ITHEA

# International Journal
# INFORMATION THEORIES & APPLICATIONS

2003 Volume 10 Number 1

International Journal
INFORMATION THEORIES & APPLICATIONS
ISSN 1310-0513
Volume 10 / 2003, Number 1



IJ ITA is official publisher of the scientific papers of the members of the Association of Developers and Users of Intellectualized Systems (ADUIS).

IJ ITA welcomes scientific papers connected with any information theory or its application. Original and non-standard ideas will be published with preferences.

Papers must be written in English.

Responsibility for papers published in IJ ITA belongs to authors.
Please get permission to reprint any copyrighted material before you send it to IJ ITA.

*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 .

The camera-ready copy of the paper should be received by e-mail: foi@nlcv.net

International Journal "INFORMATION THEORIES & APPLICATIONS"
Vol.10, Number 1, 2003
Printed in Bulgaria
Edited by the Institute of Information Theories and Applications FOI ITHEA, Bulgaria
Publisher: FOI-COMMERCE - Sofia, 1000, P.O.B. 775, Bulgaria
www.foibg.com e-mail: foi@nlcv.net

 ISSN 1310-0513

*About ...*

# I J I T A

*Ten successful years !*

*Verba volant, scripta manent !*

*The progress in Information and Computer Sciences is fuelled by the results of research work and the accumulation of practical experience. The concept "Information Theories & Applications" (ITA) represents the synthesis of this knowledge. The field of ITA is progressing rapidly and is constantly creating new challenges for professionals involved in it.*
*It is clear that there is continuing need for international forums for exchange of knowledge, experience and creative inspiration among ITA professionals in order to share and stimulate solutions.*

*The "International Journal on Information Theories and Applications" (IJ ITA) has been established as an independent scientific media in 1993.*
*IJ ITA is a successor of the scientific co-operation organised within 1986-1992 by international workgroups (IWG) researching the problems of data bases and artificial intelligence. As a result of tight relation between these problems in 1990 in Budapest appeared the scientific group of Data Base Intellectualisation (IWGDBI) integrating the possibilities of databases with the creative process support tools.*
*Heads of the IWGDBI are R.Kirkova (Bulgaria) and V.Gladun (Ukraine).*

*Founder and Editor in chief of IJ ITA is Krassimir Markov.*
*During the years main co-editors of IJ ITA have been V.Gladun, R.Kirkova, P.Barnev, and Kr.Ivanova. The IJ ITA International Editorial Staff also includes the members of the IJ ITA International Conferences Program Committees.*
*IJ ITA Publisher is FOI-COMMERCE Co., Sofia, Bulgaria.*

*For ten years IJ ITA became as well-known international journal. Till now, including this volume, more than 300 papers from more than 500 authors have been published. IJ ITA authