infeasible for a human to classify or cluster the video scenes, to find appropriate events. In contrast to searching data in a relational database, a content based video retrieval (CBVR) requires the search of similar objects as a basic functionality of the database system. There is a number of approaches on summarization, video data management, streaming media analysis, video coding, video indexing, video abstraction, video information retrieval, etc. [Hanjalic, 2004, Snoek, 2008].

One of the most important video analysis issues is an automat identification of semantic events without an operator having to view the video. Video content analysis consists of motion, style and object detection, events and objects recognition, etc. Multimedia content analysis of video data so far has relied mostly on the information contained in the raw visual, audio and text signals. Content-based (Concept-based) video retrieval techniques strive to accomplish this goal by using low level image features, such as colors, textures, shapes, motions, etc. [Snoek, 2008, Hanjalic, 2004, Geetha, 2008]. But for now more and more video content analysis is considered as a capability of video analysis with the view of detection and determination of temporal events not based on a single image (or single frame). Hereupon great attention is spared to the analysis of separate shots of video, rather to their semantic relations and changes in time. Such dependences are especially important at a content search in video data bases, discovering features unbalances which are produced by scene changes in input video data.

Acceptable mathematical model to search video is an expression in a form of multidimensional time series, describing scene changes in a feature space. Such approach allows to analyze video scene changes in time. Thus, we can find changes at some time intervals by an analysis of frames sequence. Determinations of features or descriptions for each frame and changes comparisons in these descriptions give possibilities to draw a conclusion about changes happening in the time series or differently speaking in video data. This approach has got a wide development in a number of video data processing problems, and especially in the areas related to data segmentation in time series [Liniker, 2000, Rao, 2000]. But there are a large number of issues that remain to be investigated, in particular problems of multidimensional time series processing are not fully described in respect to video understanding.

Mathematical models of multidimensional time series under video shot search

For cases when the number of observations is not fixed and grows in time, to find shot boundaries in video stream via analysis of arbitrary feature space, mathematical models of multidimensional non-stationary time series are discussed in the current section.

There exist a number of models introduced for description of multidimensional sequences or time series [Izermann, 1984, Nikifora, 1991, Bassevile, 1993, Kerestencioglu, 1993] that generally can be presented by two basic forms. The first one is structural

$$\sum_{l=0}^{p} B_l x(k-l) + Dz(k) = \eta(k) . \tag{1}$$

Here B_l are matrix coefficients at intrasystem (endogenous) variables, B_0 is an nonsingular matrix at the endogenous variables of current time, D is a matrix of coefficients at exogenous variables, $z(k)$ is a vector of exogenous variables, including and their retarding values, $\eta(k)$ is a vector revolting signal with a zero maximal expectation and restricted second moments. The second one is normalized

$$x(k) = -B_0^{-1}(\sum_{l=1}^{p} B_l x(k-l) + Dz(k) - \eta(k)) \tag{2}$$

or

$$x(k) = CZ(k) + \xi(k) \tag{3}$$

where a vector $Z(k)$ plugs in itself both exogenous variables and retarding values of endogenous variables

$$\xi(k) = B_0^{-1}\eta(k).$$

The important problem is an authentication of parameters (1-3). Note, it often suffices to use indirect two-stage or three-stage least-squares methods, but they are intended for manipulations only with given dimensionality and predetermined data set [Izermann, 1984]. Algorithms intended for work in the sequential mode, such as [Nikifora, 1991] relaxation, recursive or method of fixed-point, are characterized by insufficient rate of convergence in order to provide signal processing in real-time.

In [Badavas, 1993] the model of multidimensional time series is examined

$$x(k)=\sum_{l=1}^{p}B_l x(k-l)+\sum_{p=1}^{q}D_p z(k-p)+F\psi(k-1)+\xi(k)\,, \tag{4}$$

or

$$B(z^{-1})x(k)=D(z^{-1})z(k-1)+F\psi(k-1)+G(z^{-1})\eta(k) \tag{5}$$

where unknown coefficients, describing behavior of observed sequence, enter either into matrices B_l, D_p, F of corresponding dimensions or in matrix polynomials $B(z^{-1})$, $D(z^{-1})$, $G(z^{-1})$ from the backward shift operator z^{-1}, $\psi(k)$ that is a determinate function, describing a trend being in a signal $x(k)$. In addition it should be emphasized that the same time series $x(k)$ can be described by infinite great number of multidimensional equals (4) or (5) [Badavas, 1993].

There exist three basic parameters evaluation methods of these expressions: maximum likelihood method, Bayes approach and method of the restricted information. If first two from them are realized in a packet form, then third is a form of recurrent least-squares method which can process sequentially incoming data. Unfortunately, standard recurrent least-squares method, being in fact, an identifier with infinite memory, inherently is unsuitable for analysis of substantially non-stationary objects whose features can hardly change properties.

For finding out the changes of properties of multidimensional series, compact effective and easy-to-use so-called vector autoregressive models (VAR models) had been proposed in [Pouliezos, 1994, Juselis, 1994].

Generally VAR models links the past and current supervisions of vector signals $x(k)$ in the form

$$x(k)=B_0+\sum_{l=1}^{p}B_l x(k-l)+\xi(k) \tag{6}$$

where $B_0=\{b_{0i}\}$ is $(n\times 1)$ vector of mean values, $B_l=\{b_{lij}\}$ are $(n\times n)$ matrices of parameters, p is a model order. Except the formula (6) of VAR model can be compact described in a state space

$$\begin{cases} x(k)=\Pi x(k-1)+\Pi_0+E(k), \\ y(k)=Cx(k) \end{cases} \tag{7}$$

where $x(k)=\begin{pmatrix} x(k) \\ x(k-1) \\ \vdots \\ x(k-p+1) \end{pmatrix}$, $\Pi_0=\begin{pmatrix} B_0 \\ \vec{0} \\ \vdots \\ \vec{0} \end{pmatrix}$, $\Pi=\begin{pmatrix} B_1 & \cdots & B_{p-1} & B_p \\ I_n & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_n & 0 \end{pmatrix}$, $E(k)=\begin{pmatrix} \xi(k) \\ \vec{0} \\ \vdots \\ \vec{0} \end{pmatrix}$,

$C=(I_n,0,\ldots,0)$, $x(k)-(np\times 1)$ is a vector of the states, Π is $(np\times np)$ translational matrix, $\vec{0}$ and 0 are $(n\times 1)$ and $(n\times n)$ zero vector and matrix accordingly.

Relations (7) allow to use the powerful mathematical tools of Kalman’s filtration for the analysis of multidimensional signals.

Detection of properties changes of multidimensional time series $x(k)$ is related to analysis of each of its components $x_i(k), i=1,2,\ldots,n$ and there arise three possible situations:

i) change of mean values of $l \le p$ components

$$b_{0i}(k) = \begin{cases} b_{0i}, \text{if } k < k_a, \\ b_{0i}^a, \text{ if } k \ge k_a, \end{cases}$$

ii) change of descriptions (dispersions) $l \le p$ of perturbation $\xi_i(\sigma_i^2)$

$$x_i(k) = \begin{cases} b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij} x_j(k-l) + \xi_i(k), \text{if } k < k_a, \\ b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij} x_j(k-l) + \xi_i^a(k), \text{ if } k \ge k_a, \end{cases}$$

iii) change of coefficients b_{lij}, causing the change of autocorrelation properties of non-stationary time series

$$x_i(k) = \begin{cases} b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij} x_j(k-l) + \xi_i(k), \text{if } k < k_a, \\ b_{0i} + \sum_{l=1}^{p} \sum_{j=1}^{n} b_{lij}^0 x_j(k-l) + \xi_i(k), \text{ if } k \ge k_a \end{cases}$$

where k_a is the instant time when measuring is executed.

Detection of multidimensional time series properties changes via principal components analysis

At the analysis of largescale (both on a volume and on a dimension) observations set in form of time series an important task lies in compression with the purpose of selection of latent factors qualificatory the underlying structure of the controlled signal, that pursues an aim to do initial time series more simply interpreted from the point of view of finding out of the properties changes.

One of the most effective going near the decision of this problem is a vehicle of factor analysis, within the framework of that the most wide distribution was got by the main components method especially in the problems of patterns recognition, image processing, spectrology etc. and known yet as Karhunen-Loeve’s transform.

A $(k \times n)$ matrix of observations

$$X = \begin{pmatrix} x_1(1) & x_2(1) & \ldots & x_n(1) \\ x_1(2) & x_2(2) & & x_n(2) \\ \vdots & & & \\ x_1(u) & x_2(u) & \cdots & x_n(u) \\ \vdots & & \ddots & \\ x_1(k) & x_2(k) & \cdots & x_n(k) \end{pmatrix} \tag{8}$$

that is generated by an array from the k n-dimensional vectors containing observations $x(u) = (x_1(u), x_2(u), ..., x_n(u))^T$ and also its $(n \times n)$ cross-correlation matrix

$$R(k)=\frac{1}{k}\sum_{u=1}^{k}(x(u)-\bar{x})(x(u)-\bar{x})^{T}=\frac{1}{k}\sum_{u=1}^{k}x^{C}(u)x^{CT}(u)$$

where $x^{c}(u)=x(u)-\bar{x}$ centered relatively basic data mean are input information for an analysis.

The kernel of PCA underlies in projection of observed data from input n-dimensional space into m-dimensional ($n>m\geq 1$) output and is reduced to the system $w^{1},w^{2},...,w^{m}$ of orthogonal eigenvectors of matrix $R(k)$ such that $w^{1}=(w_{1}^{1},w_{2}^{1},...,w_{n}^{1})^{T}$ corresponds to the greater eigenvalue λ, matrices $R(k)$, $w^{2}=(w_{1}^{2},w_{2}^{2},...,w_{n}^{2})^{T}$ corresponds to the second-largest eigenvalue λ etc. In any case, the problem is searching for matrix equation solution

$$(R(k)-\lambda_{l}I)w^{l}=0$$

such that $\lambda_{1}\geq\lambda_{2}\geq...\geq\lambda_{n}\geq 0$ and $\left\|w^{l}\right\|^{2}=1$.

Thus, in algebraic terms the problem solving is closely related to the goal seeking of eigenvalues and the rank of cross-correlation matrix; in geometrical sense it is a problem of passing to less dimension space with the minimal information loss; in statistical sense it is the search for orthonormal vectors set in input space assuming maximal data variations; and, finally, in algorithmic sense it is the problem of successive determination (excretions) of set of eigenvectors $w^{1},w^{2},...,w^{m}$ by optimization of each local functionals providing global criteria

$$I_{w}(k)=\frac{1}{k}\sum_{l=1}^{m}\sum_{u=1}^{k}(x^{CT}(u)w^{l})^{2}$$

with restrictions

$$\begin{cases}w^{lT}w^{p}=0,\text{ with } l\neq p,\\ w^{lT}w^{p}=1.\end{cases}$$

The first principal component bearing a maximum of information about the processed signal can be found by maximization of local criterion

$$I_{w}^{1}(k)=\frac{1}{k}\sum_{u=1}^{k}(x^{C}(u)w^{1})^{2}$$

by standard undetermined Lagrange multipliers.

Further, the projection on the first principal component is subtracted from every vector $x^{C}(u)$ and the first principal component of remains, being the second principal component of basic data and orthonormal to first one, is calculated.

The third principal component is calculated by projection of each input vector on the first and the second principal components, projection subtraction from each $x^{C}(u)$ and search for the first principal component of remains. Eventually we arrive at the third principal component of basic data. Other principal components are calculated recursively in concordance with described procedure.

To the present time the developed mathematical and programming tools have the same disadvantage – the necessity to know matrix X with the fixed dimension to implement Karhunen-Loeve transform. If data acquisition is consequent, standard factor analysis procedures become out of operation in real time.

In this connection the use of recurrent procedures is reasonable to find eigenvectors of matrix $R(k)$ by the sequential processing of observations $x(1), x(2),..., x(k), x(k+1)...$ of multidimensional time series without a calculation cross-correlation matrix in order to achieve real time.

In [Cichocki, 1993] an artificial neuron is described on the basis of adaptive linear associators for the calculation of the first principal component in real time. On fig. 1 modified neuron for finding out the properties changes in a multidimensional signal on the basis of analysis of principal components is proposed.

For the preliminary centered data the learning algorithm looks like

$$\begin{cases} w^1(k+1) = w^1(k) + \eta(k+1)(x(k+1) - y(k)w^1(k))y(k+1), \\ y(k+1) = x^T(k+1)w^1(k), w^1(0) \neq 0, y'(1) = x^T(1)w^1(0) \end{cases} \quad (9)$$

where $\eta(k+1)$ is a step parameter of adaptation which is chosen enough small to provide the algorithm stability. Also algorithm (9) gives vector $w^1(k)$ normalization i.e.

$$\left\| w^1(k) \right\|^2 = 1$$

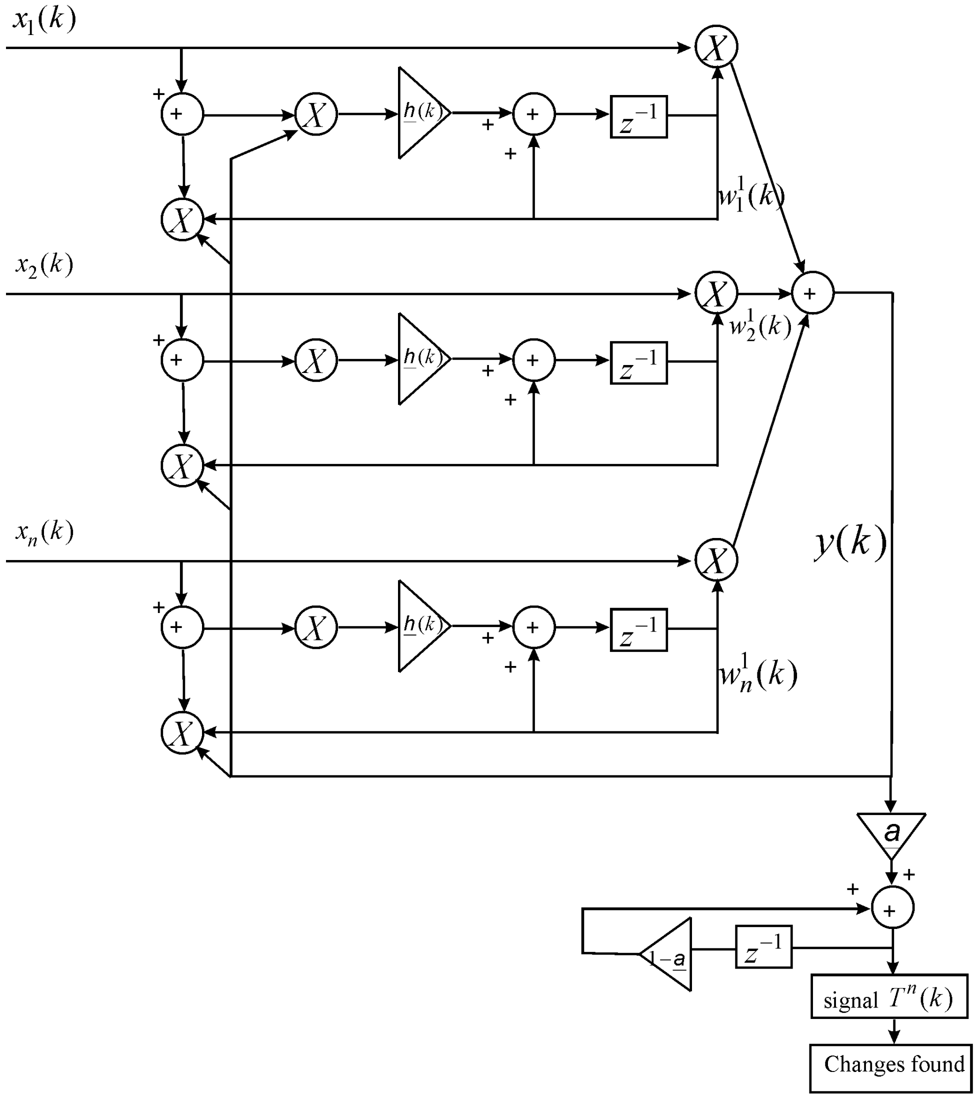


Fig.1 The modified neuron for finding out of principal component changes in multidimensional time series

Note the vector $w'(k)$ is the eigenvector of matrix $R(k)$, corresponding to the maximal eigenvalue and an output signal $y(k)$ is characterized by maximally possible dispersion, i.e. contains a maximum of information about a multidimensional input signal $x(k)$.

Further, an output signal $y(k)$ is processed with the exponential smoothing, filtering noise components $\xi(k)$, and finding out the properties change is produced by means of relation [Trigg, 1967]

$$T^{TL}(k) = \frac{T_i'(k)}{d_i(k)}$$

where $T_i'(k) = (1-\beta)\left|e_i(k)\right| + \beta T_i'(k-1)$, $d_i(k) = (1-\beta)\left|e_i(k)\right| + \beta d_i(k-1)$, $e_i(k)$ is a current estimation error, β is a smoothing parameter.

Discussion of experiments analysis

Experimental data set consists of different endoscopic videos each of which is composed by 550 frames. The goal of the experiment is to detect certain changes in the video streams. For this purpose at the first stage we built segmentation of initial video data frame by frame. Examples of input frame and its segmentation are presented on fig 2.

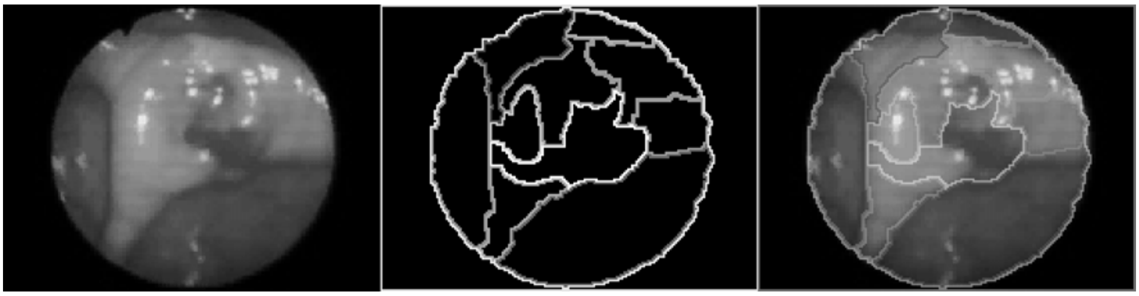


Fig. 2 Input frame and its segmentation

Extensively used Jseg algorithm [Comaniciu, 1997] was chosen for image spatial segmentation. The first part of experiments had the goal to find an influence of Jseg algorithm parameters (Scales, Quantresh, Merge) on the results of temporal segmentation of video in form of multidimensional time series. It is necessary to emphasized that the decline of threshold selection brings to appearance of more shallow regions and to an increase of its total number. Thus, it is possible to vary initial parameters in order to get more exact or more rough regions what is important for application areas, particularly for analyzed endoscopic video.

It should be noted that the change of some parameters insignificantly influences on the results of video data processing. In particular, if the parameter of Scales is varied and other parameters are invariable (see fig. 3), one can see the parameter values change, but the general trend of charts coincided.

Quantitative descriptions of the geometrical descriptions of regions that are produced by spatial segmentation were chosen to search video data changes. It should be noted that there exist a number of different descriptions of segmented images, in particular, shapes signature, polygonal approximation, moments, scale-space methods, shape transform domains etc [Mingqiang, 2008].

Numerous experiments allow us to assert that under considered frame region based video retrieval it suffices to find the integrated changes of such parameters for each region as area, perimeters, diameters of approximating circles, minimal and the maximal orthogonal projections, angle of slope to an axis and descriptions that are invariant to geometric transformations namely traditional moments features.

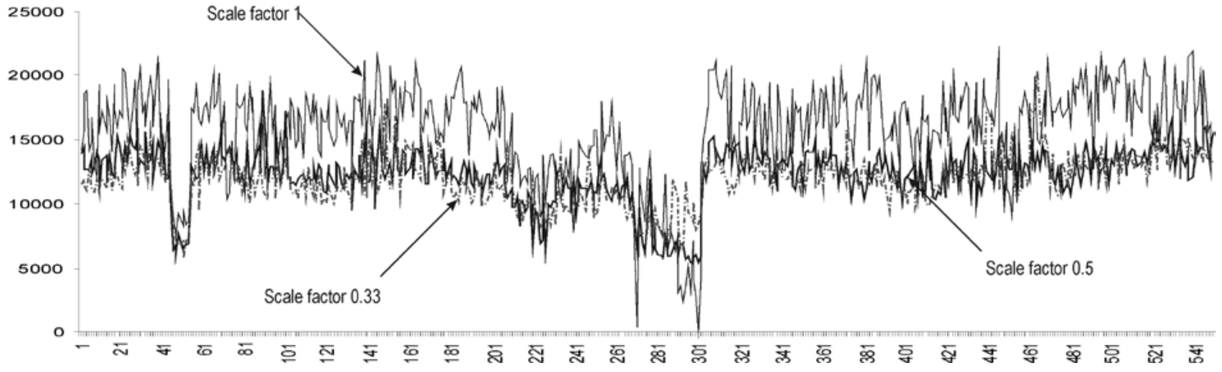


Fig. 3 Influence of segmentation parameters on video data processing

do not absolutely suit us they fully can appear unsensible to the substantial changes of initial video data. The incurrence of the chosen descriptions is equal to 11. After founding of these descriptions for all segments, the matrix of basic data (8) is formed, where as a line are values n of descriptions for one segment, and an amount of lines is a total regions number k at a frame. Thus, for each frame of video data we have a matrix of dimension $(k \times n)$ which we process in obedience to described methods. Further, using neural network (presented on fig. 1) we found one output parameter for each frame. An example of results is illustrated by fig. 4.

We can indicate that at the substantial changes of basic data, such as in a range a from 45 to 55 frame, we have shot boundaries produced by video artefacts or dramatic changes of basic data, or as in a district 300 frame we have the clear network output value overfalls, that also allows to see substantial changes of video. However in case of the small changes, for instance, related to shaking of camera, transition of eyeshot and other insignificant changes, search for shot boundaries remains enough difficult task.

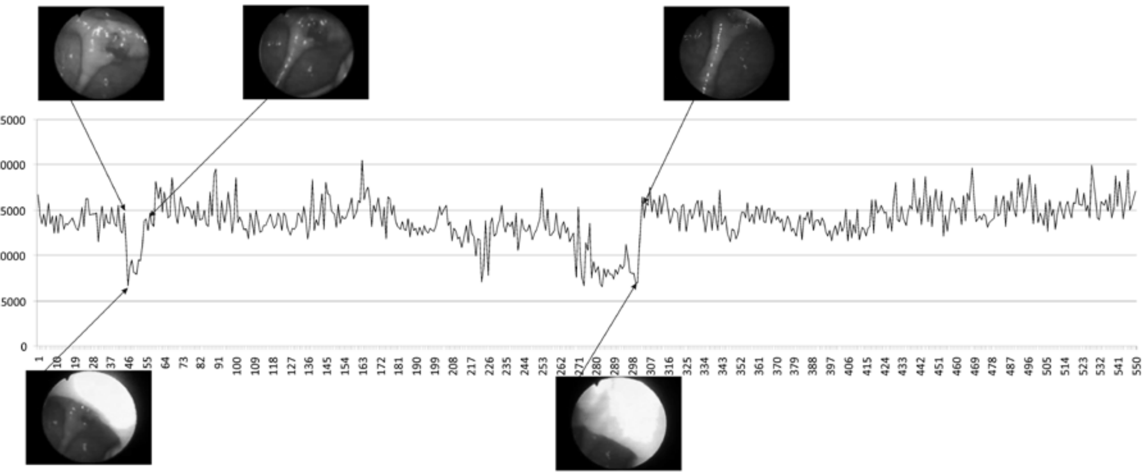


Fig. 4 Result of video data analysis

Conclusion

Content-based video retrieval systems whose effectiveness determines, in general, the success or failure in obtaining the required information are undergoing explosive growth due to the monotonic increase of accessible video data warehouses.

Based on low-level features the segmentation, breaking up the video sequences into temporally homogeneous segments (shots), is relatively simple and as a rule can be done automatically by shot-change detection algorithm. The development of shot-boundary detection algorithms has the longest and richest history in the video content analysis. But partitions a video stream into a set of meaningful and manageable segments video with a small intra-shot and large inter- shot variability generally were based on low-level spatial and rarely temporal features. Since humans perceive video as a complex interplay of cognitive concepts, a necessity of an access at the semantic level is obvious.

PCA based neural network temporal segmentation intended for real time has been proposed. Spatial segmentations of each frame or more exactly features extracted for each produced region of field of view in fact correspond to multidimensional time series describing video. All the more such partitions enough adequately reproduce ‘spatial content’ of an image and consequently frame sequences. Numerous experiments confirm a validity of theoretic results, however; offered approach solely gives possibility to find substantial changes in video data with filtration of marginal changes that can be important in some applications.

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS (www.aduis.com.ua).

Bibliography

[Badavas, 1993] P.C. Badavas. Real-Time Statistical Process Control. Eaglewood Cliffs, N.J.: Prentice-Hall, 1993.

[Bassevile, 1993] Bassevile M., Nikifora I. Detection of Abrupt Changes. Theory and Application. – Eaglewood Cliffs, N.J.: PTR Prentice-Hall, 1993.

[Cichocki, 1993] A. Cichocki, R. Unberhauen. Neural Networks for Optimization and Signal Processing. Stuttgart: Teubner, 1993.

[Comaniciu, 1997] D. Comaniciu, P. Meer. Robust Analysis of Feature Spaces: Color Image Segmentation In: IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[Geetha, 2008] P. Geetha, V. Narayanan. A Survey of Content-Based Video Retrieval. In: Journal of Computer Science, Vol.4, No 6, 2008.

[Hanjalic, 2004] A. Hanjalic Content-Based Analysis of Digital Video. Boston: Kluwer Academic Publishers, 2004.

[Izermann, 1984] Izermann R. Process Fault Detection Based Modeling and Estimating Methods – a Survey. In: Automatica, 20, No4, 1984.

[Juselis, 1994] Juselis K. The Cointegrated VAR-Model. N.Y.: Oxford University Press, 1994.

[Kerestencioglu, 1993] Kerestencioglu F. Change Detection and Input Design in Dynamical Systems. – Taunton, UK: Research Studies Press. – 1993.

[Liniker, 2000] F. Liniker, L. Niklasson Time Series Segmentation Using an Adaptive Resource Allocating Vector Quantization Network Based on Change Detection. In IEEE-INNS-ENNS International Joint Conference on Neural Networks, Vol. 6, 2000.

[Liu, 2007] Y.Liu, D. Zhanga, G. Lua, W.-Y. Ma. A Survey of Content-Based Image Retrieval with High-Level Semantics. In: Pattern Recognition. Vol. 40, No. 1, 2007.

[Mingqiang, 2008] Y. Mingqiang, K. Kidiyo, R. Joseph. A Survey of Shape Feature Extraction Techniques. In: Pattern Recognition Techniques, Technology and Applications, Vacuvar: Intech, 2008.

[Natsev, 2006] A. (P). Natsev. Multimodal Search for Effective Video Retrieval. In: Image and Video Retrieval. Berlin-Heidelberg: Springer-Verlag, Lecture Notes in Computer Science, Vol. 4071, 2006.

[Nikifora, 1991] Nikifora I.V. Sequential Detection of Changes in Stochastic Processes In: Prep. 9-th IFAC/IFORS Symp. ‘Identification and System Parameter Estimation’, Vol.1, 1991.

[Petkovic, 2004] M. Petkovic, W. Jonker. Content-Based Video Retrieval: A Database Perspective (Multimedia Systems and Applications). Boston-Dordrecht-London: Kluwer Academic Publishers, 2004.

[Pouliezos, 1994] A.D. Pouliezos, Y.S. Stavrakalis. Real Time Fault Monitoring of Clustering Processes. Dordrecht: Kluver Academic Publishers, 1994.

[Rao, 2000] Rao Y.N., J.C. Principe. A Fast On-line Generalized Eigendecomposition Algorithm for Time Series Segmentation. In: Adaptive Systems for Signal Processing, Communications, and Control Symposium, 2000.

[Shanmugam, 2009] T.N. Shanmugam, P. Rajendran An Enhanced Content-Based Video Retrieval System Based on Query-Clip. In: International Journal of Research and Reviews in Applied Sciences, Vol. 1, No 3, 2009.

[Snoek, 2008] C.G.M. Snoek, M. Worring. Concept-Based Video Retrieval. In: Foundations and Trends in Information Retrieval, Vol. 2, No. 4, 2008.

[Trigg, 1967] D.W. Trigg, A.G. Leach. Exponential Smoothing with an Adaptive Response Rate In: Operational Research, Vol. 18, No 1, 1967, P. 53-59.

Authors' Information

PhD Dmitry Kinoshenko *– Informatics department, Kharkiv National University of Radio Electronics, Lenin Ave. 14, Kharkiv, Ukraine, kinoshenko@kture.kharkov.ua*

Major Fields of Scientific Research: Video data analysis

PhD Sergey Mashtalir *– Informatics department, Kharkiv National University of Radio Electronics, Lenin Ave. 14, Kharkiv, Ukraine, mashtalir_s@kture.kharkov.ua*

Major Fields of Scientific Research: Video data analysis, Image processing

Dr. Andreas Stephan *– CURATYS International, Nonnengasse, 5a, 99084, Erfurt, Germany, Stefan@curatys.de*

Major Fields of Scientific Research: Time series analysis

Dr. Vladimir Vinarski *– University of Applied Sciences and Arts, Fachhochschule Hannover Ricklinger Stadtweg 120 30459 Hannover, Germany, vladimir.vinarski@fh-hannover.de*

Major Fields of Scientific Research: Video data analysis

ASTRONOMICAL PLATES SPECTRA EXTRACTION OBJECTIVES AND POSSIBLE SOLUTIONS IMPLEMENTED ON DIGITIZED FIRST BYURAKAN SURVEY (DFBS) IMAGES

Aram Knyazyan, Areg Mickaelian, Hrachya Astsatryan

Abstract: *The process of spectra extraction into catalogs from astronomical images, its difficulties and usage on the Digitized First Byurakan Survey (DFBS) plates are presented. The DFBS is the largest and the first systematic objective prism survey of the extragalactic sky. The large amount of photometric data is useful for variability studies and revealing new variables in the observed fields. New high proper motion stars can be discovered by a comparison of many observations of different observatories having large separation in years. The difficulty of DFBS images extraction is that extraction tools and programs are not adapted for such kind of plates. Astronomical images extraction process with usage of the Source Extractor (SE) tool is presented in this paper. The specificity of DFBS plates is that objects are presented in low-dispersion spectral form. It does not allow extraction tools to detect the objects exact coordinates and there is need of coordinates' correction. Apart this, it is required to configure SExtractor for current type of the plates so, that the output results be as close to real as possible. The extraction of DFBS plates will allow the creation of astronomical catalogs' database, which can be cross-correlated with known catalogs for investigation of the changes on sky during the years.*

Keywords: *IVOA, ArVO, DFBS, Plates extraction, SExtractor, VizieR, Astronomical catalogs*

ACM Classification Keywords: *I.4.1 Imaging geometry, Scanning I.4.3 Geometric correction, H.2.8 Data Mining, Scientific databases.*

Introduction

New generation of astronomical instruments produce terabytes of images and catalogs, which fundamentally change the way astronomy. In this context Moore's law is driving astronomy even further. Now Armenian Astronomers are facing different kind of problems, such as modeling and simulation of physical observations, data reduction, sharing and analysis, etc. It has been clear for many years that a national approach, to building such a data infrastructure is insufficient. The datasets and services which astronomers wish to use are spread all around the world. The good solution of data management, analysis, distribution and interoperability are virtual observatories (VO). About seventeen VO projects, including Armenian VO (ArVO), are now funded through national and international programs, and all projects work together under the International Virtual Observatory Alliance (IVOA) to share expertise and develop common standards and infrastructures for astronomical data (images, spectra, catalogs, literature, etc.) exchange and interoperability.

ArVO is a project of the Byurakan Astrophysical Observatory (BAO) (in collaboration with the Institute for Informatics and Automation Problems of NAS RA), which being developed since 2005. ArVO is aimed at construction of a modern system for data archiving, extraction, acquisition, reduction, use and publication. The

ArVO was created to utilize the Digitized First Byurakan Survey (DFBS) as an appropriate spectroscopic database.

The DFBS is the largest and the first systematic objective prism survey of the extragalactic sky. It covers 17,000 deg^2 in the Northern sky together with a high galactic latitudes region in the Southern sky. The FBS has been carried out by B.E. Markarian, V.A. Lipovetski and J.A. Stepanian in 1965-1980 with the Byurakan Observatory 102/132/213 cm (40"/52"/84") Schmidt telescope using 1.5 deg. prism. Each FBS plate contains low-dispersion spectra of some 15,000-20,000 objects; the whole survey consists of about 20,000,000 objects. A number of different types of interesting objects may be distinguished in the DFBS [DFBS; Mickaelian, 2007].

The same fields of the sky are observed by different observatories of the world, in different time periods. The large amount of photometric data is useful for variability studies and revealing new variables in the observed fields. New high proper motion stars can also be discovered by a comparison of many observations of different observatories having large separation in years. Hence, investigation of those observations differences is very useful. This investigation is being done by astronomical plate catalogs cross-correlation between each other. The VizieR Catalogue Service [VizieR] is an astronomical catalog service provided by the Strasbourg astronomical Data Center (CDS) [CDS], which collects and distributes astronomical data catalogs, related to observations of stars and galaxies, other galactic and extragalactic objects, solar system bodies and atomic data. It includes hundreds of catalogues from the all world and provides web access to those catalogs.

The use of high performance computing resources is crucial to fulfill the needs of astronomers that use numerical simulations for their research activity. Recent years many astronomical communities use the Grid infrastrcutuctures, as ”cyber-infrastructures” that support Astronomers in any aspect of their research activity, from data discovery and query to computation, from data storage to sharing resources and files. As an example, EuroVO covers interconnection between operational data and service grid, and supports interoperability between EuroVO and European Grid Initiative. In Armenia this problem has been successfully faced in the framework of the International Projects funded by the International Science and Technology Center.

The main purpose of this article is the presentation of DFBS astronomical plates cataloging objectives and its possible solutions based on investigation and testing results.

DFBS astronomical plates extraction into catalogs

Source Extractor (SE) [SE] is a program that builds a catalogue of objects from an astronomical image (photographic plates as well as CCDs). It is used for the automated detection and photometry of sources in files of FITS format. [Bertin, 1996] SE is particularly oriented towards reduction of large scale galaxy-survey data, but it also performs well on moderately crowded star fields. It is chosen for DFBS images extraction to catalogs.

SE builds catalogs in ASCII format and its package works in a series of steps. It determines the background and whether pixels belong to background or objects. Then it splits up the area that is not background into separate objects and determines the properties of each object, writing them to a catalog. All the pixels above a certain threshold are taken to belong to an object. If there is a saddle point in the intensity distribution (there are two peaks in the light distribution distinct enough), the object is split in different entries in the catalogs. Photometry is done on these by dividing up the intensity of the shared pixels. There is an option “clean” the catalog in order to eliminate artifacts caused by bright objects. Afterward, there is a list of objects with a series of parameters measured (ellipticity, size etc.). These are classified into stars (point-like objects) and galaxies (extended objects, everything non-star) by a neural network.

SE gets the positional information from the FITS header but most part of parameters must be specified in its input configuration file, which is an ASCII file (plain text) with the name of the parameters and the value on separate lines. Hence, it is very essential to choose the parameters such, that the output results be as close to real as possible. The threshold parameters indicate the level from which SE should start treating pixels as if they were part of objects, determining parameters from them. The main threshold parameter is counted by this formula:

$$T = k * \sigma \qquad (1)$$

where σ is the standard deviation of the image von, k isa coefficient which is given in input configuration. The bigger is k, the fewer objects will be detected.

Before the detection of pixels above the threshold, there is an option of applying a filter. Filter essentially smoothes the image. There are some advantages to applying a filter before detection. It may help detect faint, extended objects. However, it may not be so helpful if the data is very crowded. There are four types of filter (each one has its subtype) – Gaussian (the best for faint object detection), Mexican hat (useful in very crowded star fields, or in the vicinity of a nebula), Tophat (useful to detect extended, low-surface brightness objects, with a very low threshold) and “block” function of various sizes (is a small “block” function for re-binned images), all normalized.

As DFBS plates objects magnitudes are <18 (the objects are faint), the best filters for its extraction are Gaussian filters. It is also obtained by tests results.

De-blending is the part of SE where a decision is made whether or not a group of adjacent pixels above threshold is a single object or not. For example if there is a little island of adjacent pixels above the threshold it is not clear if it is an object or maybe several really close next to each other. This is set by the de-blend threshold parameter. This option is more essential especially for DFBS plates, as the objects are presented in spectral forms, their von is not homogeneous and the same object can be detected twice.

These are the most important parameters of SE configuration file. During the simulations and multiple tests the best configuration of SE for DFBS plates is found. But the main objective of DFBS plates’ extraction by SE is detected objects positions precision. This is illustrated on fig. 1.

SE sets detected object position by creating an ellipse around it and then found the center by programming calculations. The same is doing for DFBS spectrums and hence the coordinate is set somewhere on the center of spectrum both by x and y coordinates. Position by x is detected correctly, as spectrum is stretched only by y, by x it is the same as for usual images. As can be seen on fig.2, the real position by y is on the center of drawn circle, but SE finds it much lower. SE has options to find also detected object x and y minimum and maximum coordinates, which is very helpful. y(real) coordinate can be calculated with the formula below:

$$y(real)=y(max)-[x(max)-x(min)]/2 \qquad (2)$$

Where x(max) and x(min) are the detected object maximum and minimum coordinates by x, y(max) is maximum coordinate by y.

Thus, detected objects y coordinate will be reset by running a script on SE output ASCII catalog, which will calculate y(real) coordinate. But the coordinates can be calculated only by pixels using this method, and then it has to be converted to astronomical coordinates (RA and DEC) to be able to cross-correliate it with other catalogs.

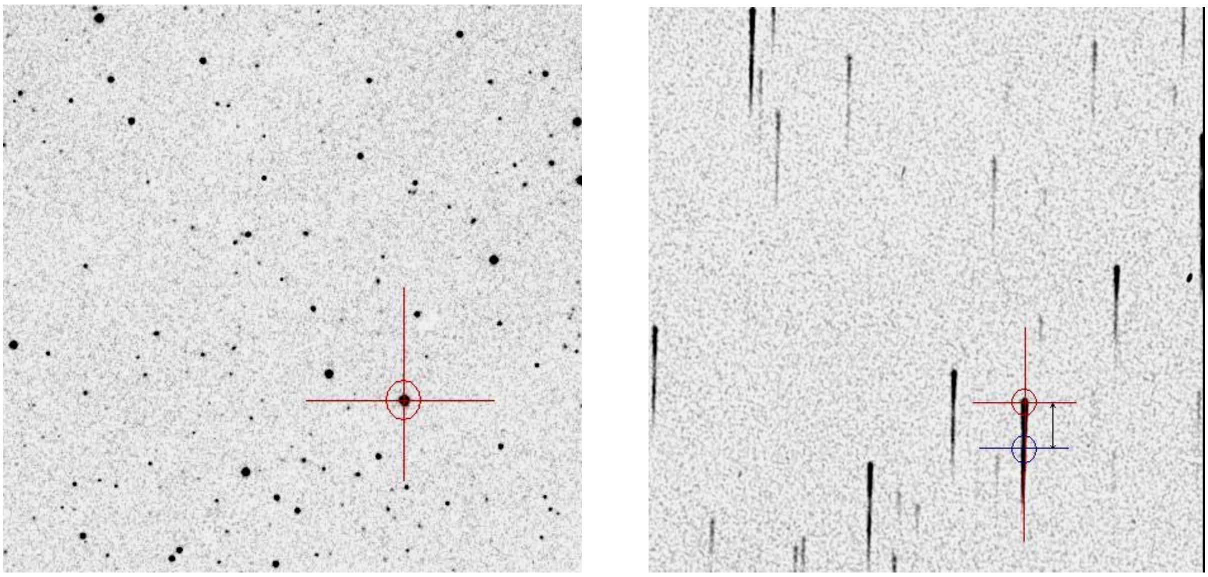


Fig. 1: Illustration of detected object position on the same sky field (left – DSS2 survey image, right - DFBS image)

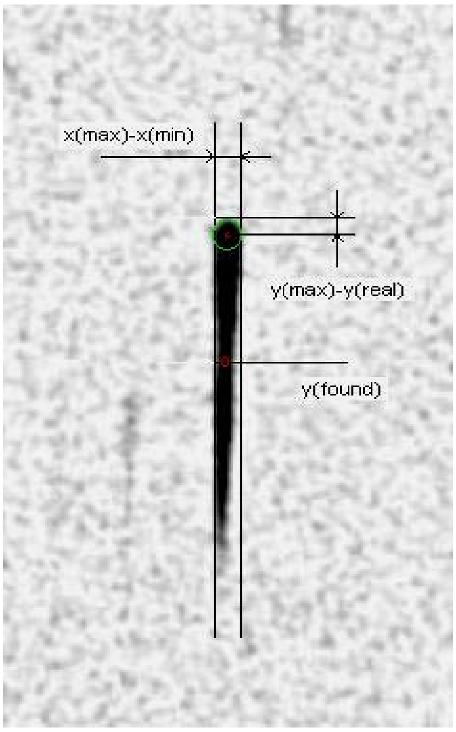


Fig.2: Illustration of DFBS detected spectrum coordinate correction

Conclusion

Astronomical catalogs research is very essential for discovery of new variables and high proper motion stars by a comparison of many observations of different observatories having large separation in years. In the future the aim is the extraction of DFBS all plates to catalogs, its integration in ArVO Portal and sharing with other virtual observatories. The cross-correlation of DFBS catalogs with known calalogs will allow to investigate the changes on sky during the years, as well as develop a classification scheme to estimate the nature of each object.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[IVOA] International Virtual Observatories Alliance, http://www.ivoa.net

[ArVO] Armenian Virtual Observatory, http://arvo.sci.am

[BAO] Byurakan Astrophysical Observatory, http://www.bao.am

[DFBS] Digitized First Byurakan Survey, http://byurakan.phys.uniroma1.it

[Mickaelian, 2007] Mickaelian, A. M.; Nesci, R.; Rossi, C.; Weedman, D.; Cirimele, G.; Sargsyan, L. A.; Erastova, L. K.; Gigoyan, K. S.; Mikayelyan, G. A.; Massaro, E.; Gaudenzi, S.; Houck, J.; Barry, D.; D'Amante, L.; Germano, P., The Digitized First Byurakan Survey – DFBS, Journal Astronomy and Astrophysics, Vol 464, pp. 1177-1180, 2007

[VizieR] The VizieR Catalogue Service, http:// vizier.u-strasbg.fr

[CDS] Strasbourg astronomical Data Center, http://astro.u-strasbg.fr

[SExtractor] SExtractor software, http://www.astromatic.net/software/sextractor

[Bertin, 1996] Bertin, E., Arnouts, S., SExtractor: Software for source extraction, Astronomy & Astrophysics Supplement 317, 393, 1996

Authors' Information

Aram Knyazyan *- Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia, 1, P. Sevak str., Yerevan 0014, Armenia; e-mail: aram.knyazyan@ipia.sci.am*

Areg Mickaelian *- Byurakan Astrophysical Observatory, National Academy of Sciences of the Republic of Armenia, Byurakan 0213, Aragatzotn province, Armenia; e-mail: aregmick@aras.am*

Hrachya Astsatryan - *Institute for Informatics and Automation Problems, National Academy of Sciences of the Republic of Armenia, 1, P. Sevak str., Yerevan 0014, Armenia; e-mail: hrach@sci.am*

MEMBRANES DISTRIBUTION USING GENETIC ALGORITHMS

Miguel Ángel Peña, Juan Castellanos

***Abstract**: Membrane computing is an area of natural computing, which solves NP-complete problems simulating permeability of live cells membranes. Different researchers have developed architectures to distribute membranes in clusters. They have studied, at theoretical level, the system behavior and the minimum time it would take to executing. In this paper proposes the use of genetic algorithms to distribute membranes in processors, thanks to their evolving capacities, they achieve distributions better than random distribution. Theoretical results are compared with a set of examples, noting improvement that genetic algorithms produce on these systems and how architectures are beneficial from execution viewpoint.*

***Keywords**: Distributed Communication, Membrane Computing, Membrane Dissolution, P-Systems Architectures, Genetic Algorithm*

***ACM Classification Keywords**: F.1.2 Modes of Computation, I.6.1 Simulation Theory, H.1.1 Systems and Information Theory, C.2.4 Distributed Systems*

***Conference topic**: Distributed and Telecommunication Systems*

Introduction

Membrane computing, inspired in “*basic features of biological membranes*”, was introduced by Gheorge Paun [Paun, 2000] to solve *NP*-Complete problems in polynomial time. As original model -Transition P System- as remaining models emerging from its; they are an abstract representation of hierarchical structure and non-deterministic behavior of biological membranes. A membrane is a region compounds by other membranes and chemicals (objects) that uses chemical reactions (evolution rules) generated another chemicals. Each membrane has a permeability capacity that enables chemicals (objects) to move between membranes (communication). Chemicals reactions produced in membranes can dissolve it. This process implies that contained object and membranes to become part of parent membrane.

In base to this behaviour, P-System are systems that can be executed in vitro or simulated, using hardware implementations ([Petreska, 2004], [Fernandez, 2005] or [Martinez, 2007]), using software simulations ([Suzuki, 2000] or [Arroyo, 2003]) or even two researcher using a real cluster of processors ([Ciobanu, 2004] and [Syropoulos, 2004]). Currently, researchers focused on simulations by distributed software, to alleviate the sequential nature of processors, to obtain lower running time. Must be distinguish two main steps in this system evolution: applying rules (with proper selection of it), and object communication between membranes. To apply rules, [Frutos, 2009] proposed to create decision trees to determine possible rules to apply according to membrane context, and [Gil, 2009] proposed algorithms to distribute objects among rules and its application.

Objects communication between membranes depends on used distributed architecture. Since Ciobanu detected that network congestion produce higher response time, future studies have focused on searching architectures to eliminate network collision.

For each architecture, its author presents a study about time it would takes in evolving a P System, using the membranes number and communication and application time of the cluster used, like input parameters. Nevertheless, is not considering P System topology, like one of these parameters and so it is not given a formula to membranes distribution into different cluster processor. Due to internal properties of existing architectures and diversities of tree topologies possible on the system –even with the same membranes number- it is not exists some algorithm able to indicate what is the better distribution of P-System on elements in the processor network that will be used for its evolution.

Genetic algorithms, based on a idea that genes of better adapted individuals tends to survivor, are a good option to seek better membranes distributions on processors of clusters, since in the algorithm execution will be finding better distributions will reduce the time evolution of the system. This paper pretends to show how genetic algorithms improve results of evolution global time, with better membranes distributions on processors. It also pretends to study the improvement percentage of this technique and possible differences between theoretical times of each architecture with real times obtaining after a realistic distribution, like are showed by some genetic algorithms. By last, it will compared different architectures to see if the differences found in the theory are also maintained when make practical distributions.

P System definition

The first definition of a P System was published by Paun [Paun, 2000], who defined a Transition P System as:

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

V is an alphabet (composed of objects).

μ is the membrane structure with n membranes.

ω_i are the multiset of symbols for the membrane i.

R_i are the evolution rules for the membrane i. A rule is a sorted pair (u,v) where u is a string over V, and $v = v'$ or $v = v'\delta$ and v' is a string over $V_{TAR} = V \times TAR$ with $TAR = \{here, out\} \cup \{in_j \mid 1 \leq j \leq n\}$. δ is a special symbol no include in V, and represent the dissolution of the membrane.

ρ_1 are the priority of rules for the membrane i.

i_0 indicates a membrane, which is the system output membrane or skin membrane.

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

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

2. To identify the set of applicable rules: It will be necessary to evaluate all the evolution rules of the membrane to identify those that meet the following constraint: its predecessor is contained in the multiset of objects of the region.
3. To build the set of actives rules: Intersection of two previous sets are the input group to this micro-step. Each one of the rules belonging to the set must be processed, to determine which meets the condition of active rule. To determine if a rule meets such condition, it is necessary check if there is not another rule with higher priority that belongs to the useful and applicable rules set.

4. Non deterministic distribution of objects of the region between its active rules and application: In this micro-step, copies of present objects in the multiset of the region are distributed between active evolution rules. Copies of objects that are assigned to each rule, match with those of the multiset that results from scalar product of a number between minimum and maximum bound of applicability of those rule and its predecessor. This distribution process is made on a non-deterministic way. Moreover, at the end of it, objects no assigned to any rule forms a multiset and they will be characterized because they do not contained to any predecessor of rules. The result of the distribution is a multiset of actives rules, where multiplicity of each rule defines the times that would be applied, and therefore, indirectly through its predecessors, objects are assigned to the multiset of the region. Objects used are eliminated and generate new objects that are destined for a membrane.
5. Transferring of multiset of generated objects or communication phase: In this micro-step, the new objects generated on the previous micro-step whose target membranes was *in* or *out*, must be transferred to its corresponding membranes. Each membrane, will have unused objects in its application, together with those that result for the applying of rules and that have this membrane as destiny.
6. Membranes dissolution: This action means that membranes who have applied some rule with dissolution capacity, they must send containing objects at this time to their nearest antecessor.
7. Composition of new objects multiset: At finish this last micro-step, on each membrane it should generate a new multiset of objects that will be used in the next step of evolution. In this way, multiset of objects of delimitated region by a membrane identified as *j* will be formed by:
   a. Objects do not assigned to any rule at the non-deterministic distribution.
   b. Objects created in the membrane whose target identifier, is the membrane itself.
   c. Objects created in daughter membranes whose target identifier was *out*.
   d. Objects coming from mother membrane, whose target identifier was in_j.
   e. Objects coming from dissolution of daughter membranes of membrane *j* and all objects whose their destiny was dissolved *j* daughter membrane.

Different researchers have reduced these theoretical microsteps to only two general steps, in function of number of involved membranes, without losing the characteristics of original system. In particular, the steps are: rule selection and application, and communication between membranes. The first step, which occurs on each membrane individually and without the need of it knows about the rest of P System is divided into rules selection and the subsequent application. [Frutos, 2009] proposes the rules selection based on decision tree, and [Gil, 2009] proposed an algorithm for non-deterministic assignation of objects in the selected rules, thus like application of these rules. On step of communication between membranes, like its name describe, are involved several membranes, and thus, in clustered deployments is necessary to know the network topology that forms the P-System, and that form the different processors. Because that, the communication depends on the used architecture.

P-System Architectures

The first architecture designed specifically to membranes distribution in diverse processors of a network or a cluster was proposed by Tejedor [Tejedor, 2008]. This architecture eliminates collisions previously found in experiments when they don't take into account the topology of P System [Ciobanu, 2004]. The architecture named "partially parallel evolution with partially parallel communication" originally and subsequently known like Peer-to-Peer (P2P) it distributed several membranes in each one of processors that used.

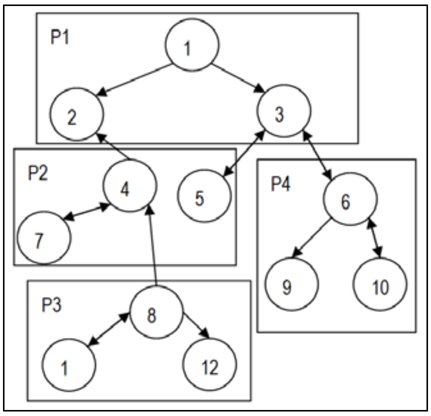


Fig. 1 P System example with 12 membranes distributed on 4 processors.

Rules selection and application step is made in parallel on each processor. So, all processors, at the same time, execute sequentially the rules selection and application on each one of its membranes. In the next step (communication between membranes), each processors, through its proxy communicates with the rest of adjacent processors according to tree topology used on membranes distribution. The communication order is determinate by topology, and in each moment only one processor is communicating. This order is observed on figure 2(a). Tejedor suggest that to know the evolution time of a P-System is necessary make a uniform distribution of membranes in processors, and this time is:

$$T_{P2P} = \left\lceil \sqrt{\frac{2MT_{com}}{T_{apl}}} \right\rceil T_{apl} + 2(\left\lceil \sqrt{\frac{MT_{apl}}{2T_{com}}} \right\rceil - 1)T_{com} \tag{1}$$

Where T_{apl} is the time that take each membrane in its application step (considering constant for all membranes), and T_{com} is the necessary time for each processor to communicate with other. M is the membranes number.

Based on the same idea of use the tree topology and the proxy like communicator element, Bravo proposes an improvement over this architecture using the superposition of communications, so in every moment it can communicate various processors without collisions. This architecture is known like Hierarchical Peer-to-Peer (HP2P). This topology was further enhanced [Pena, 2011] to allow membranes can be dissolve and with only one message can getting lower times. The order communication is showed on figure 2(b), and the needed time for each evolution is:

$$T_{HP2P} = \frac{T_{apl} M(A-1)}{A^L - 1} + (LA + L - 2)T_{com} \qquad (2)$$

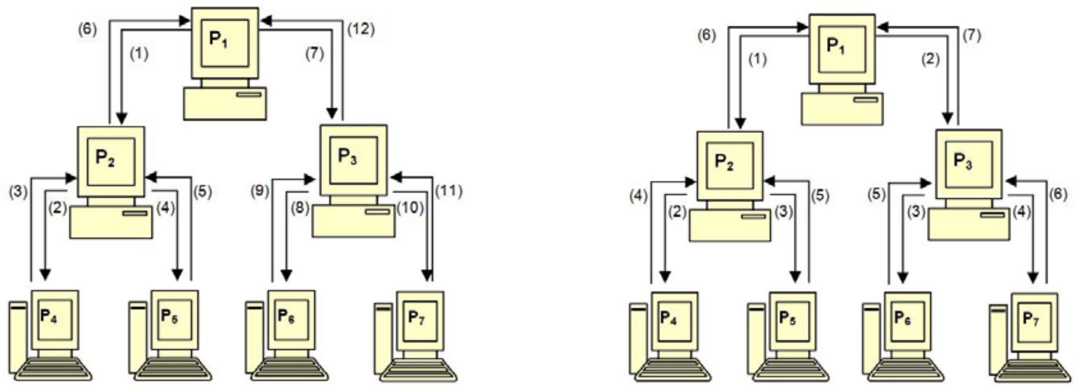
Where *A* is amplitude or processor number of children and *L* is number of processors tree levels, taking root like *L=1*. *L* and *A* values are determined by T_{com} y T_{apl} to minimize the function.

Third architecture that eliminates the restriction that processors form a tree is known like Master-Slave (MS) [Bravo, 2007b]. This architecture, where communication also is made by proxies (Figure 2(c)), each evolution step needs:

$$T_{MS} = \left\lceil \sqrt{\frac{MT_{com}}{T_{apl}}} \right\rceil T_{apl} + \left(\left\lceil \sqrt{\frac{MT_{apl}}{T_{com}}} \right\rceil + 1 \right) T_{com} \qquad (3)$$

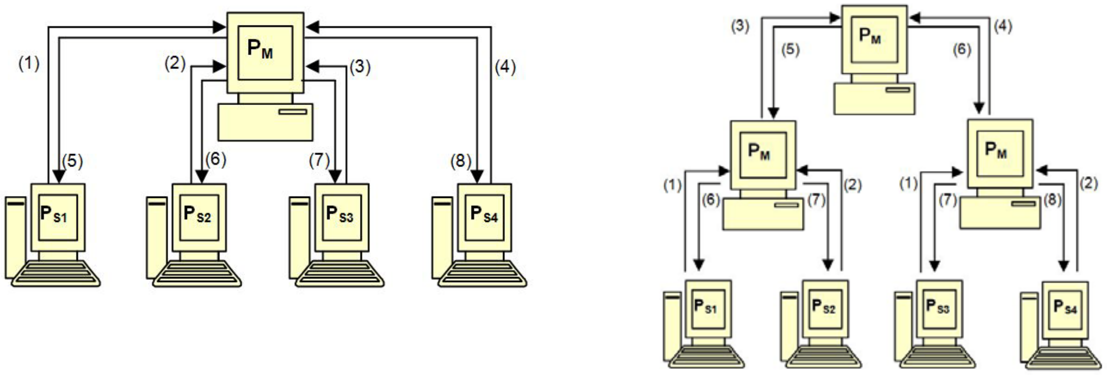
Unifying characteristics of the HP2P and MS architectures, it arises an architecture [Bravo, 2008] that uses the parallelism of HP2P architecture with elimination of restriction to make a distribution on a tree shape showed on MS architecture. Later, with introduction of membranes dissolution in P-System evolution, it was observed that behavior of HP2P and HMS is similar (figure 2(d)), but HMS besides not having a distribution in tree form, allows the overlap of steps. The time needed for each evolution depends if steps overlapping is total or partial:

$$T_{HMS} = \begin{cases} (\mathrm{AL} + \mathrm{L} - 2)T_{com} \quad if \quad \dfrac{MT_{apl}}{A^{L-1}} \le (A-1)T^{com} \\ \dfrac{MT_{apl}}{A^{L-1}} + (LA + L - A - 1)T_{com} \quad otherwise \end{cases} \qquad (4)$$



a) Peer to Peer architecture

b) Hierarchical Peer to Peer architecture



c) Master-Slave architecture

d) Hierarchical Master-Slave architecture

Fig. 2. Different distributed P-System architectures including the object communication order

Genetic algorithm

Genetic algorithms (GAs) are adaptive methods that can be used to resolve problems of searching and optimization. They are based on a genetic process of live organisms. Over generations, natural populations evolve in line with the principles of natural selection and survival of the fittest, proposed by Darwin [Darwin, 1859]. In imitation of this process, genetic algorithms are capable to creating solutions to real world problems. The evolution of these solutions to optimal values of problem depends on an appropriate codification of it. Basic principles of genetic algorithms were established by Holland [Holland, 1975], and are good described on text of Goldberg [Goldberg, 1989] or Davis [Davis, 1991]. The abstract algorithm, shows in a global way how must be implemented a genetic algorithm (Figure 3) and the parts that compose it.

```
BEGIN AGA
  Make initial population at random.
  WHILE NOT stop DO
    BEGIN
      Select parents from the population.
      Produce children from the selected parents.
      Mutate the individuals.
      Extend the population adding the children to it.
      Reduce the extend population.
    END
  Output the best individual found.
END AGA
```

Fig. 3. The pseudocode of the Abstract Genetic Algorithm (AGA).

To use a genetic algorithm on solution of a problem is needed define the characteristics and algorithms that it will use. For example, representation of individuals can be done in binary, integer or floating-point. To crossover can be used one-point crossover [Holland, 1975], n-point crossover or uniform crossover [Syswerda, 1991]...

Distribution using genetic algorithm

To use a genetic algorithm for membranes distribution, it must to decide representation to be used. It is a phenotype. As there are *m* membranes, are used *m* numbers, representing the *i-th* processor that goes into *i* membranes. Each of these numbers is an integer positive number.

It must distinguish two types of architectures, P2P and MS (HP2P and HMS are similar respectively, only changing the fitness function). To P2P distributions will be uses crossover and mutation oriented to tress and for MS are valid any evolutionary method (it will use one point crossover). On the same way, the initial population will be created on randomly way, but conserving the tree topology in case of P2P architecture.

It is used a method of proportional selection to objective function, forming a new population with new individuals, and 20% of more adapted of parents. Mutations always are performed with a 50% of probability that an individual mutates. Experimentally it has found that mutation increases the searching in the solutions space. It will stop on population *$log_{1.01}(m)$*. Population size will be *2m*.

With these parameters, has been proved suitability of genetic algorithms on set of 1000 P Systems of 100 membranes each one, and on 122 documented cases of 1000 membranes. Obtained results are showed on figures 4 and 5.

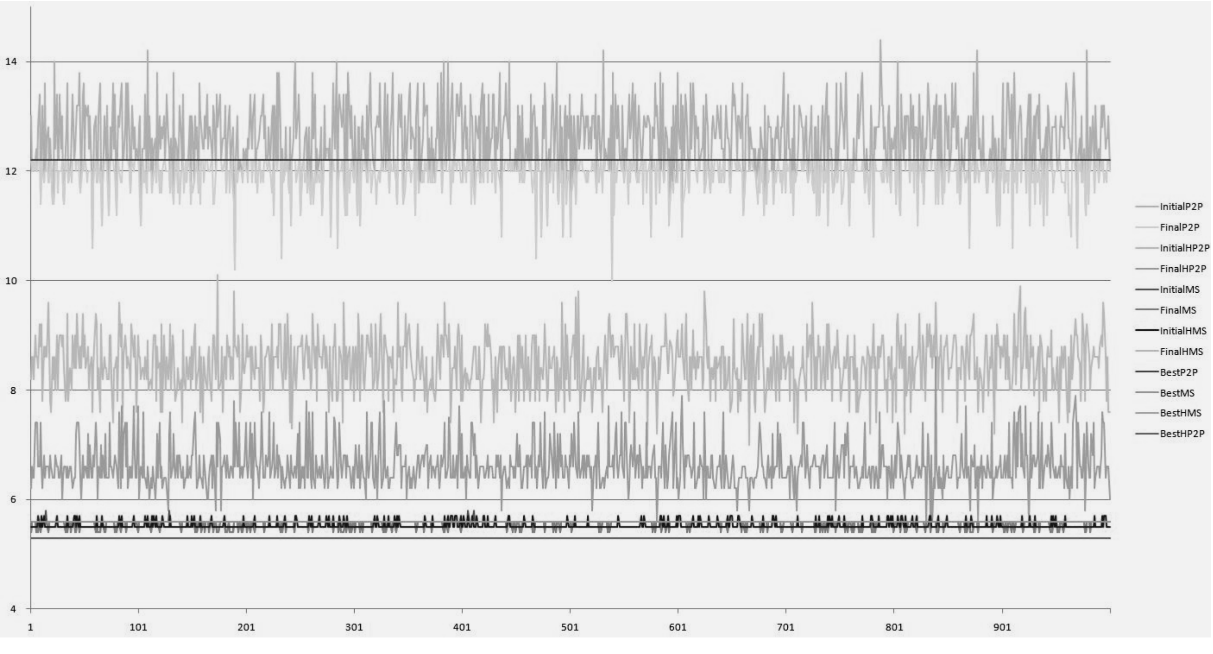


Fig. 4. Evolution times with 100 membranes

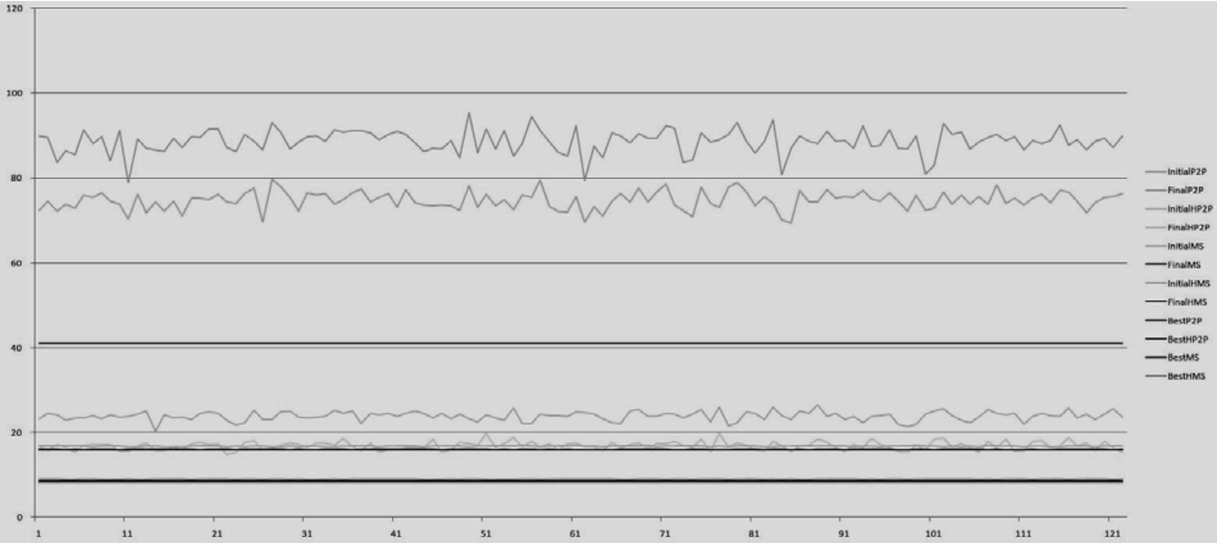


Fig. 5. Evolution times with 1000 membranes

On each graph is showed, to each P System (X axis), the needed time for its evolution (Y axis). Different lines showed: the best theoretical values for each architecture (BestP2P, BestHP2P, BestMS, BestHMS) and times used to make the evolution if is used the better solution of the initial population (initialP2P, initialHP2P, initialMS, initialHMS) or using the solution given by the algorithm (finalP2P, finalHP2P, finalHMS, finalHMS).

The first conclusion to be drawn is that there is a great difference between theoretical time of architectures and real times obtained after making the distribution. However, MS and HMS architectures time are similar because they are not penalized external communications.

It is very important pointed how genetic algorithms improve the evolution times of P System on 20%, given better distributions than the random distributions. And it is expected that with more evolutions of the genetic algorithm, the time improves even more.

Also it is needed mentioned how some solutions are even better that those theoretically calculated. It is showed in the P2P architecture for P System of 100 membranes. It is because in the theoretical calculus, rounding is applied because the processors number and the membranes number in each processor must be an integer, but in practice there are better solutions that used less processors or membranes on each processor getting better results.

Regarding to architectures, we see that the HMS architecture is not only better in theory, but there are distributions that achieve the same time and they need the least time.

Conclusion

Although, in theoretical studies shown four architectures for membranes distribution, none indicates how distributions must be done. This paper has shown the validity of genetic algorithms from random distributions, obtaining better distributions that approximates to optimal time for each architecture presented.

It has been analyzed results on architectures and was obtained that Hierarchical Master-Slave architecture is not only the best at theoretical level, but also it is easy to make membranes distributions that achieve the theoretical time; they are the best of the four architectures.

Bibliography

[Arroyo, 2003] Arroyo, F., Luengo, C., Baranda, A. V., and de Mingo, L. (2003). A software simulation of transition p systems in haskell. Membrane Computing, 2597:19-32.

[Bravo, 2007a] Bravo, G., Fernández, L., Arroyo, F., and Frutos, J. A. (2007). A hierarchical architecture with parallel communication for implementing p systems. In Kr. Markov, K. I. E., editor, Proceedings of the Fifth International Conference Information Research and Applications i.TECH 2007, volume 1, pages 168-174.

[Bravo, 2007b] Bravo, G., Fernández, L., Arroyo, F., and Tejedor, J. (2007b). Master-slave distributed architecture for membrane systems implementation. In Aggarwal, A., editor, Proceedings of the 8th WSEAS International Conference on Evolutionary Computing - Volume 8, pages 326-332, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

[Bravo, 2008] Bravo, G., Fernández, L., Arroyo, F., and Peña, M. A. (2008). Hierarchical master-slave architecture for membrane systems implementation. In Sugisaka, M. and Tanaka, H., editors, Proceedings of the 13-th International Symposium on Artificial Life and Robotics (AROB 2008), pages 485-490, Beppu, Japan.

[Ciobanu 2004] Ciobanu, G. and Guo, W. Y. (2004). P systems running on a cluster of computers. Membrane Computing, 2933:123-139.

[Darwin, 1859] Darwin, C. (1859). On the Origin of Species by Means of Natural Selection. John Murray, London.

[Davis, 1991] Davis, L. (1991). Handbook of Genetic Algorithms. Van Nostrand Reinhold.

[Fernandez, 2005] Fernandez, L., Martinez, V. J., Arroyo, F., and Mingo, L. F. (2005). A hardware circuit for selecting active rules in transition p systems. Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Proceedings, 1:415-418.

[Frutos, 2009] Frutos, J. d., Fernandez, L., and Arroyo, F. (2009). Decision trees for obtaining active rules in transition p systems. In Gheorghe Paun, Mario J. Perez-Jimenez, A. R.-N. n., editor, Tenth Workshop on Membrane Computing (WMC10), pages 210-217.

[Gil, 2009] Gil, F. J., Tejedor, J., and Fernandez, L. (2009). Fast lineal algorithm for active rules. International Journal "INFORMATION THEORIES & APPLICATIONS", 16(3):222-232.

[Goldberg, 1989] Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization & Machine Learning. Addison-Wesley, Reading, MA.

[Holland, 1975] Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

[Martinez, 2007] Martinez, V., Arroyo, F., Gutierrez, A., and Fernandez, L. (2007). Hardware implementation of a bounded algorithm for application of rules in a transition p-system. SYNASC 2006: Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Proceedings, 1:343-349.

[Paun, 2000] Paun, G. (2000). Computing with membranes. Journal of Computer and System Sciences, 61(1):108-143.

[Pena, 2011] Pena, M.A., Bravo, G., de Mingo, L.F. Membrane Dissolution in Distributed Architectures of P-Systems. Recent researches in applied computer and applied computational science, pages 229-234.

[Petreska, 2004] Petreska, B. and Teuscher, C. (2004). A reconfigurable hardware membrane system. Membrane Computing, 2933:269-285.

[Suzuki, 2000] Suzuki, Y. and Tanaka, H. (2000). On a lisp implementation of a class of p systems. In Romanian Journal of Information Science and Technology, volume 3, pages 173-186.

[Syropoulos, 2004] Syropoulos, A., Mamatas, E. G., Allilomes, P. C., and Sotiriades, K. T. (2004). A distributed simulation of transition p systems. Membrane Computing, 2933:357-368.

[Syswerda, 1991] Syswerda, G. (1991). Schedule optimization using genetic algorithms. In Davis, L., editor, Handbook of Genetic Algorithms, pages 332-349, New York. Van Nostrand Reinhold.

[Tejedor, 2008] Tejedor, J., Feranndez, L., Arroyo, F., and Bravo, G. (2008). An architecture for attacking the communication bottleneck in p systems. Artificial Life and Robotics, 12:236-240.

Authors' Information

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

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

VIRTUAL MEMBRANE SYSTEMS

Alberto Arteta, Angel Castellanos, Nuria Gómez

Abstract: *Within the membrane computing research field, there are many papers about software simulations and a few about hardware implementations. In both cases, algorithms for implementing membrane systems in software and hardware that try to take advantages of massive parallelism are implemented. P-systems are parallel and non deterministic systems which simulate membranes behavior when processing information.*

This paper presents software techniques based on the proper utilization of virtual memory of a computer.

There is a study of how much virtual memory is necessary to host a membrane model.

This method improves performance in terms of time.

Keywords: P-systems, Parallel systems, Natural Computing, evolution rules application, set of patterns, Virtual structure.

ACM Classification Keywords: D.1.m Miscellaneous – Natural Computing

Introduction

Membrane computing is a parallel and distributed computational model based on the membrane structure of living cells [Pãun 2000]. This model has become, during these last years, a powerful framework for developing new ideas in theoretical computation. The main idea was settled in the base of connecting the Biology with Computer Science in order to develop new computational paradigms. P-systems are the structures that reproduce the information processing occurring in the living cells. Nowadays, it can be found that several P Systems simulators are much elaborated [Pãun 2005]. At this point, there is a problem with the parallelism synthesized in [Pãun 2005] by: “parallelism a dream of computer science, a common sense in biology”. This is the reason why, Pãun avoids “to plainly saying that we have ‘implementations of P systems”, because of the inherent non-determinism and the massive parallelism of the basic model, features which cannot be implemented, at least in principle, on the usual electronic computer.

There are several uses of the p-systems. P-systems can be used to increase performance when dealing with known problems; for example, the knapsack problem [Pan, Martin 2005].

Also, Membrane computing is currently used to model other parallel and distributed systems as Grid [Qi, Li, Fu, Shi, You 2006].

An overview of membrane computing software can be found in [Ciobanu, Pérez-Jiménez, Ciobanu, Pãun 2006], or tentative for hardware implementations [Fernández, Martínez, Arroyo, Mingo 2005], or even in local networks [Ciobanu, Wenyuan 2004].

Work structure

This paper is structured as follows:

We will define some concepts as: P-systems, Extinguished multisets, multisets of objects, multiplicity of an object and evolution rules.

Reviewing the concept of patterns [Arteta, Castellanos, Martinez 2010].

Creation of a main n-dimensional structure in a virtual auxiliary n-dimensional structure. That main structure is an application that establishes a link between the initial multisets, and the number of times that each evolution rule should be applied in order to obtain an extinguished multiset.

Definitions

Here are some helpful definitions to understand this work. More info about these definitions can be found in [Arteta, Castellanos, Martinez 2010]

Definition Transition P Systems

A Transition P System of degree *n*, *n* > 1 is a construct

= $\Pi = (V, \mu, \omega_1, .., \omega_n, (R_{1,}\rho_1), ..(R_{n,}\rho_n), i_0)$

Where:

V is an alphabet; its elements are called objects;

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

$\omega_i \; 1 \le i \le n$, are strings from V^* representing multisets over *V* associated with the regions *1,2,..,n* of *µ*

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

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

Definition Multiset of objects

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

$$M : U \to N$$
$$a_i \to u_1$$

Where a_i is an object and u_i its multiplicity.

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

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

Definition Evolution rule with objects in *U* and targets in *T*

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

$m \in M(U), c \in M(UxT)$ *and* $\delta \in \{$*to dissolve*, *not to dissolve*$\}$

From now on '*c*' will be referred a *s* the consequent of the evolution rule '*r*'

Definition The set of evolution rules

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

Definition Multiplicity of an object in a multiset of objects *M(U)*

Let $a_i \in U$ be an object and let $m \in M(U)$ be a multiset of objects. The multiplicity of an object is defined over a multiset of objects such as:

$$| \; |_{a_i} : U \times M(U) \to N$$

$$(a_i, m) \to |m|_{a_i} = n \mid (a_i, n) \in m$$

Definition Multiplicity of an object in an evolution rule *r*

Let $a_i \in U$ be an object and let $R(U,T)$ be a multiset of evolution rules. Let $r = (m, c, \delta) \in R(U,T)$ where $m \in M(U), c \in M(UxT)$ *and* $\delta \in \{$*to dissolve*, *not to dissolve*$\}$

The multiplicity of an object is defined over an evolution rules such as:

$$| \; |_{a_i} : U \times R(U,T) \to N$$

$$(a_i, r) \to |m|_{a_i} = n \mid (a_i, n) \in m$$

Let C_i be the consequent of the evolution rule r_i. Thus, the representation of the evolution rules is:

$$\begin{array}{l} r_1 : a_1^{u_{11}} a_2^{u_{12}} ..a_n^{u_{1n}} \to C_1 \\ r_2 : a_1^{u_{21}} a_2^{u_{22}} ..a_n^{u_{2n}} \to C_2 \\ .... \to .... \\ r_m : a_1^{u_{m1}} a_2^{u_{m2}} ..a_n^{u_{mn}} \to C_m \end{array}$$

Observation

Let $k_i \in N$ be the number of times that the rule r_i is applied. Therefore, the number of symbols a_j which have been consumed after applying the evolution rules a specific number of times will be:

$$\sum_{i=1}^{m} k_i \cdot u_{ij} \tag{1}$$

Definition Extinguished Multisets

Given a region *R*, let *U* be an alphabet of objects $U = \{a_i \mid 0 < i \leq n\}$.

Let $M(U) = \{(a_i, u_i) \mid a_i \in U\ u_i \in \mathrm{N}\ 0 < i \leq n\}$ the set of all the multisets over *U*. Let $x \in M(U)$ be a multiset of objects over *U* within *R*. Let $R(U,T) = \{r_i\ i \in \mathrm{N} \mid \exists m \in \mathrm{N}\ i \leq m\}$ be a set of evolution rules. $r_j = (x_j, c_j, \delta) \in R(U,T)\ and\ x_j = \{(a_{ji}, u_{ji})\ \ i, j \in \mathrm{N} \mid \exists n, m \in \mathrm{N}\ i \leq n, j \leq m\}$

Let $k_j \in \mathrm{N}$ the number of times that the rule $r_j \in R(U,T)$ is applied over *x*. We say *x* is an Extinguished multiset if and only if :

$$\bigcap_{j=1}^{m} \left[\bigcup_{i=1}^{n} \left(u_i - \sum_{j=1}^{m} (k_j \cdot u_{ji}) \leq u_{ji} \right) \right]$$

Observation

In other words, *m* is an Extinguished multiset if and only if is not possible to apply any more evolution rules over it.

Patterns created from an evolution rules multiset.

In this section we show the patter definition. More information about patters can be found in [Arteta, Castellanos, Martinez 2010]

Definition Pattern

$N \otimes N \to A \quad A \subseteq \mathrm{P}(\mathrm{N})$

$[i, j] = \{k \in N \mid i \leq k \leq j\ \ i, j \in N\}$,

Definition Additive translation of a pattern

n+[*i,j*]=[*n*+*i,n*+*j*]

Definition Subtractive translation of a pattern

n-[*i,j*]=[*n*-*i,n*-*j*]

$n - i = 0 \quad n \leq i$

Definition Set of pattern $S \subseteq \mathrm{P}(\mathrm{N})$ is a Set of patterns is defined as the set:

$\{[\mathrm{a_i}, \mathrm{b_j}] \mid \exists n, m \in \mathrm{N}, i \leq n\ j \leq m\ \mathrm{a_i}, \mathrm{b_j}, i, j \in \mathrm{N}\}$

Definition Order in a set of patterns.

Let $\mathrm{S} = [[a_1, a_2], ..[a_3, a_4], .., [a_{n-1}, a_n]]$

We define:

$$\begin{aligned} f_{number}\ &: \mathrm{N} \otimes \mathrm{P}(\mathrm{N}) \to\ \mathrm{P}(\mathrm{N}) \\ &(i, S) \quad \to [a_i, a_{i+1}] \quad S \in \mathrm{P}(\mathrm{N}), i \in \mathrm{N} \end{aligned}$$

$\Rightarrow f_{number}(i, \mathrm{S}) = [a_i, a_{i+1}]$

Note: From now on, we will denote $f_{number}(i, \mathrm{S}) \quad as \quad S[i]$

Definition Inclusion in a set of patterns

Let S be a set of n patterns; and let $x = [x_1, x_2, ..., x_n] \in \mathrm{N}^n \quad x \in S \Leftrightarrow x_i \in S[i] \; \forall i \in \mathrm{N} \quad i \le n$

Definition Set of set of patterns

A Set of set of patterns $SS = \{S_i \quad i \in \mathrm{N} \; S_i \; \textit{is a set of patterns} \mid \exists n \in \mathrm{N} \; i \le n\}$.

Observation

Given a region *R* and alphabet of objects *U*, and *R* (*U, T*) set of evolution rules over *U* and targets in *T*.

$$
\begin{array}{l}
r_1 : a_1^{u_{11}} a_2^{u_{12}} ..a_n^{u_{1n}} \rightarrow C_1 \\
r_2 : a_1^{u_{21}} a_2^{u_{22}} ..a_n^{u_{2n}} \rightarrow C_2 \\
\qquad ... \qquad \rightarrow .... \\
r_m : a_1^{u_{m1}} a_2^{u_{m2}} ..a_n^{u_{mn}} \rightarrow C_m
\end{array}
$$

There is a Set of set of patterns $SS_{R(U,T)}$ associated to it. This set of set of patterns contains all the possible extinguished multisets and it is obtained by expanding the formula included in the definition of extinguished multiset:

$$\bigcap_{l=1}^{m} \left[\bigcup_{i=1}^{n} \left(u_i - \sum_{j=1}^{m} (k_j \cdot u_{ji}) \le u_{li} \right) \right] \quad \text{[Arroyo, Luengo 2003]}$$

$$
\left(
\begin{array}{c}
[[0,u_{11}],[0,u_{12}]...[0,u_{1n}]],[[0,u_{11}],[0,u_{12}]...[0,u_{2n}]],..,[[0,u_{11}],[0,u_{12}]...[0,u_{mn}]], \\
... \\
[[0,u_{11}],[0,u_{22}]...[0,u_{1n}]],[[0,u_{11}],[0,u_{22}]...[0,u_{2n}]],..,[[0,u_{11}],[0,u_{22}]...[0,u_{mn}]], \\
... \\
[[0,u_{11}],[0,u_{m2}]...[0,u_{1n}]],[[0,u_{11}],[0,u_{m2}]...[0,u_{2n}]],..,[[0,u_{11}],[0,u_{n2}]...[0,u_{mn}]], \\
... \\
[[0,u_{21}],[0,u_{12}]...[0,u_{1n}]],[[0,u_{21}],[0,u_{12}]...[0,u_{2n}]],..,[[0,u_{21}],[0,u_{12}]...[0,u_{mn}]], \\
... \\
[[0,u_{21}],[0,u_{22}]...[0,u_{1n}]],[[0,u_{21}],[0,u_{22}]...[0,u_{2n}]],..,[[0,u_{21}],[0,u_{22}]...[0,u_{mn}]], \\
... \\
[[0,u_{21}],[0,u_{m2}]...[0,u_{1n}]],[[0,u_{21}],[0,u_{m2}]...[0,u_{2n}]],..,[[0,u_{21}],[0,u_{n2}]...[0,u_{mn}]], \\
... \\
[[0,u_{m1}],[0,u_{12}]...[0,u_{1n}]],[[0,u_{1n}],[0,u_{12}]...[0,u_{2n}]],..,[[0,u_{m1}],[0,u_{21}]...[0,u_{mn}]], \\
... \\
[[0,u_{m1}],[0,u_{22}]...[0,u_{1n}]],[[0,u_{1n}],[0,u_{22}]...[0,u_{2n}]],..,[[0,u_{m1}],[0,u_{22}]...[0,u_{mn}]], \\
... \\
[[0,u_{m1}],[0,u_{m2}]...[0,u_{1n}]],[[0,u_{1n}],[0,u_{m2}]...[0,u_{2n}]],..,[[0,u_{m1}],[0,u_{2n}]...[0,u_{mn}]], \\
\end{array}
\right)
$$

Observation:

The cardinal of $SS_{R(U\ ,T)} \quad |SS_{R(U\ ,T)}| = n^m$ where n=|U| and m=|$R(U,T)$|

is

Definition Translation of a Set of set of patterns

$$\mathrm{N}^n \otimes \mathrm{N}^{n^m} \to \mathrm{N}^{n^m}$$

$$(x_1, x_2, .., x_n) - \begin{pmatrix} [[0,u_{11}],...[0,u_{1n}]],..., [[0,u_{11}],..[0,u_{mn}]], \\ ... \\ [[0,u_{m1}],..[0,u_{mn}]],..., [[0,u_{m1}],..[0,u_{mn}]], \\ \end{pmatrix} = \begin{pmatrix} [[x_1 - u_{11}, x_1],...[x_n - u_{1n}, x_n]],..., [[x_1 - u_{11}, x_1],..[x_n - A_{mn}, x_n]], \\ ... \\ [[x_1 - u_{m1}, x_1],..[x_n - u_{mn}, x_n]],..., [[x_1 - u_{m1}, x_1],..[x_n - u_{mn}, x_n]], \\ \end{pmatrix}$$

Evolution rules linear functions

In [Arteta, Castellanos, Martinez 2010] several functions are defined to build a structure in the physical memory of a computer. Our work defines these virtual functions to build the structure in virtual memory. During this part we will define the necessary function that will allow us to create the virtual structure. The structure is placed on virtual memory. The function will be referred to as “virtual” evolution rules linear isomorphism. These functions will contain itself other function as follows.

Virtual linear evolution rules function

Based on the previous definitions, it is possible to define a function which will be the key for building a linear structure and then to allocating it in virtual memory. This function will be the composition of two different functions.

Virtual linear Multisets function

The function is defined as follows:

$$\Phi_{1virt} : \mathrm{N}^n \quad \to \quad \mathrm{P}(\mathrm{N}^m)$$

$$[x_1, x_2, .., x_n] \to \begin{bmatrix} k_1, k_2, .., k_m \\ k_1^{'}, k_2^{'}, .., k_m^{'} \\ ... \\ k_1^{'p}, k_2^{'p}, .., k_m^{'p} \end{bmatrix}$$

Given an input $x = (x_1, x_2, .., x_n) \in \mathrm{N}^n$, it returns the set of numbers $k = (k_1, k_2, .., k_m) \in N^m / \varphi_1(k) = x$

Virtual linear Pattern function

$$\Phi_{2virt} : \qquad \mathrm{P}(\mathrm{N}^m) \qquad \to \qquad \mathrm{N}^n$$

$$\begin{pmatrix} [[0,u_{11}],...[0,u_{m1}]],..., [[0,u_{11}],..[0,u_{mn}]], \\ ... \\ [[0,u_{12}],..[0,u_{m1}]],..., [[0,u_{12}],..[0,u_{mn}]], \\ \end{pmatrix} \to \begin{bmatrix} x_1, x_2, .., x_n \\ x_1^{'}, x_2^{'}, .., x_n^{'} \\ ... \\ x_1^{'p}, x_2^{'p}, .., x_n^{'p} \end{bmatrix}$$

Given a set of set of patterns as the input it returns a set of numbers $X=(x_1,x_2,..,x_n)\in \mathrm{N}^n$. The elements of this resulting set are all the combinations of all the possible

$x^j=(x_1,x_2,..,x_n)\in \mathrm{N}^n$ where $x_i^j \in pattern \quad (i)\in SP\ (j)$ of a set of patterns contained in the matrix of set of patterns.

Building virtual linear structures from the multisets isomorphism.

Creation of the virtual linear structure $L_{\Phi\ virt}$

Given *V*={$\{X_i \mid i=1,..n\}$ be a multiset of symbols and given *R*(*U*,*T*), a multiset of evolution rules. Given the set {$k_i \in N$ the number of times that the evolution rule r_i is applied over the initial multiset}. We build the evolution rules function. $\Phi_{virt}=(\Phi_{1virt}\circ\Phi_{2virt})\,(P_{R(U,T)}(A_{ij})$

Once the virtual function has been constructed, the new virtual linear structure must be built this way:

$$L_{virt}\left[\Phi_{2virt}\left(P_{R(U,T)}(A_{ij})\right)\right]=\Phi_{1virt}\circ\Phi_{2virt}\left(P_{R(U,T)}(A_{ij})\right)append[L_{virt}\left[\Phi_{2virt}(M)\right]]\ \_$$

When $L_{\Phi\ virt}$ already has a value, then the new values are included and are appended to the existing ones. This means that in each entry of the virtual linear structure, there will be different values $(k_1,k_2,..,k_m)append[(k_1^{'},k_2^{'},..,k_m^{'})]append...append[(k_1^{'p},k_2^{'p},..,k_m^{'p})]$.

A concatenation of $(k_1,k_2,..,k_m)$ values will be included in each cell of the virtual linear structure. This linear structure has to comply with having all possible numbers up to a combination of benchmarks reasonably high. Each symbol $\{X_i \mid i=1,..n\}$ will have a benchmark. The combination of all the benchmarks will define the number of entries that the linear structure has. Each entry will store a concatenation of values $\{k_1.,..k_m\}$. These values will indicate the number of times that an evolution rule should be applied to an initial multiset in order to obtain an extinguished multiset. After proving the consistency of the physical structure [Arteta, Castellanos, Martinez 2010] proving the consistency of the virtual one is trivial.

Now there is enough information to build the physical linear structure L_Φ in this example. A way to build the structure could be:

At this moment we are able to build the virtual linear structure. Each cell a concatenation of values will be stored instead of just one value.

The structure is built as follows:

$L_{\Phi virt}(x)=$

$L_{\Phi virt}(0,1)=0,0,1\ L_{\Phi virt}(1,0)=0,0,1\ L_{\Phi virt}(1,1)=0,0,1\ L_{\Phi virt}(2,0)=0,0,2$

$L_{\Phi virt}(2,1)=0,1,0\ L_{\Phi virt}(2,2)=0,1,0 \quad 0,0,2\ L_{\Phi virt}(3,2)=0,1,0\ L_{\Phi virt}(3,3)=0,1,0$ In this way, the

$L_{\Phi virt}(4,3)=1,0,0 \quad 0,1,1L_{\Phi virt}(5,4)=1,0,1 \quad 0,1,2 \quad 0,2,0....$

structure will be created and allocated in the virtual memory.

Algorithm

```
$X, Y \leftarrow Multiplicity(R(U, T))$
$BEGIN$
$output(L_{\Phi}(X, Y))$
$END$
$(L_{\Phi}(X, Y)) \leftarrow random(L_{\Phi virt}(X, Y))$
```

The algorithm search in the virtual structure the position(*X,Y*) which are the input values corresponding to the multiplicities of the initial multiset. When the value is returned, a new value coming from the virtual structure overwrite the value stored in the position (*X,Y*).

Conclusion

This paper is a continuation of [Arteta, Castellanos, Martinez 2010]. Here, the author analyzes the resources to use to build a structure in the RAM of the computer. This paper describes a method to do so in a virtual memory device. When there are limitations on physical memory, a proper use of the virtual one is recommended. The way to build the structure determines the proper performance of the method.

Bibliography

[Arroyo, Luengo 2003] F. Arroyo, C. Luengo, A.V. Baranda and L.F. Mingo. A software simulation of transition P systems in Haskell. International Workshop Membrane Computing, Curtea de Arges (Romania), August 2002, Springer-Verlag, Vol 2597, pp. 19-32, Berlin, 2003.

[Arteta, Castellanos, Martinez 2010] A. Arteta, A, Castellanos, A. Martinez. Membrane computing: Non deterministic technique to calculate extinguished multisets. International Journal "Information Technologies and Knowledge", Vol. 4, Number 1, pp. 30-40. 2010.

[Arteta, Fernandez, Gil 2008] A. Arteta, L.Fernandez, J.Gil. Algorithm for Application of Evolution Rules based on linear diophantine equations. Synasc 2008. Timisoara Romania September 2008.

[Ciobanu, Pérez-Jiménez, Ciobanu, Păun 2006] M. Pérez-Jiménez, G. Ciobanu, Gh. Păun. Applications of Membrane Computing, Springer Verlag. Natural Computing Series, Berlin, October, 2006.

[Ciobanu, Wenyuan 2004] G. Ciobanu, G. Wenyuan. A P system running on a cluster of computers. Proceedings of Membrane Computing. International Workshop, Tarragona (Spain). Lecture Notes in Computer Science, Vol 2933, Springer Verlag, pp. 123-150, Berlin, 2004.

[Fernández, Martínez, Arroyo, Mingo 2005] L. Fernández, V.J. Martínez, F. Arroyo, L.F. Mingo. A Hardware Circuit for Selecting Active Rules in Transition P Systems. Proceedings of International Workshop on Theory and Applications of P Systems.Timisoara (Romania), September, 2005.

[Pan, Martin 2005] L. Pan, C. Martin-Vi de. Solving multidimensional 0-1 knapsack problem by P systems with input and active membranesl. Journal of Parallel and Distributed Computing Volume 65 , Issue 12 (December 2005)

[Păun 2000] Gh. Păun. Computing with Membranes. Journal ofComputer and System Sciences, 61(2000), and Turku

Center of Computer Science-TUCS Report nº 208, 1998.

[Păun 2005] Gh. Păun. Membrane computing. Basic ideas, results, applications. Pre-Proceedings of First International

Workshop on Theory and Application of P Systems,Timisoara (Romania), pp. 1-8, September , 2005.

[Qi, Li, Fu, Shi, You 2006] Zhengwei Qi, Minglu Li, Cheng Fu, Dongyu Shi, Jinyuan You. Membrane calculus: A formal . method for grid transactions. Concurrency and Computation: Practice and Experience Volume18,Issue14 , Pages1799-1809. Copyright © 2006 John Wiley & Sons, Ltd.

Authors' Information

Alberto Arteta *–Associate ProfessorTechnical University of Madrid. aarteta@eui.upm.es*

Angel Castellanos *–Departamento de Ciencias Basicas aplicadas a la Ingeniería Forestal. Escuela de Ingeniería Técnica Forestal. Technical University of Madrid, Avda. de Ramiro de Maeztu s/n 28040 Madrid, Spain. angel.castellanos@upm.es*

Nuria Gómez Blas*–Departamento de Organización y Estructura de la Información. Escuela Universitaria de Informática. Technical University of Madrid, Crta DE Valencia KM. 7, 28031 Madrid, Spain. ngomez@eui.upm.es*

HIGAIA METHODOLOGY

A. Anguera, A. Gutierrez, M.A. Diaz

Abstract: At present there is a deficiency in the field of scientific theories that support software development. On the other hand, the few existing scientific theories do not provide methodological support for all phases in software development. It is necessary to combine both aspects and develop a methodology, supported by a scientific theory, which extends these methodological support to all phases of software life cycle.

The proposed Software Development methodology combines Holons and Informons theory with GAIA, a well known methodology in the field of MultiaAgent Systems (MAS).

The elements defined in scientific theory are used in the description of the software development phases included in GAIA, extending them to complete the software life cycle.

Analysis, Architectural Design and Detailed Design phases of GAIA have been completed with Requirements Elicitation and Implementation phases, the latter based on the AUML standard.

In this way we obtain a complete methodology supported by a scientific theory that allows develop software systems based on Holonic Integrated Systems (HIS).

Keywords: HIGAIA,HIS, HIP, Holon, Models, MAS, Software Development Methods.

ACM Classification Keywords: C.2.4 Dsitributed Systems – Distributed applications, D.2.11 Software Architectures-Languages

Introduction

There are several schools of thought on the current state of scientific theories that support software development. All of them describe a slightly optimistic panorama.

Some authors understand that while the scientific theory is not developed, this is not the same with the engineering component, which itself is developed but in many cases without the required scientific theoretical support. The solution might be to address the software development taking into account that is treated as an engineering discipline but in any case must be supported by an underlying scientific theory.

There is a emerging scientific theory proposal to support the software development made by Alonso [Alonso 2004] where are introduced the holon and informon concepts as basic elements of this theory. Later, Martinez [Martínez, 2008] develops these concepts in the theory of holons and informons for software development , which is used by Suárez [Suárez, 2009] to formulate a specific proposal on the scientific theory on computational and conceptual model of software development.

According to [Suárez, 2009], the basic element of information that makes sense for a holon and allows to make decisions and implement appropriate actions is called "informon". The informon can take the form of facts, news or knowledge.

On the other hand, a "holon" can be generically defined as an element in its own right which has an autonomous, cooperative and self-organized behavior.

Autonomy means the ability of an entity to create their own behavior, that is, create their own plans and behavior strategies and monitoring execution as well as their own state.

Cooperation in this context is a process by which a set of entities develops plans, commonly accepted, that run in a distributed manner. Finally, it should be noted that one of the main characteristics of holons is their self-replication, which provides the ability of recursion and self-organization.

In building software terms, a holonic computational model consists of different levels [Martinez 2008], as detailed below:

- *Instruction*: Primary holons that are their own and cooperatives entities. They deal only with data and produce new data or simple news. They are specialized and perform basic operations.
- *Component*: The result of a holarctic structure (hierarchical/heterarchical) of primary level "instruction" holons. The holon component has a functionality greater than the sum of "instruction" holons and has capacity to produce more elaborate news and/or knowledge.
- *Entity*: Holarctic relationships formed by holon components. An "entity" holon presents beliefs, motivations and intentions and change their behavior based on previous experience. Incorporates proactive property, so it is able to act on their own initiative to achieve goals that itself is able to generate. An "entity" holon deals with news and knowledge informons. It can act as a software agent.
- *Organization*: A holonic organization is a collaborative holarchy entity. An "organization" holon means a formal and stable group of entity holons and provides a well defined interface for external communication. The activities of each organization are determined by cooperation processes with other entities or organizations. Presents a common goal and the entity holons, which have their their own goals, have a cooperative behavior led to organization common goal. An organization holon is similar but not equal, to a MultiAgent System (MAS).
- *Domain*: Is the logical space in which entity or organization holons, each with their own goals within the domain, operate and communicate with each other. Providing the context in which these holons may locate, contact and interact with others. A domain can be simple, anyone committed to working with any other following a set of established interaction protocols; or complex, designed to achieve a goal shared by several holons. In the domain, the holons are viewed as members of a society whose global activities should be contemplated. The social structure of the domain is determined by roles, social relationships identified between holons and social laws that govern these relationships. Holons are incorporated in the domain to achieve their own goals, and accept or reject their entry depending on which role they used to participate, if appropriate for this domain and accept conventions and social laws.

Software engineering has gone through different stages marked by the programming paradigms. Since the structured approach or object-oriented approach to the Unified Process. Today there are great challenges to these standards resulting from new technologies such as distributed Web services and agents. Agents are computational entities that differ from objects in several respects, such us autonomy, social ability and proactivity among others [Wooldridge, 1995], [Franklin, 1996], [Hernandez, 1999]. All these properties defined for the agents, and more, can also be attributed to the holon element defined in the holon and informons theory.

The presented scientific theory provides the basic tools for modeling computer systems being a formal support for the development of such models, but is in early maturity stage and does not include some software life cycle phases needed to complete development of such systems. This paper presents a methodology having its roots in the abovementioned theory approach to the software design phase. Taking as its starting point the main element of this theory, the holon, to make a comparison between an entity holon organization and a software agent. It can be inferred that they share more features than differ. If we accept this assertion and focus on a more complex relationship form, holons in Organizations and agents in MultiAgent Systems (MAS) found that both forms of organization share many similarities. An holon organization may seem like a MultiAgent System (MAS), although they are not alike.

HIGAIA

The general diagram that defines the phases of the HIGAIA methodology builds on the general model proposed in GAIA [Wooldridge, 2000], [Zambonelly, 2000] within the field of development methodologies Multiagent Systems (MAS) [Henderson, 2005]. Developing and adapting the methodology to the elements of the holons and informons theory for software development. The following figure represents the GAIA phases taken as a starting point.

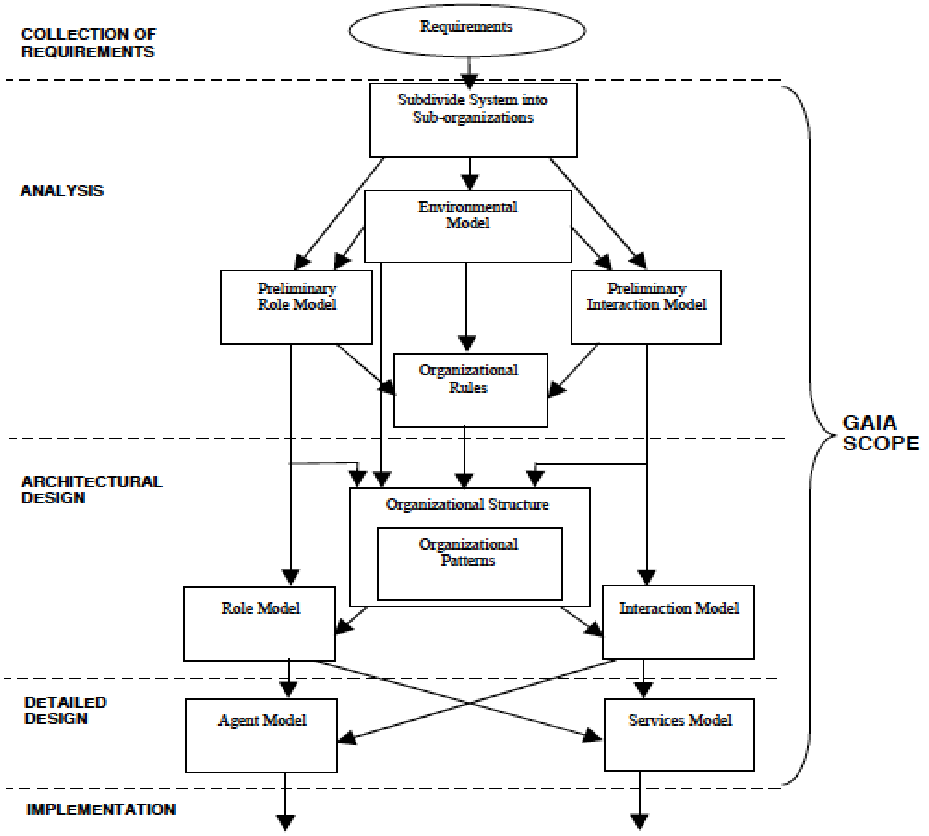


Figure 1: Phases of software develop.

As shown in Figure 1, the model consists of five phases outlined below.

1. *System Requirements elicitation phase*: The goal of this phase is to obtain a detailed specification of the information system that meets the users information needs and serve for further analysis and system design.

2. *Analysis phase*: The analysis stage output systematically documented all the functionalities and features that the system has to express, together with the characteristics of the operational environment in which it is located. Therefore, the main objective of the analysis phase is to organize the specifications and system requirements to become an environmental model. The preliminary roles, interaction patterns, and a set of rules of organization, for each of the sub-organizations that make up the system in general. This phase is divided in turn into five phases.

   2.1. Determination of the organizations. The first phase of HIGAIA analysis deals with the identification of multiple organizations that exist in the system and interact to become an Integrated Holonic System (IHS). The identification of these organizations is reasonably easy if (i) system identifies naturally or (ii) the system mirrors the structure of the real world, and it involves multiple agencies that interact. An organization consists of a set of related roles, these roles will be assigned to holons depending on skills and abilities of the latter. A set of holons can be grouped into sub-organizations, each of which will have its own goals and cooperate to achieve the overall goal of the organization.

   2.2. Environment Model: A IHS is always in an environment and we believe this should be considered as a primary abstraction for the analysis and design phases. In general terms, identify and model the environment is to identify all entities and the resources the IHS can exploit, control or consume when you are working toward the goal of the organization. This model consists of a list of abstract computational resources that can be drawn at this stage of design. Collect the system inputs, outputs and resources associated with sub-organizations identified in the previous phase. It should include active components that interact with the organization.

   2.3. Preliminary model of roles: The analysis phase does not result in the design of the current SHI organization (this is the purpose of the subsequent architectural design phase). However, even without knowing what the structure of the organization will, it is possible, in most cases, be able to identify some characteristics of the system that remains independently the final organizational structure. This phase detects the basic skills (preliminary roles) that organization requires to achieve its goals. To represent these preliminary roles uses a semiformal and abstract description to describe their capabilities and expected behaviors. Using two main class features respectively, permissions and responsibilities. The latter are subdivided into two types of properties, liveness and safety. The atomic components of an vitality expression are protocols and activities. Respectively are short-term actions that require interaction with other roles and can be done without third-party roles.

   2.4. Preliminary interaction model: Define the interactions by a reference to an institutionalized model of exchange of messages, such as FIPA (The Foundation for Intelligent Physical Agents), FIPA-Request [FIPA 1964]. This model captures the dependencies and relationships between the different organization roles, specifying the characteristics of the protocols concerned. Interaction organization models describe the protocols that govern the interactions between roles. Moreover, the interaction model describes the characteristics and dynamics of each protocol (for example, when, how and by whom a protocol should be executed). As the role model is still preliminary at this stage, the protocols for this model must necessarily be preliminary, for the same reasons.

2.5. Organizational rules: preliminary functions and interaction models capture the basic features, functions, and patterns of interaction that must realize in the MAS, irrespective of any pre-defined organizational structure. But at this stage can be captured functions and protocols to the organization. Those the rules that control the overall relations between the different roles of the organization and interactions between different protocols. Among others, the organizational rules establish the order of execution of the roles, time constraints in the execution of the same role for different entities and the maximum number of times you can run a role.

3. *Design phase*: *Architectural Design*. The specifications provided in the analysis phase should be structured and used in architectural design to identify an efficient and reliable SHI's organizational structure, and thus complete the preliminary tasks and interaction models. It is worth emphasizing that although the analysis phase is primarily concerned with understanding what the SHI will be, the design phase is where decisions are made to bear on the actual characteristics of SHI. At this stage, there are three different phases:

3.1. Organizations structure: The choice of the organizational structure is a very critical stage in the development of SHI, which affects all subsequent stages. Unfortunately, as always with the architectural design, it is not possible to identify a precise and formal methodology with which to obtain the "best" design. However, a design methodology can and should provide guidelines to help designers make decisions. In this case seeks to achieve maximum simplicity and organizational efficiency. As a general rule, choose the simplest topology that enables the organization to properly handle both the computational complexity and the complexity of coordination between different actors. Another factor that impacts the choice of organizational structure is the need for the SHI should respect rules of organization and be able to adopt them for implementation. We analyze each of the organizations that have been described in the analysis phase, as well as control schemes adopted (dependence, equality, or control) between the roles that are part of each, in order to maximize efficiency and simplicity of the system. If you increase significantly the number of members of different organizations, would opt for a multilevel hierarchical system in which some roles may partially control the actions of other roles.

3.2. Ultimate role model, complete the set of roles defined in the initial roles model with different organizational structures designed in step 3.1.

3.3. Final interaction model: the incorporation of new roles in the previous phase, implies a redefinition of the initial interaction model.

4. *Design Phase: Detailed Design*. This phase is responsible for the eventual identification of the type of holons and service model that, in turn, act as guidelines for actual implementation of the agents and their activities. The objective of this phase is to define two models:

4.1. Holons model. HIGAIA context distinguish between holon and holonic organization. A holon is a software entity that runs a set of roles. Therefore, the definition of holon model will determine what kinds of holon are defined to perform specific roles and how many instances of each of them must be implemented in the real system. Normally there will be a one-to-one relationship between the roles and holons so that a given role is taken by a holon responsible for implementing that role. The adoption of a role for a particular holon depends on the skills of the holon that can deliver the goals that represent the role.

An holonic organization shall consist of holons which are in turn dependent on the acceptance of social rules of the organization and should develop cooperative behavior led to the common goal. This model is described by one side each of the holons which are part of the system using Table 2 and the other holonic organizations that conform the system as seen in Table 1.

holonic organization class		Organization identifier.
Servicces	Name	Organization name.
	Goals (description)	Organization objective.
	Skills (description)	Holons capabilities as part of the organization to realize its goals
	preconditions	Conditions to be satisfied in the system just before the start of the execution of any element that is part of the organization.
	postconditions	Declare what should be true in the system when is to successfully complete the runnig of the organization.
	input	External ítems that the organization needs to running its behavior.
	output	items to get de organizations after running the organization behavior.
Strategies		Descriptions of algorithms or procedures that represent the behavior of the organization to reaching its goals.
Set of Roles		Roles that are part of the organization at all times.

Table 1: Clase Organización Holónica

Holon class		Holón identifier.
Services	Name	Holon name.
	Goals (description)	Holón objetive.
	Skills (description)	Holon capabilities to reach their goals. A skill is always linked to a goal.
	preconditions	Conditions to be satisfied in the system just before the start of the running of the holon
	postconditions	Declare what should be true in the system when is to successfully complete the running of the holon.
	input	Elements takes to run your behavior holon
	output	items to get the holon after running the holon behavior
Strategies		Strategies
Set of Roles		Set of Roles

Table 2: Holón Class

4.2. Service model. This model allows us to identify the services associated with each holon in the system, or equivalent, with each of the roles that run the holon. Thus, the service model applies in the case of static allocation of functions to holons as well as in the case of functions they can assume dynamically. Must be defined for each of the holons that form part of the system (as defined in the previous phase, 4.1), indicating the inputs and outputs, as well as necessary preconditions or postconditions needed to perform the role assigned to holon. The services that compose a holon are derived from the list of protocols, activities, responsibilities and liveliness properties of the functions it implements. At one extreme, there is at least one service for each parallel activity that holon has to execute. However, even for sequential execution activities, there may be a need to introduce more services to represent the different stages of holon implementation.

5. *Implementation phase*. GAIA notations, probably due to their simplicity, are poor and far from being widely accepted as industry solutions (unlike UML in object oriented software engineering). This seems to be quite clear in the specification of the interactions between holons within HIGAIA methodology. In fact, the Gaia protocol model, despite rigorously specify the actors and the inputs and outputs of the protocol, use informal natural language to specify the semantics, including dependence and interaction of the actors involved. AUML is not itself a complete and comprehensive methodology, however, is based on the acknowledged success of UML to support software engineering for industrial processes. The AUML core are

interaction protocols between agents (AIP). The proposed methodology use an adaptation of AUML AIP interaction model to capture the interaction between holons (HIP- Holons Interaction Protocol). Will, therefore, the central element of the specification of the HIP, completing the present model in GAIA.

The following figure shows a generic class hierarchy, which can be used to implement any holonic system (SHI). As can be seen are established the essential elements of the organizational model as well as the relationships established between each of them.

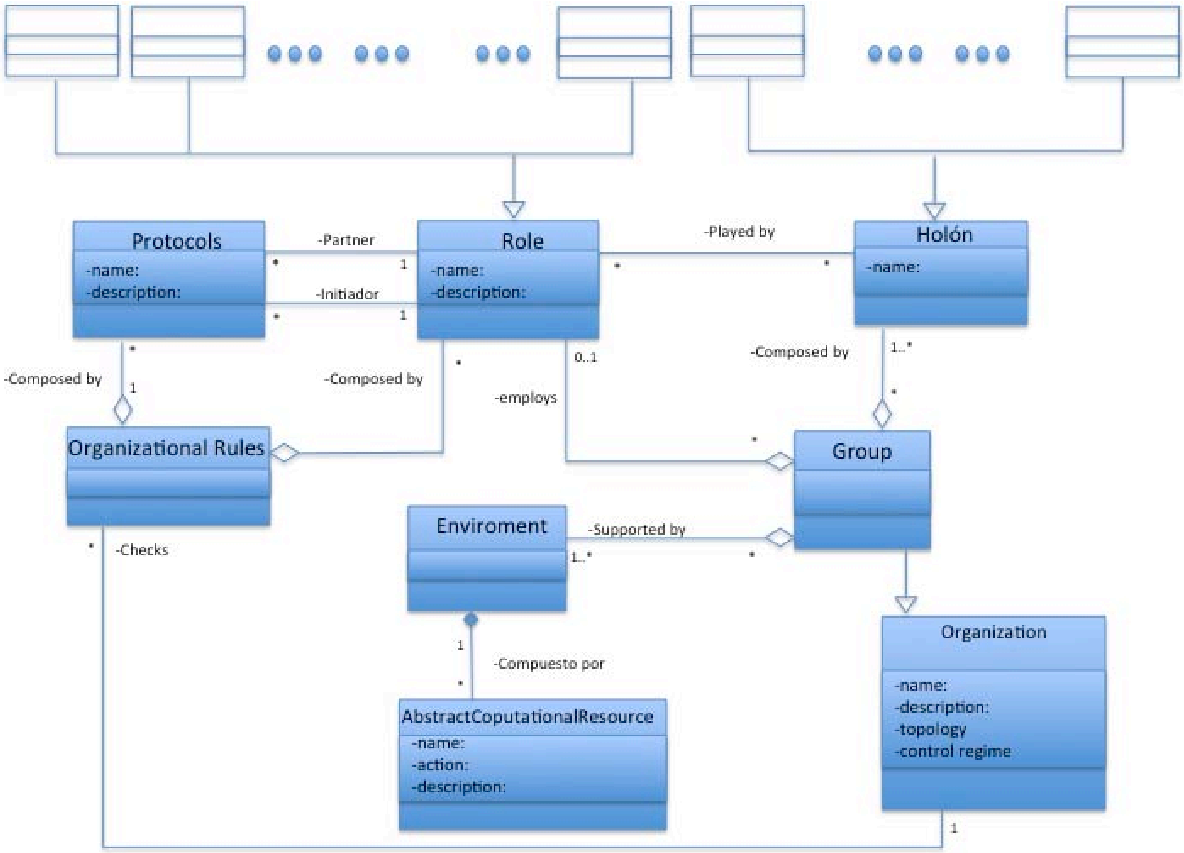


Figure 2: Generic hierarchical class model

Conclusions

This paper develops a methodology based on the scientific theory of holons and informons that unlike existing theories covers all phases of software development. Completing the scientific theories with development methodologies. Achieving a balance between the two sides of software, the science base and its development as seen from the point of software engineering.

HIGAIA methodology is proposed in order to be applied in any field where multi-agent systems (MAS) can be used. HIGAIA is a versatile methodology that allows easy application and adjustment to the software development industry. Finally, the HIGAIA methodology is independent of programming language chosen to implement the system.

Bibliography

[Alonso, 2004] Alonso F., López G., Pazos J., Rodríguez-Patón A., Silva A., Soriano F.J., Fundamental Elements of a Software Design and Construction Theory: Informons and Holons, Proceedings of the International Symposium of Santa Caterina on Challenges in the Internet and Interdisciplinary Research (SSCCII), Italia, 2004, pp. 21-35.

[FIPA, 1964] FIPA, 1964 The FIPA Specification Repository, http://www.fipa.org/repository/index.html Fletcher, R., & C.M. Reeves, "Functions minimization by conjugate gradients", Computer Journal, vol. 7, pp. 149-154, 1964

[Franklin 1996] Franklin, S., Graesser, A.: Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages. Springer-Verlag (1996).

[Henderson, 2005]. Brian Henderson-Sellers, P.G.Agent-Oriented Methodologies, 2005, Idea Group.

[Martínez, 2008] Martínez, M.A.: Una Teoría para el Desarrollo Software construída mediante técnicas y modelos de Gestión del Conocimiento. Tesis Doctoral. Universidade da Coruña. Mayo 2008.

[Suárez, 2009] Suárez, S.: La experimentación Crucial Basada en una Teoría de Holones e Informones para el Desarrollo del Software. Tesis Doctoral. Universidade da Coruña. Junio 2009.

[Wooldridge 1995] M. Wooldridge and N. R. Jennings, editors. Intelligent Agents --- Theories, Architectures, and Languages. Volume 890 of Lecture Notes in Artificial Intelligence, Springer-Verlag, January 1995.

[Wooldridge et al. 00] Wooldridge, M., Jennings, N. R., and Kinny, D., The Gaia Methodology for Agent-Oriented Analysis and Design, Journal of Autonomous Agents and Multi-Agent Systems, vol. 15 2000.

[Zambonelly, 2000] Zambonelly, F., Wooldridge M., and Jennings N.R., Organisational Rules as an Abstraction for the Analysis and Design of Multi- Agent Systems, International Journal of Software Engineering and knowledge Engineering, 2000

Authors' Information

Aurea Anguera de Sojo Hernández- Associate professor U.P.M. Crtra Valencia km 7, Madrid 28031, Spain;

e-mail: aanguera@eui.upm.es

Miguel Angel Díaz Martínez- Associate professor U.P.M. Crtra Valencia km 7, Madrid 28031, Spain;

e-mail:mdiaz@eui.upm.es

Abraham Gutiérrez Rodríguez- Asocciate professor U.P.M. Crtra Valencia km 7, Madrid 28031, Spain;

e-mail:abraham@eui.upm.es

DIGITAL ARCHIVE AND MULTIMEDIA LIBRARY FOR BULGARIAN TRADITIONAL CULTURE AND FOLKLORE

Radoslav Pavlov, Galina Bogdanova, Desislava Paneva-Marinova,
Todor Todorov, Konstantin Rangochev

Abstract: *In this paper we present investigation of methods and techniques for digitization and security in digital folklore archive - an archive that consists of unique folklore artifacts stored and annotated in the National center for non-material cultural heritage, Institute of Folklore, Bulgarian Academy of Sciences. The research is separated in several basic aspects. First we investigate techniques for digitization of different multimedia types - text, images, audio and video. We use this research to selected collections of artifacts. Second we describe several methods applied for securing the intellectual property and authors' rights. These include digital watermarking and error-correcting codes. The paper also presents the functional specification, implementation and testing procedures of the Bulgarian folklore digital library, where the digital folklore archive is kept.*

Keywords: *multimedia digital libraries, digital archive, systems issues, user issues, online information services, watermarking.*

ACM Classification Keywords: H.3.5 Online Information Services – Web-based services, H.3.7 Digital Libraries – Collection, Dissemination, System issues, K.6.5 Security and Protection.

Introduction

Information and multimedia technologies that have been developed during the past couple of years introduced new innovative approaches of documentation, maintenance and distribution of the huge amounts of information materials. Preliminary stages start with the content digitization and preservation with proven security methods. After that is the creation of content digital archives and its on-line publication in multimedia digital libraries: environments maintaining diverse hypertext-organized collections of information (digital objects such as text, images, and media objects) organized thematically and managed by complex specialized services for: content structuring, indexing, semantic annotation of digital resources and collections, content grouping, search (semantic-based search, multilayer and personalized search, context-based search), information retrieval, resources and collection management, metadata management, personalization and content adaptivity, multilinguality, content tracking, etc. [Pavlov et al., 2006].

In an attempt to answer the need of wider accessibility and popularization of Bulgarian folklore culture, a team from the Institute of Mathematics and Informatics has developed the "Bulgarian Folklore Heritage" archive and the Bulgarian folklore digital library (BFDL) within the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" national research project (Folknow). The "Bulgarian Folklore Heritage" archive aims to present Bulgarian folklore treasure kept in the funds of the Institute for Ethnology and Folklore and Ethnographic Museum (IEFEM) at the Bulgarian Academy of Sciences. The Bulgarian Folklore digital library (BFDL, available at: http://folknow.cc.bas.bg/) represents a complete web-based environment for registration, documentation, access and exploration of a practically unlimited number of Bulgarian folklore artefacts and specimens digitally included in the "Bulgarian Folklore Heritage" archive.

This paper presents the Folknow project, its vision and ideas. The digitization process and the approaches and tools for building the Bulgarian Folklore archives are included. The paper describes techniques for intellectual property rights digital protection of image and text data in the archive. The paper also presents the functional specification, implementation and testing procedures of the Bulgarian folklore digital library.

Folknow Project Overview

The "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" project started 6 years ago with fundamental research on contemporary technologies for virtual exposition of intangible cultural heritage. Main project's tasks are the creation of the "Bulgarian Folklore Heritage" archive using Funds of IEFEM at the Bulgarian Academy of Sciences by modern methods of digitization and analysis of selected collections and the release of a prototype of multimedia digital library for the registration, documentation, and access to archive's folklore objects. The digital archive of folk materials is created by knowledge-based technologies of storage, protection and effective access to data in order the modern presentation in a virtual form of valuable artifacts of the Bulgarian folklore heritage. The project aims to give various social applications of the folklore knowledge such as interactive distance learning/self-learning, research activities in the field of Bulgarian traditional culture, and for cultural tourism and ethno-tourism in Bulgaria [Bogdanova et al., 2006].

Digitization process

Digitization or digitize is creation of an object, image, audio, document, or a signal (usually an analog signal) from a discrete set of its points or samples. The result is called "digital image", or more specifically "digital images" of the object and "digital form of the signal. Digitization is the process of creating digital files by converting analog media which can be traditional archival materials (documents, plans, photographs) or three-dimensional objects such as museum artifacts or even works of architecture. Process of transformation of information flows from other types of traditional and digitized. As digitization is a method implementation of new approaches and means to convert some other sources in using information and communication technologies [Project Calimera].

The tasks of digitalization can be synthesized in certain key areas:

- retention of funds and records - many of digitalized objects are fragile or brittle structure is influenced by weather conditions and over time their digitization is the only hope for preservation;
- simultaneous access to materials - most objects are subject to the digitization of rare and unique items of historical past and have a priceless value, the process of digitization will allow more users involved to touch them;
- conservation funds in digital formats - archives, websites, digital libraries. Strengthening international exchange and promotion;
- providing access via computer networks - easy access to digital archives, access to records of persons with disabilities

- provide new opportunities to work with digitalized materials funds - all the functionality is available to users of web space to be copied, multiplied, forwarded, etc., without jeopardizing the its integrity and strength;
- full text search - digital archives organization contributes to easier detection of the searched object among all the crowd, advanced search, unification of search results;
- classification of digital funds via metadata - entire photo meta data wealth funds may be accompanied by important information about copyright, creation date, identification number, etc.

Following advantages of digital libraries can be listed:

- volume density - on one server machine can have the information contained in many conventional libraries;
- accessible - via a computer connected to the Internet, no matter what point of the Earth one has an instant access to desired information;
- quality - regardless of how long certain information stays - it does not change - and the text colors do not fade, the sound is not muted;
- speed - within seconds, minutes can crawl the entire array of data and information needed to remove certain parameters
- security - digital (electronic) data are independent of the medium and within minutes can be transferred to another medium, which makes them physically decoupled from the owner of the information and material damage;
- adequacy - data can be added and changed in the digital library, without stopping or rearrange work and record in real time;
- flexibility - possible structure of the digital archive can be changed without affecting the data and resources;
- simplicity - creating an intuitive and simple interface, work with the digital library will not require specific technical knowledge;
- performance - with minimal effort the user receives the desired information, reports, charts, etc.;
- diversity - the ability to depict different media such as photographs, text, sound and video recordings, three-dimensional images and animation;
- communication - user himself can choose the information he seeks;
- interactivity - an appropriate interface, a variety of information and accessibility of digital library expands the circle of users (children, students, people in difficulty, curious consumers, etc.);
- multitasking - multiple terminals, different users can simultaneously carry out different data operations: view, search, add, update, etc. simultaneously.

Development of digital technology hit the storage and processing of information. Today, almost every unit of information is created digitally: digital photography, digital sound recording, digital communications, text, saved in a file, videos, movies, multimedia presentations stored on digital media, etc. Storage of such digital multimedia data in digital archives became subject to the same challenges, such as archives, we know before the invention of digital computing devices.

Requirements for the parameters obtained after digitization of objects

Studies conclude that we must work in two directions. Next are given requirements for digitization with priority storage as much information content (left side of the tables) and digitization with priority WEB and network usage (left side of the tables).

TEXT: Text files do not occupy a lot of resources and there are no requirements for size or volume of supporting information. As requirements relating to criteria for accessibility and multiplatform could be specified: cyrillized table of characters is - Unicode, Windows cyrillic 1251 or UTF-8 and use the following file formats: TXT, DOC, PDF, RTF, HTML.

PHOTO: Following parameters to preserve the digital image quality can be used:

File size - from 2 MB to 50 MB	File size - less than 1 MB
Resolution - from 1200x800 px to 6000x4000 px	Resolution - 800x600 px
Resolution - 600 dpi to 2000 dpi	Resolution - 600 dpi
Number of colors - more than 16 bit color for each RGB channel of the scheme;	File formats: PNG or JPG
File formats: BMP, TIFF, PNG or JPG	

VIDEO: Parameters for digital video recording are:

File formats - A system for streaming video – PAL	File formats - Stream: MPEG-4, FLV
File size - up to 20 MB	Resolution of the picture - 300x200 px
Number of frames per second - 25	File size - up to 20 MB
Resolution of the picture - 720x576 px	
Ratio of the picture - 4:3	
Bitrate - from 2000 Mbps to 4000 Mbps	
File formats - DVD-Video, DV-AVI, MPG, AVI.	

AUDIO: To preserve the integrity of auditory information we use the following file format parameters:

File formats – WAV	File formats –MP3 (MPEG-1 Layer 3)
Audio format - PCM;	Sample rate - 44 100 Hz
Sample rate - 44 100 Hz	Bitrate - from 128 kbps to 320 kbps;
Bit resolution - 16 bit;	File size - up to 10 MB
File size - up to 100 MB	

Initial analysis of the volume in the archives of the National Center for Non-material Cultural Heritage (NCNCH) in IEFEM shows that there are about 10 million text documents, audio – 50 000 hours video – 20 000 hours, 100 000 photo images. It is expected that in the near future the folklore archive will increase the volume of stored information and the number of digital copies. With such a volume of data, following organization of the digital archive is needed:

- Tree file structure;
- Matrix of signature file name;
- META additional textual data for indexing of media files;
- Description for digital archive unit.

In our research we also consider some previous experience in the same area [Bogdanova et al., 2006], [Luchev et al., 2008], [Paneva et al., 2007]. We use several techniques for creation and analysis of digital archive [Boganova et al., 2008a], [Boganova et al., 2008b] [Noev, 2010].

Tree file structure - In selecting the file structure of the organization we keep in mind that the funds hold millions of NCNCH materials cataloged with signatures and future digitization should keep optimized the file organization. We choose the following organization for digitalized artifacts:

- Home folder is "NCNCH IF archive" – ("National Center for Non-material Cultural Heritage, IEFEM, archive");
- Next are folders to individual files, divided by type of media and content, before digitizing them: paper based, photo archive, audio records, video archive and CD archive;
- Next step is a division of media file format (WAV, mp3, AVI, etc.).
- Due to the large number signatures are divided into groups of 100 pcs. in separate folders;
- Next step is a folder for each signature number (archival unit), where digitalized materials can be founded.

Matrix for signatures of file names - For determining the name of files are taken into account:

- Signature numbers in NCNCH funds;
- In media files some information should be listed in brackets;
- Number of digital resource, if there is more than one recording;
- If the name of execution exists, it is written in Latin letters;
- At the end is extension of file format (bmp, jpg, WAV, mp3, AVI, pdf, doc, txt, etc.).

Description for digital data unit - Describing the content of an archival unit, whether digital or not, is essential for organizing the archive. We formulate following text fields as mandatory for description of digital objects from the NCNCH fund: name of the organization that has archive unit (in this case: NCNCH), signature number in the archive, place of recording, date of entry, recorder, theme of the study. And additionally for description of an files: name of the file, parameters of the digital record, date of digitization, underlying digitization, type content of the material.

Establishment of the archive using Fotoware FotoStation - In establishing the digital archive we should keep in mind that it has a tree file structure where all objects are arranged. FotoStation program allows an archive to a multitude of primary and sub directories (see http://www.fotoware.com/en/Products/FotoStation/). The program has built in powerful file editor providing all functions needed for working with files located on a computer system.

Techniques for Intellectual Property Protection in the Bulgarian Folklore Heritage Archive

With the development of digital technologies increasingly part of the audio, video and any other information is available for fast, easy and high quality copying. This fact entails the problem of protecting information from unauthorized distribution. Research in this area is considered in several aspects. One of the most important of these is steganography [Gribunin et al, 2002]. Unlike other subfields such as cryptography, dealing with the concealment of the message itself, steganography tries to hide the fact that the built in message exists. Like steganography, watermark protection, aims carry hidden information but have more resistance against attempts to remove the embedded information. Digital watermark is visible or preferably invisible to the identification code that is permanently embedded into digital data and maintained a presence in them after extracting it [Cox et al, 1997]. Next we describe methods that we used to digitally watermark image, audio and text data in the Bulgarian Folklore Heritage archive.

Methods for image watermarking in the spatial domain

In these methods data are incorporated directly into the original image. The watermark is embedded by changing the illumination or color components. An example of this method is the method of Kutter [Kutter, 1998]. Let s is the bit is that we want to incorporate in the image $I = (R;G;B)$ (color channels; R - red, G - green, B-blue), and $p = (i; j)$ is a pseudo (obtained by the random number generator) position in I. This position depends on the secret key K, which is used as the core for a generator of random numbers. The bit s is embedded by modifying the blue channel B at position p, using illumination $L = 0, 299R + 0,587G + 0, 114B$ by $B_{ij} = B_{ij} + (2s\text{-}1)L_{ij}q$, where q is a constant. The value of q is determined so as to achieve the best balance between stability and invisibility. To derive the integrated information we should made an assumption based on a linear combination of pixels around p. To derive the value of the embedded bit is calculated difference between assumed value and actual value of the pixels. The sign of the difference determines the value of the embedded bit. Extracting bits is done without the knowledge of the original message. Sustainability can be improved with the use of a code error correction. The method is robust to filtering, JPEG compression and geometric transformations.

Methods for audio watermarking with amplitude modulation

A basic approach to watermarking is to encode the information into the least-significant bits of the audio data. There are two basic classes of ways to do this: you can replace the lower order bits completely with a PN-sequence, or you can embed a PN-sequence into the existing low-order bitstream. This technique works in the time domain by changing the amplitude of the audio data in a way that can be recovered given the PN-sequence. Variations on this approach include adaptively attenuating the amplitude of the embedded sequence to match the sound level of the current sound passage, and shaping the PN-sequence itself to match the underlying psychoacoustic masking characteristics to further bury the signal.

Methods for text watermarking using encoding by row offset

This is a method, in which lines of text are displaced vertically so that the document can be uniquely encoded [Brassil et al, 1994]. In most cases, the decoding can be performed without the use of the original document, if it is known that the primary document is the same distance between rows. The method is easily noticeable, but resistant to noise.

We use techniques described above to improve security of data in the archive with digital watermarks. As an improvement to these watermarking methods we use error-correcting codes [Baudry et al., 2001]. Because of the specificity of protection with watermark, this problem remains open. Its solution requires the use of code that is as compact and resistant to different types of attacks [Katzenbeisser et al., 2000]. As an improvement to these watermarking methods we use error-correcting codes. We improve performance of the codes by using our own coding method [Berger et al., 2008]. This encoding makes embedded watermarks more resistible to attempts to remove the embedded information.

Functional Specification of the Bulgarian Folklore Digital Library

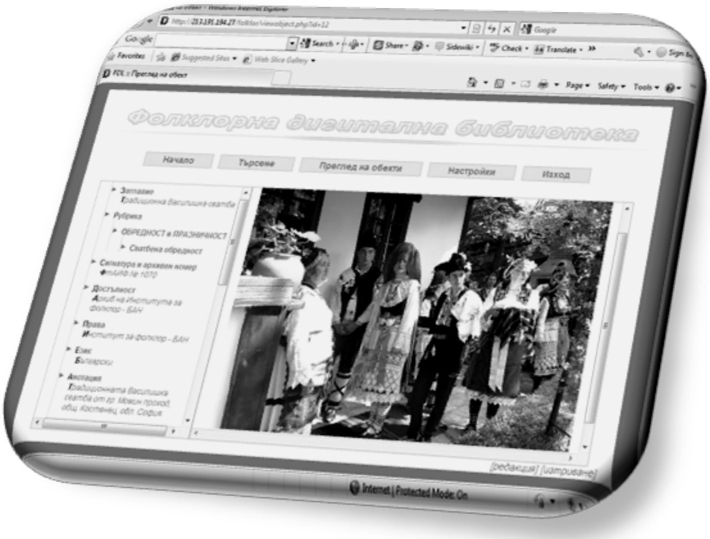
The Bulgarian Folklore digital library represents a complete web-based environment for registration, documentation, access and exploration of a practically unlimited number of Bulgarian folklore artefacts and specimens digitally included in the "Bulgarian Folklore Heritage" archive. It provides a rich knowledge base for the Bulgarian traditional culture and folklore, enabling its usage for content annotation, preview, complex search, selection, group and management [Paneva-Marinova et al., 2010] [Pavlov et al., 2010] [Paneva-Marinova et al., 2009] [Pavlova-Draganova et al., 2006].

The key for the current release of BFDL is the efficiency and the provision of strictly designed functionalities, powered by a long-term observation of the users' preferences, cognitive goals, and needs, aiming to find an optimal functionality solution for the end users. In BFDL we also follow the requirements of experts in the area of Bulgarian folklore and the accepted functional specification for a digital library. Following them the basic BFDL functional modules are:

- A module for adding and editing folklore objects. For the semantic annotation of the objects in this module is used special ontology describing the Bulgarian folklore domain [Paneva et al., 2007] [Luchev et al., 2008]. The "add object" form expects as an input two types of objects: simple folklore objects and complex folklore objects, strictly specified in the ontology.
- A module for viewing the content of folklore objects (according to their base type and rubric to which they belong or by different descriptive characteristics). Figure 1 shows a snapshot of a folklore object.
- A module for searching by: signature and archive number; keywords of the following categories: name, language, annotation, type of the folklore object/rubric; file type; record information (simultaneously or one by one): by situation, by reporter name, by recorder name, by record date and by recording location; extended search – it provides the option for searching through all the object characteristics;
- A module for managing the user data;
- A module for monitoring the user's actions, which keeps track of the following: a) Actions related to working with the system: registration, logging in the system, unsuccessful log-in attempts, logging out, changing of the user password, e-mail address change, etc.; b) Actions related to the object

manipulation: adding an object, editing an object, deleting an object, adding a file, deleting a file; c) Actions related to the content viewing: review of objects by their characteristics, view of a single object, searching for objects by characteristics; d) Other administrative actions: changing the user’s level, deleting a user, generation of an XML copy of the data in the system;

- A module for file format conversion;
- A module for generation of XML copies of the objects in the system.



Заглавие
Традиционна Василишка сватба

Рубрика
ОБРЕДНОСТ и ПРАЗНИЧНОСТ
Сватбена обредност

Сигнатура и архивен номер
АИФ № 1070

Достъпност
Архив на Института за фолклор - БАН

Права
Институт за фолклор - БАН

Език
Български

Анотация
Традиционната Василишка сватба от гр. Момин проход общ. Костенец, обл. София

Figure 1: Folklore Object Preview

The module for viewing the content of folklore objects is available to all users of the library, except the administrators. The reason is that the administrators of such systems are often people who don’t have any relation to their content; they only do support tasks. The module itself was implemented similar to the Windows OS file browser and KDE, so that it is closer to the familiar user interfaces for viewing hierarchical information. The left side shows a tree of all classes, which inherit “Type of folklore object”, and the right side shows a list of objects of the selected class in the tree.

The module for creating and editing folklore objects is used for adding new objects and modifying the information of already created objects. Through it, one can add more multimedia files to an object or delete existing ones.

Searching for information is the most frequent search and therefore the most important operation in a digital library. This is why there are several modules for searching by different criteria:

- Searching by a signature or archive number – This search module is useful for finding objects by their archive number (for example, AIF No 200, folder 1, page 57). In general, there is only one search result. In case of incorrect data, it is possible to have several objects as a result.
- Search by a keyword in the object properties – by name, language, annotation and type of the folklore object – Searching is performed simultaneously over all these properties. It is expected that this module is the most frequently used one. This is why special attention has been paid to its optimization.

- Searching by record information – This module is used to find all the objects which cover some of the following conditions: all the objects recorded in a given situation, for example an interview, chat/conversation, etc.; all the objects recorded by a given person; all the objects recorded by a given informer; all the objects recorded in a given period of time; all the objects recorded in a given location.
- Searching by file type – This module allows getting a list of all the objects to which there is a multimedia file attached – audio, video or images. This type of searching uses the database in which the administrative information is stored instead of the OWL file that contains the ontology.
- Complex search on all fields semantically describing the folklore object. Using this search simple and complex folklore objects could be found, tracking their semantic metadata records.

Most types of searching use SPARQL (SPARQL Protocol and RDF Query Language). This is a language for requests to the RDF and OWL ontologies. The language is in a standardization process by RDF Data Access Working Group as an official recommendation of the World Wide Web Consortium. The SPARQL syntax is similar to the most widespread language for database requests – SQL.

The module for monitoring the user's actions is intended to keep logs of the objects modified and deleted by the users, so that in case of data deleted by mistake or entered wrongly, the responsible user can be found. There is also a log of search requests, whose purpose is to enable statistical reports about the search types that are used least and most often. It would allow the removal of the rarely used search types and the priority optimization of the ones that are used most often.

The module for file format conversion was developed to provide the ability to present every file which is unsuitable for internet preview in a "light" and convenient form for web preview. The module recognizes the "inconvenient" files, tries to covert them and on success replaces the original file with the new "lighter" file; on failure, the module keeps the original file in the library. The module for generating an XML copy of the data is available only to the system administrators. The purpose behind this module is creating a copy, which can be used as an archive copy on one hand and on the other hand it may serve as raw data for other systems using information from the library.

The presented BFDL functionality aims to serve a wide range of users – specialist and non-specialist. The group of specialists is composed by scientists who study Bulgarian folklore professionally and search for specialized information on the observed folklore objects. The group of non-specialists has interests and wants only to learn more about the classical Bulgarian folklore objects. The BFDL system supports several users' levels: administrators, folklore content editors, specialist viewers and non-specialists viewers. Their individual characteristics, needs, interests, motivation, and preferences are discussed in [Pavlov et al., 2006].

Implementation of the Bulgarian Folklore Digital Library

The implementation specifics of the functional modules of the BFDL are the following:

A module for adding objects to the BFDL – Adding objects is implemented through filling and sending a form to the web server. Because of the great number of fields to fill, the form is not generated completely. Only the fields necessary for the creation of the objects are generated, following the semantic descriptions presented in the BFO, built at the first stage of module 3 of the project. The technology used for the implementation is AJAX. The user interface passes a request to the server, in which it requires only that part of the form which according to the user is necessary to create the object. The server processes the request and returns the required fields as a result, which is visualized in the user interface. After all the fields are filled, the user submits the form. The server validates the data and if everything is correct, it adds the object to the data storage. If there is something wrong, it returns a message to the user, relative with the error (usually, an empty field or unacceptable field value). After the server adds the information from the form to the data storage, there follows a check for attached files in the user request. If there are attached files, the server checks if there are file formats which are unsuitable for web presentation (for example, wav, .doc, .mpg, .avi, .mpeg, etc.) and if it finds such files, the system refers to the module for file format conversion to formats suitable for web preview. For each of these files, the module for file format conversion tries to convert them. Upon success, it adds the converted file to the library. On failure (which can occur if the added file has any specifics which the system cannot recognize), it adds the original file to the library. At the end of the object adding procedure, the system refers to the module for monitoring the user actions, where it adds an “object added” event and records the author (the user who created the object) and the event date.

A module for editing objects in a BFDL – The module for editing objects works almost in the same way as the module for adding objects. The difference is that the system doesn’t add information about a new object, but replaces the existing information about an object with the new information, provided by the module for editing. Again, the system checks the form for errors, processes the files (if there are new files added), changes its data and finally adds an event for modified object through the module for monitoring the user’s activity.

A module for viewing the content of folklore objects – This module takes a request from a user, in which the user specifies the property by which a folklore objects must be found. The module refers to the data storage and makes a request for selecting and sorting the objects by this property. The module for monitoring the users’ actions records the “view objects by” event and adds data about the date, the user and the property by which objects are listed. The storage processes the request and returns a result, which the system processes and sends to the user. The user interface visualizes the result in a proper manner.

A module for searching – This module allows the user to set a property or properties by which objects are found. The following algorithms are used:

The algorithm for searching by a single property – The user interface sends a request to the data server specifying the property and its needed value. The module for searching refers to the data storage of semantic metadata with a query for selection and sorting the objects with the needed value of the specified property. The module for monitoring the user actions records the “search” event with the provided search parameters, the date and the user, who performs the search. The storage processes the request and returns a result, which is then processed by the search module and displayed in a proper manner by the user interface.

The algorithm for searching by more than a single property – The algorithm is parallel to the one described above, with the only difference that the query to the data storage is more complicated – there are multiple selections of objects for each search property and the result is a sorted section of the selection results.

After an analysis of the means and standards in the technological implementation of the library environment and the functional modules, the following software was chosen: Operating system: Microsoft Windows Server 2008 x64 Standard; Web server: Apache HTTP Server v 2.2, PHP v 2.2.9; Database management system: MySQL v

5.1 Standard; Tools for the additional modules: FFMPEG, vwWare, HTML, JavaScript, AJAX; Database query language: SPARQL.

Testing Procedures in the Bulgarian Folklore Digital Library

The functional components of the architecture of the BFDL were implemented and tested for errors and speed on a server platform with the following hardware configuration: CPU: 2 x Intel QuadCore 2.8 GHz; RAM: 8GB DDR3; HDD: 4 x 500GB, RAID 10 SATA II; LAN: 2 x 1000Mbit.

Testing the functional module for adding/editing a folklore object – Server response time (average of 50 attempts): 0.0058 s, i.e. in theory the functional module for adding/editing an object can process about 172 requests per second for each processor core, which makes 172*8=1376 requests.

Testing the module for viewing folklore objects – Time for server response: 0.009 seconds per request, i.e. 888 requests per second.

Testing the module for searching by a single property – Time for server response: 0.008 seconds per query, i.e. 1000 requests per second.

Testing the module for searching by several properties – The test was performed with 25 properties (it will happen very rarely). Time for server response: 0.01 seconds per query, i.e. 800 requests per second.

Testing the module for file format conversion – Converting video files: the server sends a response before it converts the video file, because the process is relatively slow. The average time of processing a video file is about 30 seconds, i.e. you can add about 16 video objects per minute. In this way, after adding a video object, its actual recording in the BFDL happens in 30 seconds.

Converting audio files – The server responds before the file is actually processed. The average time for processing an audio file is about 10 seconds, i.e. in theory a system with such a configuration can process about 48 audio files per minute.

Converting MS Word (.doc) files – The conversion takes place in real time. The average server response time is 0.04 seconds per request, which are about 200 requests per second.

Conclusion and Future Work

Digitizing and presenting our traditional culture and folklore in virtual exposition through digital libraries we enable "any citizen to access human knowledge anytime and anywhere, in a friendly, multi-modal, efficient, and affective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices" [Brainstorming report, 2001]. Moreover, during the presented project we give various innovative social applications of the folklore knowledge: interactive distance learning/self-learning, research activities in the field of Bulgarian traditional culture, cultural tourism and ethno-tourism in Bulgaria, dissemination of the national folklore treasure through social networks, etc.

During the last project year the work continues with the development of a Bulgarian folklore information artery, providing ability for users to collaborate and interact with each other in a dynamic, user-centred environment, maintaining social media dialogue in the folklore domain.

Acknowledgements

This work is supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant №IO-03-02/2006 “Development, Annotation and Protection of a Digital Archive “Bulgarian Folklore Heritage”” and №IO-03-03/2006 “Development of Digital Libraries and Information Portal with Virtual Exposition “Bulgarian Folklore Heritage”” from the project “Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage”.

Bibliography

[Baudry et al., 2001] Baudry, S., Delaigle, J.F., Sankur, B., Macq, B., Maitre, H., Analyses of Error Correction Strategies for Typical Communication Channels in Watermarking. Signal Processing, pp. 1239-1250, 2001.

[Berger et al., 2008] Berger, T., Todorov, T., Improving the Watermarking Process with Usage of Block Error-Correcting Codes. Serdica Journal of Computing, 2008, Vol. 2, pp. 163-180.

[Bogdanova et al., 2006] Bogdanova, G., Pavlov, R., Todorov, G., Mateeva, V., Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage. Advances in Bulgarian Science Knowledge, National Center for Information and Documentation, 2006, Vol. 3, pp. 7-15.

[Bogdanova et al., 2008a] Bogdanova, G., Georgieva, Ts., Using Error-correcting Dependencies for Collaborative Filtering. Data and Knowledge, 2008, Vol. 66, Number 3, Elsevier.

[Bogdanova et al., 2008b] Bogdanova, G., Todorov, T., Georgieva, Ts., New Approaches for Development, Analyzing and Security of Multimedia Archive of Folklore Objects. Computer Science Journal of Moldova, 2008, 16, 2(47), pp. 183-208.

[Brainstorming report, 2001] Digital Libraries: The Future Directions for European Research Programme, Brainstorming Report, San Casioano, Italy, 2001.

[Brassil et al, 1994] Brassil, J., Low, S., Maxemchuk, N., O'Gorman, L., Electronic Marking and Identification Techniques to Discourage Document Copying. In the Proceedings of IEEE INFOCOM '94, 1994, Vol. 3, pp. 1278-1287.

[Cox et al, 1997] Cox, I., Kilian, J., Leighton, T., Shamoon, G., Secure spread spectrum watermarking for multimedia. In Proceedings of the IEEE International Conference on Image Processing, 6, 1997.

[Gribunin et al, 2002] Gribunin, G., Okov, I., Turincev, I., Cifrovaia Steganographia. Solon-Press, 2002.

[Katzenbeisser et al., 2000] Katzenbeisser, S., Peticolas, F., Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, 2000.

[Kutter, 1998] Kutter, M., Digital Signature of Color Images Using Amplitude Modulation. Journal of Electronic Imaging, 1998, pp. 326-332.

[Luchev et al., 2008] Luchev, D., Paneva, D., Rangochev, K., Use of Knowledge Technologies for Presentation of Bulgarian Folklore Heritage Semantics. International Journal Information Technologies and Knowledge, Vol. 2, Number 4, 2008, pp. 307-313.

[Noev, 2010] N. Noev, Organization and Security of the Audio and Video Archive for Unique Bulgarian Bells, Mathematica Balkanica, NewSeries Vol. 24, 2010, Fasc.3-4, 2010, pp. 285-291.

[Paneva et al., 2007] Paneva, D., Rangochev, K., Luchev, D., Knowledge Technologies for Description of the Semantics of the Bulgarian Folklore Heritage. In the Proceedings of the Fifth International Conference Information Research and Applications i.Tech, Varna, Bulgaria, 2007, Vol. 1, pp. 19-25.

[Paneva-Marinova et al., 2010] Paneva-Marinova, D., Pavlov, R., Rangochev, K., Digital Library for Bulgarian Traditional Culture and Folklore, In the Proceedings of the 3rd International Conference dedicated on Digital Heritage (EuroMed 2010), 8-13 November 2010, Lymassol, Cyprus, Published by ARCHAEOLINGUA , pp. 167-172.

[Paneva-Marnova et al., 2009] Paneva-Marnova, D., Pavlov, R., Rangochev, K., Luchev, D., Goynov, M., Toward an Innovative Presentation and Creative Usage of the Bulgarian Folklore Wealth, International Journal „Information Technologies & Knowledge”, Vol. 3, Number 1, 2009, pp. 56-66.

[Pavlov et al., 2006] Pavlov, R., Paneva, D., Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, Cybernetics and Information Technologies, 2006, Vol. 6, Number 3, pp. 51-62.

[Pavlov et al., 2010] Pavlov, R., Paneva-Marinova, D., Rangochev, K., Goynov, M., Luchev, D., Towards Online Accessibility of Valuable Phenomena of the Bulgarian Folklore Heritage, In the Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech'10), June, 2010, Sofia, Bulgaria, ACM ICPS Vol. 471, pp. 329-334.

[Pavlova-Draganova et al., 2006] Pavlova-Draganova L., Georgiev V., Draganov L., Virtual Encyclopaedia of the Bulgarian Iconography, In the Proceedings of Modern eLearning Conference, Varna, Bulgaria, 1 - 5 July, 2006, pp. 165 – 170.

[Project Calimera] Project Calimera - www.lib.bg/proekti/CalimeraSummaryBG.doc

Authors' Information

Radoslav Pavlov – PhD in Mathematics, Associated Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: radko@cc.bas.bg

Major Fields of Scientific Research: Multimedia and Language Technologies, Digital Libraries, Information Society Technologies, e-Learning, Theoretical Computer Science, Computational Linguistics, Algorithmic, Artificial Intelligence and Knowledge Technologies.

Galina Bogdanova – PhD in Informatics, Associated Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: galina@math.bas.bg

Major Fields of Scientific Research: Information Society Technologies, Multimedia Digital Archives, Data Mining, Information Society Technologies, Steganographia, Coding Theory, Computer Science, Algorithms, Knowledge Technologies and Applications.

Desislava Paneva-Marinova – PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: dessi@cc.bas.bg

Major Fields of Scientific Research: Multimedia Digital Libraries, Personalization and Content Adaptivity, eLearning Systems and Standards, Knowledge Technologies and Applications.

Todor Todorov – PhD in Informatics, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: todor@math.bas.bg

Major Fields of Scientific Research: Coding theory, Steganography, Watermarking, Multimedia Digital Archives, Databases, Data Mining.

Konstantin Rangochev – PhD in Philology, Assistant Professor, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev Str., bl. 8, Sofia 1113, Bulgaria; e-mail: krangochev@yahoo.com

Major Fields of Scientific Research: Ethnology, Folklore studies, Culture Anthropology, Linguistics, Computational Linguistics, Digital Libraries.

CORRELATION ANALYSIS OF EDUCATIONAL DATA MINING BY MEANS A POSTPROCESSOR’S TOOL

Georgi Teodorov, Oktay Kir, Irina Zheliazkova

Abstract: *The paper deals with the correlation analysis as educational data technique that is easy to interpret and simple to implement. Two datasets respectively from environment for knowledge testing and for exercise tasks modelling testing are gathered. Programming of tasks for test parameters relationships, test reliability, cheat recognition, and test validation in a specialized postprocessor tool is discussed .*

Keywords: data mining, correlation analysis, dataset, test reliability, test validity, cheap recognition, postprocessor tool

ACM Classification Keywords: Computer and Information Science Education, Knowledge Representation

Introduction

In the recent review of Romero & Ventura [6] correlation analysis has been pointed out as one of the Educational Data Mining (*EDM*) techniques for extracting useful information to support reasonable decisions making in the educational environments. In table 1 all kinds of tasks (from A to K) to which this technique has been applied are listed. It is not surprising that teachers widely use the correlation analysis as it’s easy to interpret similar to the descriptive statistics and simpler for computation in comparison with other known techniques as neural, Bayesian, and Kohenen networks, rule-based systems, cluster and regression analysis. The main requirements for design of the *EDM* tools also are formulated in the same survey and concern: the user interface, visualization task, integration of the tool with an educational environment, standardization of data and models, as well as algorithms for data mining.

An earlier paper of Hernandez et al. [1] deals in depth with the task F and more precisely with cheat in online testing. There a questionnaire study is cited concerning the students’ cheat and conducted by Donald McCabe. In a representative sample of 1,800 students from nine state universities in USA, seventy percent of students admitted to cheate on exams. As a result five reasons for this undesirable student’s behavior were discovered, namely: lazy or didn’t study or prepare, to pass a class or improve a grade, external pressure to succeed, didn’t know answers, time pressure or too much work. In the above-mentioned paper Genderman, who founded four main factors associated with academic dishonesty (individual characteristics, peer group influences, instructor influences, and institutional policies) also is cited. In the everyday teaching practice some students even become masters in the art of cheating.That is why it is interesting to analyze the student’s abnormal behavior and compare it with the normal one.

Table 1. The groups of tasks for EDM using correlation analysis technique

	Objective
A. Analysis and Visualization of Data	to highlight useful information and support decision making. In the educational environment, for example, it can help educators and course administrators to analyze the students' course activities and usage information to get a general view of a student's learning.
B. Providing Feedback for Supporting Instructors	to provide feedback to support course authors/teachers/administrators in decision making (about how to improve students' learning, organize instructional resources more efficiently, etc) and enable them to take appropriate proactive and/or remedial action.
C. Recommendations for Students	to be able to make recommendations directly to the students with respect to their personalized activities, links to visits, the next task or problem to be done, etc, and also to be able to adapt learning contents, interfaces, and sequences to each particular students.
D. Predicting	to estimate the unknown value of a variable that describes the student. In education, the values normally predicted are performance, knowledge, score, or mark. This value can numerical/continuous value (regression task) or categorical/discrete value (classification task).
E. Student Modeling	to develop cognitive models of human users/students, including a modeling of their skills and declarative knowledge. Data mining has been applied to automatically consider user characteristics (motivation, satisfaction, learning styles, affective status, etc.) and learning behavior in order to automate the construction of student models.
F. Detecting Undesirable Student Behaviors	to discover/detect those students who have some type of problem or unusual behavior such as: erroneous actions, low motivation, playing games, misuse, cheating, dropping out, academic failure, etc.
G. Grouping Students	to create groups of students according to their customized features, personal characteristics, etc. Then the clusters/groups of students obtained can be used by the instructor/developer to build a personalized learning system to promote effective group learning, to provide adaptive contents, etc.
H. Social Network Analysis	Social networks analysis, aims at studying relationships between individuals, instead of individual attributes or properties. A social network is considered to be a group of people, an organization or social individuals who are connected by social relationships like friendships, cooperative relations, or informative exchange.
I. Developing Concept Maps	to help instructors/educators in the automatic process of developing/constructing concept maps. A concept map is a conceptual graph that shows relationships between concepts and expresses the hierarchal structure of knowledge.
J. Constructing Courseware	to help instructors/development process of courseware and learning contents automatically. On the other hand, it also tries to promote the reuse/exchange of existing learning resources among different users and systems.
K. Planning and Scheduling	to enhance the traditional educational process by planning future courses, helping with student course scheduling, planning resource allocation, helping in the admission and counseling processes, developing curriculum, etc.

For detecting students cheats in on-line exams Hernandez et al. [1] proposed to use Knowledge Discovery in Databases (*KDDs*) a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in databases. They viewed the *EDM* simply as an essential step in the process of *KDDs* and use *WEKA* as a Data Mining Engine *(DME*). *WEKA* contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. The machine learning algorithms can either be applied directly to a dataset or be called from the researcher own Java code. It is also well-suited for developing new machine learning algorithms. At the same time it’s easy to use and understandable and provides a comprehensive environment for testing methods against other existing methods.

To solve this *DM* task Jelev et al. applied the descriptive statistics and visualization techniques on a dataset with test results. The test was extracted by a database containing 150 questions covering the lecture material for the course “Programming structures” and created by means of the popular during the 90’s multi-media environment ToolBook. The test included 35 multiple-choice questions each one with 5 alternatives only one of which is the correct answer. The goal of the first study [4] was to determinate the test validity using one-factor analysis with Fisher’s criterion. Each of 6 test versions was analyzed according to the correct answers students gave to 7 randomly selected questions (one from each topic taught). The examination of the test difficulty through one-factor analysis shown that the different test versions do not pose a significant influence on the grades, e.g. the test is a good means for the knowledge testing. The intervals of correct answers were determined to correspond to a six-point marking scale as 10 scores to correspond to the mark “poor”. That threshold presents 0.35 from the total test scores is close to that one (0.40) accepted by Zheliazkova’s group. In the second study [5] the scores distribution in percentage of the experimental curve of the student’s marks and the theoretical curve representing a normal Gauss distribution were visualized in a common coordinate system. The conclusion was made that the two curves have approximately the same distributions with mean equal to the “good” mark (4.00). The probability for a student to give a definite number of correct answers under full lack of knowledge or random choice of answers was calculated applying the Bernoulli formula. The obtained graphical result of this probability shown that the case of 7 correct answers has the highest probability. For analysis of the dispersion measures the sigma derivation method was applied to assess if the sample of 35 questions is representative. The obtained result was approximately 33 questions as this number covers with 90% guarantee the lecture material.

Authors’ Team Previous Studies

Since 2006 Zheliazkova’s research group has been using different *DM* techniques for postprocessing students’ tests and exercise results. Two experimental studies had been conducted to assess the effectiveness of the intelligent computer-based tests in comparison with the traditional ways of testing such as multiple-choice tests, and exams [8,9]. Objects of the studies were students-bachelor (1-st year, 2-nd semester), specialties Computer Systems and Technologies (CST) and Communication Technique and Technologies (CTT) at Rousse University. A multiple-choice test (*T1*) and an intelligent test (*T2*) covering the theme “Algorithms” from the subject “Programming 1” were generated by means of a specialized environment for knowledge testing. As a tool for postprocessing the gathered datasets was used Excel. The relationship between the exam mark given previous semester by the lecturer (*M3*) and the intelligent test mark (*M2*) in traditional six-grade scale was found to be high while the relationship between *M1* and *M3*, as well as between *M1* and *M2* was moderate. Another interesting relationship between the time undertaken and *M1* and time and *M2* was calculated as lower. According to *M3* the experimental data was divided in two tables respectively with data for the students with mark “5” or “6” and for the students with mark “3” or “4”. That was made because at the end of the previous semester the students from the first group were assessed by the lecturer and released from the exam. Note that, the values of *r (time, M1*) close to 0 confirm the statement that is not objective due to some well-known reasons in contrast to *r*(*time*,*M2*) had a positive value, greater for the first group of students than for the second one.

In a more recent study of Zheliazkova’s group [10] another interesting relationship that between the test mark (*TM*) and exercise mark (*EM*) in the traditional scale also had been investigated. Two datasets were gathered from two specialized environments respectively for knowledge testing and modeling dynamic systems used in the authors’ team teaching practice. Again Excel was used as a tool for postprocessing. The value of *r* (*TM, EM*) was 0.44 that means a moderate relationship. The conclusion made was that both environments are feasible and yield to sustainable and valid results. Probably the reasons why this value was not higher is that the exercise tasks for modeling were more complex and the students used a new software environment to perform them. It took most of students 2-3 times longer to complete the first task for modeling than to complete the following ones. So, unavoidably, the exercise performed within the above-mentioned environment beside the subject specific knowledge partly measures also technological skills to use this environment.

A teacher’s tool implemented by Zheliazkova’s group for the *EDM* called postprocessor was reported from design, implementation, and user’s points of view. For ensuring the tool’s intelligence and its adaptation to the teacher a power and expressive script language called *SessionScript* was implemented. Programming of descriptive statistics, visualization, and correlation analysis techniques was demonstrated using two output data sets respectively from both above-mentioned environments. Application of the linear methods of prediction using this tool is reported in another paper submission for the present conference [3]. The description of the experimental dataset and technology of the tool using can be found there.

The present paper deals with the same experimental dataset and the same tool that’s why their descriptions are omitted here. The next paper sections focus on the correlation analysis application respectively for the following *DM* tasks: test parameters relationships, test reliability, cheat recognition, and test validation. Conclusion summarizes the methodology proposed for their application using the postprocessor.

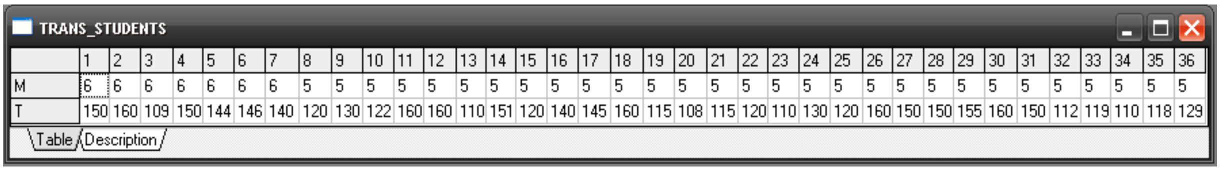
Correlation Analysis for the Test Parameters Relationships

The coefficient of correlation analysis r_{XY} can serve as a qualitative indicator for the relationship between two statistical test’s parameters, for example, X and Y with the number of the questions n and the mean indicators respectively m_x, m_y $r_{XY} = \sum_{i=1}^{n}(x_i - m_x).(y_i - m_y) / \sqrt{\sum_{i=1}^{n}(x_i - m_x)^2 . \sum_{i=1}^{n}(y_i - m_y)^2}$. The value of r_{XY} shows how strong is the relationship between the considered parameters and changes in the range from -1.00 to +1.00. In order to move from its concrete value to more clear for a non-skilled teacher (T) a five-intervals scale with a linguistics value is used. For example, if r_{XY} is in the range 0.0 ÷ 0.3 then the relationship is low; 0.3 ÷ 0.5 – moderate; 0.5 ÷ 0.7 – significant; 0.7 ÷ 0.9 – high; 0.9 ÷ 1.0 – very high. If two parameters are moving in the same way r = +1.0 and if in the opposite r = -1.0. The value 0 means that there is no relationship between the considered parameters.

The test mark in the traditional six-grades scale was computed only on the base of correct knowledge scores (P) as follows: $0 \le P \le 0.4^* P_{max}$ – “2”; $0.4^* P_{max} < P \le 0.55^* P_{max}$ – “3”; $0.55^* P_{max} < P \le 0.70^* P_{max}$ – “4”; $0.70^* P_{max} < P \le 0.85^* P_{max}$ – “5”; $0.85^* P_{max} < P \le 1.0^* P_{max}$ – “6” where P_{max} = 352 was the total test scores. The experience accumulated during the last decade by Zheliazkova’s research group has pointed out that such a non-linear scale gives a Gauss distribution of the student’ marks and is acceptable by both teachers and students.

Correlation between the student’s mark and time undertaken for test performance is one of the very interesting relationships and at the same time not well studied. The time planned for the test performance was T_{max} = 120 min but the students were told that the time for the test performance actually is unlimited and together with wrong and missing knowledge will be used as assessment indicators only for research purpose. The input table 2 contains test mark (M) and time undertaken (T) for the “very good” students, e.g. with mark “6” or “5”on the base of their correct knowledge. The coefficient of correlation calculated was 0.21, e.g. that means low correlation between the considered parameters.

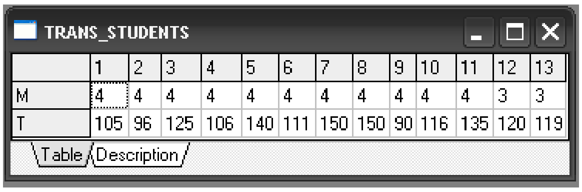
Table 2. The test mark and time undertaken for the “very good” students



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
M	6	6	6	6	6	6	6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
T	150	160	109	150	144	146	140	120	130	122	160	160	110	151	120	140	145	160	115	108	115	120	110	130	120	160	150	150	155	160	150	112	119	110	118	129

The input table 3 contains the values of M and T for the “good” students, e.g. received mark “4” or “3”. For this part of the students the coefficient of correlation was 0.02 that means no correlation between the considered parameters. The total coefficient of correlation computed was equal to 0.36 that means low rather than moderate relationship.

Table 3. The test mark and time undertaken for the “good” students



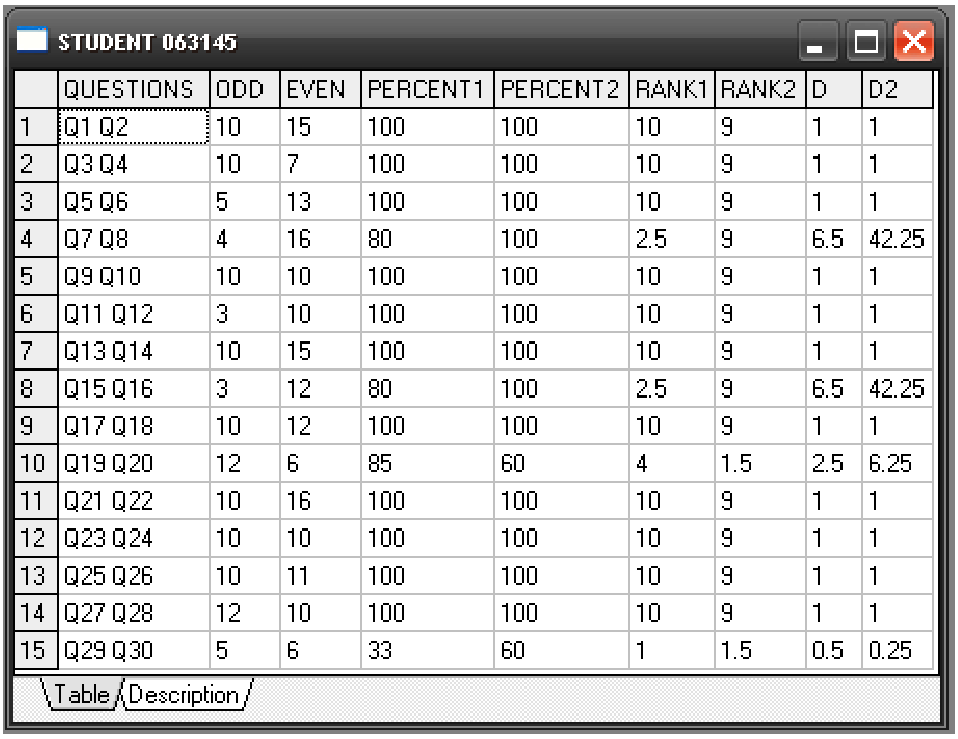
	1	2	3	4	5	6	7	8	9	10	11	12	13
M	4	4	4	4	4	4	4	4	4	4	4	3	3
T	105	96	125	106	140	111	150	150	90	116	135	120	119

Note, that the total number of students decreased from 63 to 49 due to unwilling of some students to register T in their Word documents. Among them was the single student received the mark “poor”. It is seen from the table 3 that only two students received the mark “satisfactory”. These test results also confirm the finding that the students believe in the objective and precise test assessment and go to test only if they assessed themselves at least with mark “satisfactory”.

Correlation Analysis for the Test Reliability

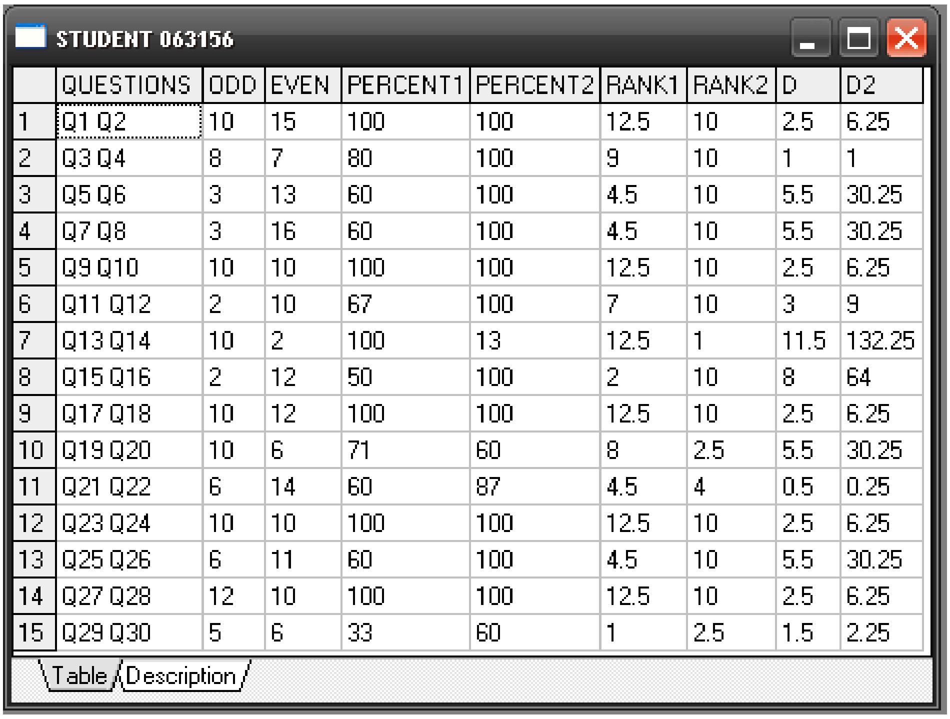
The idea for application of the correlation analysis for the test reliability belongs to Savelev et al., 1986 [7] under the assumption that all other factors are constant, and a longer test will be probably more reliable than a shorter one the indicators of the answers of the even and odd questions and the coefficients of their rank correlation have to be computed.

Table 4. Test results of the ST4



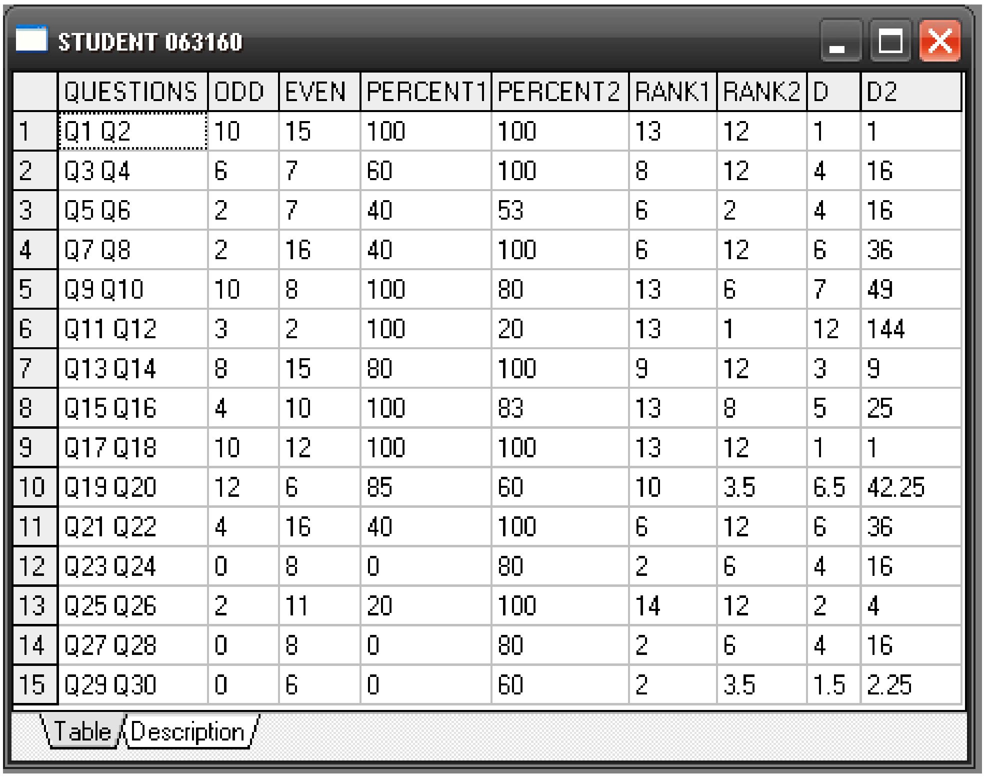
	QUESTIONS	ODD	EVEN	PERCENT1	PERCENT2	RANK1	RANK2	D	D2
1	Q1 Q2	10	15	100	100	10	9	1	1
2	Q3 Q4	10	7	100	100	10	9	1	1
3	Q5 Q6	5	13	100	100	10	9	1	1
4	Q7 Q8	4	16	80	100	2.5	9	6.5	42.25
5	Q9 Q10	10	10	100	100	10	9	1	1
6	Q11 Q12	3	10	100	100	10	9	1	1
7	Q13 Q14	10	15	100	100	10	9	1	1
8	Q15 Q16	3	12	80	100	2.5	9	6.5	42.25
9	Q17 Q18	10	12	100	100	10	9	1	1
10	Q19 Q20	12	6	85	60	4	1.5	2.5	6.25
11	Q21 Q22	10	16	100	100	10	9	1	1
12	Q23 Q24	10	10	100	100	10	9	1	1
13	Q25 Q26	10	11	100	100	10	9	1	1
14	Q27 Q28	12	10	100	100	10	9	1	1
15	Q29 Q30	5	6	33	60	1	1.5	0.5	0.25

Table 5. Test results of the ST1



	QUESTIONS	ODD	EVEN	PERCENT1	PERCENT2	RANK1	RANK2	D	D2
1	Q1 Q2	10	15	100	100	12.5	10	2.5	6.25
2	Q3 Q4	8	7	80	100	9	10	1	1
3	Q5 Q6	3	13	60	100	4.5	10	5.5	30.25
4	Q7 Q8	3	16	60	100	4.5	10	5.5	30.25
5	Q9 Q10	10	10	100	100	12.5	10	2.5	6.25
6	Q11 Q12	2	10	67	100	7	10	3	9
7	Q13 Q14	10	2	100	13	12.5	1	11.5	132.25
8	Q15 Q16	2	12	50	100	2	10	8	64
9	Q17 Q18	10	12	100	100	12.5	10	2.5	6.25
10	Q19 Q20	10	6	71	60	8	2.5	5.5	30.25
11	Q21 Q22	6	14	60	87	4.5	4	0.5	0.25
12	Q23 Q24	10	10	100	100	12.5	10	2.5	6.25
13	Q25 Q26	6	11	60	100	4.5	10	5.5	30.25
14	Q27 Q28	12	10	100	100	12.5	10	2.5	6.25
15	Q29 Q30	5	6	33	60	1	2.5	1.5	2.25

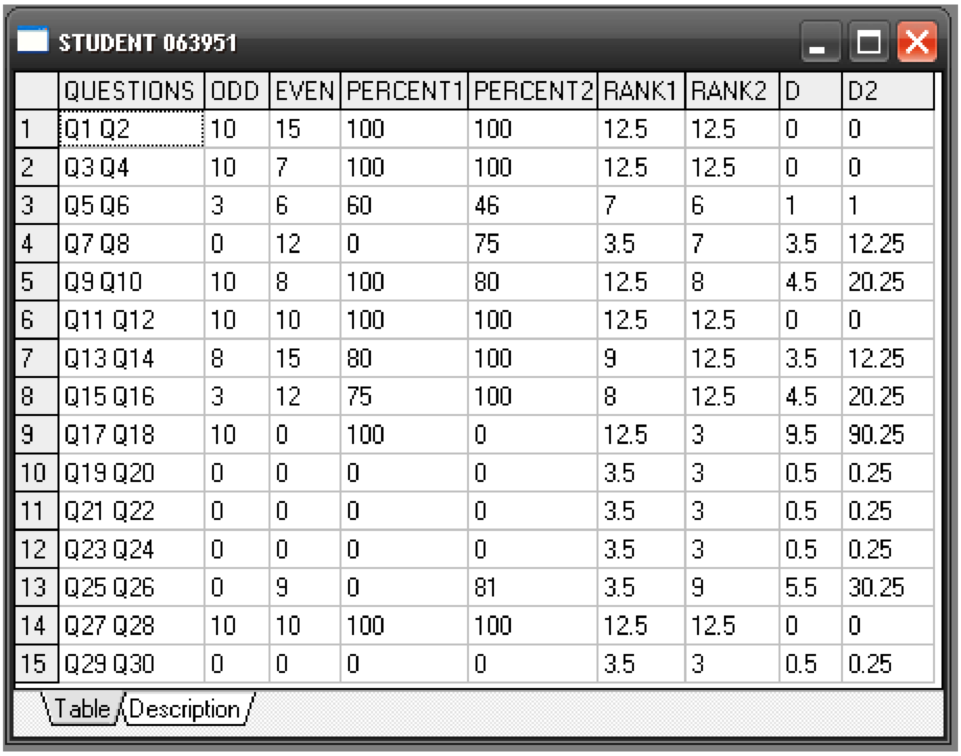
Table 6. Test results of the ST2



STUDENT 063160

	QUESTIONS	ODD	EVEN	PERCENT1	PERCENT2	RANK1	RANK2	D	D2
1	Q1 Q2	10	15	100	100	13	12	1	1
2	Q3 Q4	6	7	60	100	8	12	4	16
3	Q5 Q6	2	7	40	53	6	2	4	16
4	Q7 Q8	2	16	40	100	6	12	6	36
5	Q9 Q10	10	8	100	80	13	6	7	49
6	Q11 Q12	3	2	100	20	13	1	12	144
7	Q13 Q14	8	15	80	100	9	12	3	9
8	Q15 Q16	4	10	100	83	13	8	5	25
9	Q17 Q18	10	12	100	100	13	12	1	1
10	Q19 Q20	12	6	85	60	10	3.5	6.5	42.25
11	Q21 Q22	4	16	40	100	6	12	6	36
12	Q23 Q24	0	8	0	80	2	6	4	16
13	Q25 Q26	2	11	20	100	14	12	2	4
14	Q27 Q28	0	8	0	80	2	6	4	16
15	Q29 Q30	0	6	0	60	2	3.5	1.5	2.25

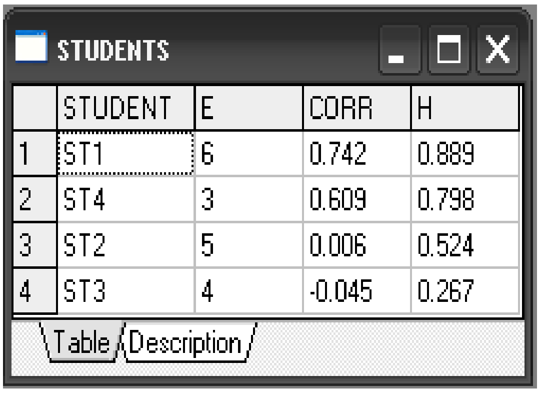
Table 7. Test results of the ST3



STUDENT 063951

	QUESTIONS	ODD	EVEN	PERCENT1	PERCENT2	RANK1	RANK2	D	D2
1	Q1 Q2	10	15	100	100	12.5	12.5	0	0
2	Q3 Q4	10	7	100	100	12.5	12.5	0	0
3	Q5 Q6	3	6	60	46	7	6	1	1
4	Q7 Q8	0	12	0	75	3.5	7	3.5	12.25
5	Q9 Q10	10	8	100	80	12.5	8	4.5	20.25
6	Q11 Q12	10	10	100	100	12.5	12.5	0	0
7	Q13 Q14	8	15	80	100	9	12.5	3.5	12.25
8	Q15 Q16	3	12	75	100	8	12.5	4.5	20.25
9	Q17 Q18	10	0	100	0	12.5	3	9.5	90.25
10	Q19 Q20	0	0	0	0	3.5	3	0.5	0.25
11	Q21 Q22	0	0	0	0	3.5	3	0.5	0.25
12	Q23 Q24	0	0	0	0	3.5	3	0.5	0.25
13	Q25 Q26	0	9	0	81	3.5	9	5.5	30.25
14	Q27 Q28	10	10	100	100	12.5	12.5	0	0
15	Q29 Q30	0	0	0	0	3.5	3	0.5	0.25

Table 8. The generated table



	STUDENT	E	CORR	H
1	ST1	6	0.742	0.889
2	ST4	3	0.609	0.798
3	ST2	5	0.006	0.524
4	ST3	4	-0.045	0.267

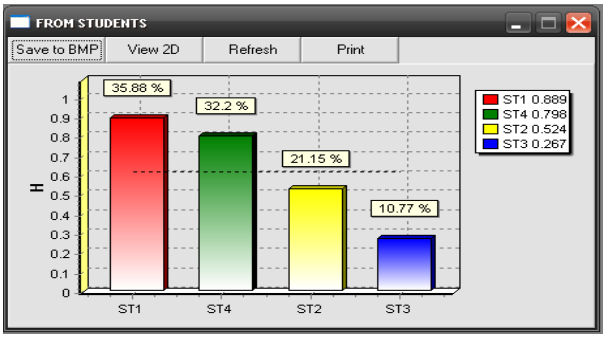


Fig. 1. The bar diagram of the Spearman-Brown coefficient

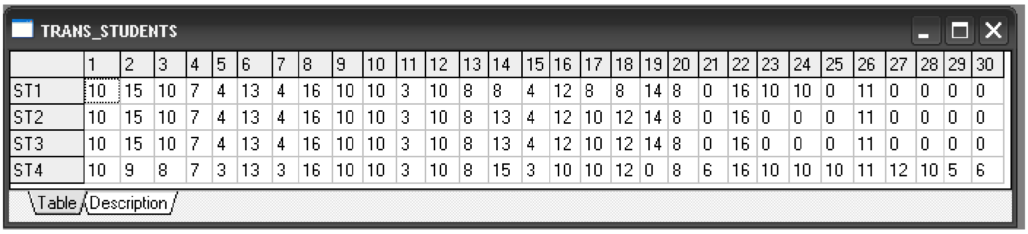
The reliability of test is measured applying the formula of Spearman-Brown: $H = 2.r_{XX} / (1 + r_{XX})$, where $r_{XX} = 1 - 6.\sum_{i=1}^{n} (x_i^{'} - x_i^{''})^2 / (n^3 - n)$. It is accepted that the test reliability is enough if $H > 0.8$. For a student's test performance the r_{XX} between the odd and even test questions (CORR) and the Spearman-Brown coefficient (H) one are given in table 8. For the first and fourth student the test is reliable enough. For the second and third students the test turned to be non-reliable. This can be explained with the fact that the test was oriented to the “4” students supposed to be a substantial part of all students. Obviously, for the “5” and “6” students, as well as the “3” students the test is more likely unreliable. The average of the H (approximately 0.6) depicted with a dotted line on fig. 1 shows that the test can be accepted as reliable.

Correlation Analysis for the Cheat Recognition

For cheat discovering Hernandez et al. [1] used a complex patterns recognition approach using test correct and inccorrect answers as dataset. The approach based on correlation analysis has three stages of proving the cheat - datasets with correct (table 9) , missing (table 10), and wrong knowledge (table 11). The observation during the test performance showed at least four students possibly attempted to cheat.

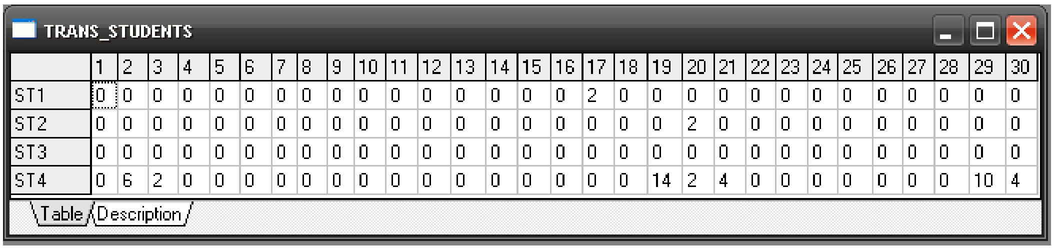
A visual comparison of the correlation between ST1, ST2, ST3 and ST4 is shown on tables 12, 13, and 14. The correlation of the ST4 with ST1, ST2, and ST3 is close to moderate that is a normal student. The highest correlation of 1.00 is between ST2 and ST3 which means one of them could be the test answers source.

Table 9. The test questions correct knowledge of four students possibly attempted to cheat



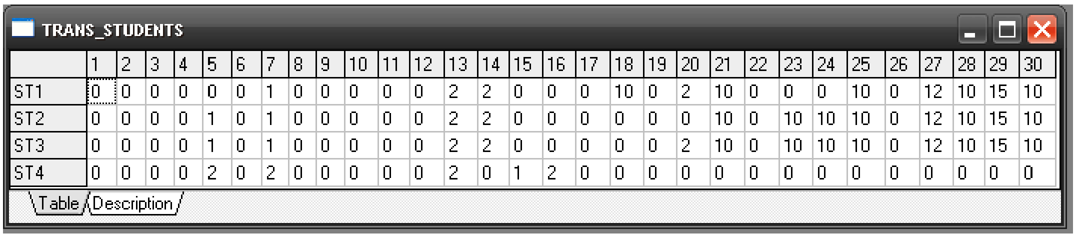
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
ST1	10	15	10	7	4	13	4	16	10	10	3	10	8	8	4	12	8	8	14	8	0	16	10	10	0	11	0	0	0	0
ST2	10	15	10	7	4	13	4	16	10	10	3	10	8	13	4	12	10	12	14	8	0	16	0	0	0	11	0	0	0	0
ST3	10	15	10	7	4	13	4	16	10	10	3	10	8	13	4	12	10	12	14	8	0	16	0	0	0	11	0	0	0	0
ST4	10	9	8	7	3	13	3	16	10	10	3	10	8	15	3	10	10	12	0	8	6	16	10	10	10	11	12	10	5	6

Table 10. The test questions missing knowledge of four students possibly attempted to cheat



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
ST1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
ST2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
ST3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ST4	0	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	2	4	0	0	0	0	0	0	0	10	4

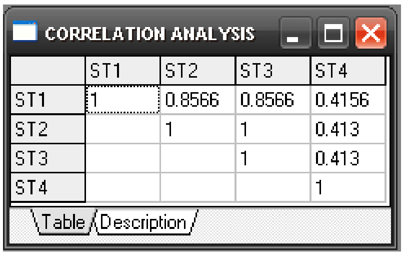
Table 11. The test questions wrong knowledge of four students possibly attempted to cheat



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
ST1	0	0	0	0	0	0	1	0	0	0	0	0	2	2	0	0	0	10	0	2	10	0	0	0	10	0	12	10	15	10
ST2	0	0	0	0	1	0	1	0	0	0	0	0	2	2	0	0	0	0	0	0	10	0	10	10	10	0	12	10	15	10
ST3	0	0	0	0	1	0	1	0	0	0	0	0	2	2	0	0	0	0	0	2	10	0	10	10	10	0	12	10	15	10
ST4	0	0	0	0	2	0	2	0	0	0	0	0	2	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0

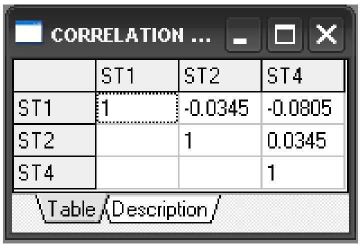
The ST3 has zero missing knowledge for all questions answers (table 10) and due to the error “division by zero” when calculating the correlation so he/she is omitted in table 13. To confirm this finding the lecturer had a talk with these students and assessed them with the test mark “3” for their honesty.

Table 12. Correct knowledge case



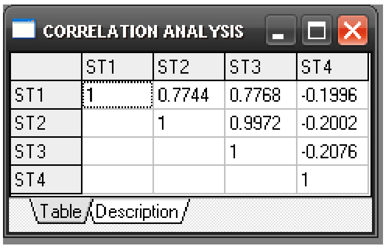
	ST1	ST2	ST3	ST4
ST1	1	0.8566	0.8566	0.4156
ST2		1	1	0.413
ST3			1	0.413
ST4				1

Table 13. Missing knowledge case



	ST1	ST2	ST4
ST1	1	-0.0345	-0.0805
ST2		1	0.0345
ST4			1

Table 14. Wrong knowledge case



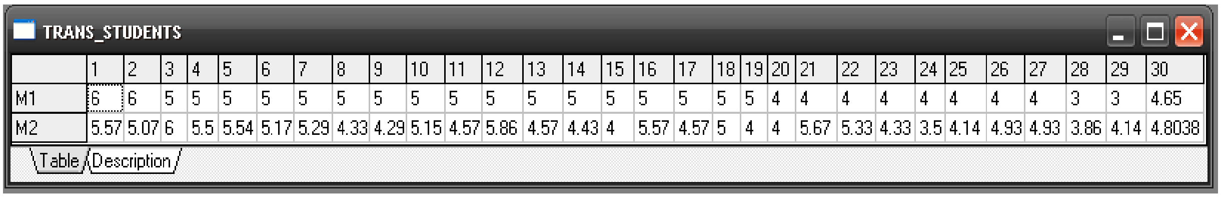
	ST1	ST2	ST3	ST4
ST1	1	0.7744	0.7768	-0.1996
ST2		1	0.9972	-0.2002
ST3			1	-0.2076
ST4				1

Correlation Analysis for the Test Validation

For the test validation correlation between the test mark (*M1*) and exercise mark (*M2*) given by the assistant at the end of the semester was computed. He/she had been told to assess each student's activity during each exercise (their number was 15) in the traditional six-grade scale. That information was brought in an Excel table and *M2* of each student was computed as an average mark with accuracy 0.25. It is assumed that the difference between both marks hasn't to exceed one interval for the students with normal behavior. The difference between both marks would exceed one interval for the students with low *M2* probably used the common device for cheat.

The results for both groups of students are shown in the input tables 15 and 16 respectively in which the last column (30 and 15) contains the corresponding average mark. Note, that the total number of the students decreased from 63 to 43 as for some students *M2* missing. A positive tendency also is confirmed that for the intelligent tests the average mark is shifted from "good" to "very good". The mean of the test mark close to "excellent" (5.35) against that for the exercise (3.71) confirms that these students have abnormal behavior.

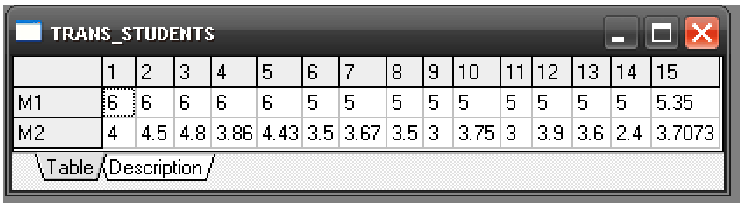
Table 15. The test and exercise mark for the students with normal behavior



TRANS_STUDENTS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
M1	6	6	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	4	4	4	4	4	4	4	3	3	4.65
M2	5.57	5.07	6	5.5	5.54	5.17	5.29	4.33	4.29	5.15	4.57	5.86	4.57	4.43	4	5.57	4.57	5	4	4	5.67	5.33	4.33	3.5	4.14	4.93	4.93	3.86	4.14	4.8038

Table 16. The test and exercise mark for the students with abnormal behavior



TRANS_STUDENTS

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
M1	6	6	6	6	6	5	5	5	5	5	5	5	5	5	5.35
M2	4	4.5	4.8	3.86	4.43	3.5	3.67	3.5	3	3.75	3	3.9	3.6	2.4	3.7073

There is a small difference (0.15) between *M1* and *M2* for the first group of students while as this difference for the second group is substantial (1.65). The coefficient of correlation for the first group *r (M1,M2)* = 0.43 is close to those in our previous study [11] and for the second group is 0.74 is much greater.

Conclusions

The correlation analysis for the test parameters relationships, test reliability, cheat recognition, and test validation has been applied using a specialized tool based on two data sets respectively for knowledge testing and exercise performing. Programming these tasks is simpler and easy to interpret by the educators. Some findings in this study are in line with some previous studies.

The following methodology for using the tool for such group of tasks is proposed: 1) Constructing the input table with columns equal to the students with normal behaviour and rows to their test and exercise marks; 2) Adding new column with the average values of both mark; 3) Calculating the correlation between these marks; 4) Repeating 1, 2, and 3 for the group of students with abnormal behaviour (if it exists);

The authors are grateful to all students participated in this risky study with help of which it becomes possible.

Bibliography

[1] Hernandez J. A., Ochoa A.,, Munoz J., Burlak G., Detecting Cheats in online students assessments using Data Mining, Conference on Data Mining, 2006, pp 204-210.

[2] Kir O., Zheliazkova I. I., Teodorov G., Educational Data Mining by Means of a Power Instructor’s Tool, Proceedings of International Conference on Entrepreneurship, Innovation, and Regional Development (ICEIRD), 2011 (Accepted).

[3] Kir O., Zheliazkova I. I., Prediction of Educational Data Mining by Means of a Postprocessor Tool, International Conference “Modern E-Learning”, Varna, 2011 (Accepted).

[4] Jelev G., Minkovska D., Approaches for Definition the Validity of the Results of the Test for Knowledge Mastering, International Conference on Computer Science, 2004, Sofia, pp. 268-273.

[5] Jelev G., Minkova Y., Determination of Representative Sample Size and Knowledge Assimilation Test Results Processing. Problems and Discussion, International Conference on Computer Science, 2004, Sofia, pp. 274-279.

[6] Romero Cr., Ventura S. Educational Data Mining: A Review of the State of the Art, IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews 2010; 40/6: 601-18.

[7] Savelev A. Ya., Novikov V.A., Lobanov Yu. I., Preparing Information for Automated Teaching Systems, Moscow, 1986.

[8] Zheliazkova I. I., Andreeva M. H., Kolev R. T., Knowledge Testing in Algorithms – An Experimental Study, International Conference “Modern E-Learning”, Varna, 2006, pp.55-62.

[9] Zheliazkova I. I., Kolev R. T., Andreeva M. H., Knowledge Testing in Algorithms by Means of a Word Technology, Proceedings of the Second National Conference on E-Leaning in the Higher Education, Kiten, Bulgaria, 2006, pp. 21-25.

[10] Zheliazkova I. I., Valkova P. A., Georgiev G. Todorov, A Computer-Based Technology for Processing and Visualization of Session’s Data, Int. Journal of Information Technologies and Control, 2011 (Accepted).

Authors' Information

Georgi Teodorov – PhD student, University of Rousse, Studentska street 8, Rousse 7017, Bulgaria; e-mail: georgi.t.georgiev@gmail.com

Oktay Kir – PhD student, University of Rousse, Studentska street 8, Rousse 7017, Bulgaria; e-mail: kir.oktay@gmail.com

Irina Zheliazkova – Associate Professor; University of Rousse, Studentska street 8, Rousse 7017, Bulgaria; e-mail: irina@ecs.ru.acad.bg

TABLE OF CONTENTS

Cross Intersection Sequel Of Discrete Isoperimetry

Levon Aslanyan, Vilik Karakhanyan 203

Adaptive Neuro-Fuzzy Kohonen Network With Variable Fuzzifier

Yevgeniy Bodyanskiy, Bogdan Kolchygin, Iryna Pliss 215

Correlation Maximization in Regression Models Based on Convex Combinations

Oleg Senko, Alexander Dokukin 224

Neural Network Segmentation of Video Via Time Series Analysis

Dmitry Kinoshenko, Sergey Mashtalir, Andreas Stephan, Vladimir Vinarski 232

Astronomical Plates Spectra Extraction Objectives and Possible Solutions Implemented on Digitized First Byurakan Survey (DFBS) Images

Aram Knyazyan, Areg Mickaelian, Hrachya Astsatryan 243

Membranes Distribution Using Genetic Algorithms

Miguel Ángel Peña, Juan Castellanos 248

Virtual Membrane Systems

Alberto Arteta, Angel Castellanos, Nuria Gómez 258

Higaia Methodology

A. Anguera, A. Gutierrez, M.A. Diaz 267

Digital Archive and Multimedia Library for Bulgarian Traditional Culture and Folklore

Radoslav Pavlov, Galina Bogdanova, Desislava Paneva-Marinova, Todor Todorov, Konstantin Rangochev 276

Correlation Analysis of Educational Data Mining by Means a Postprocessor’s Tool

Georgi Teodorov, Oktay Kir, Irina Zheliazkova 289

Table of Contents 300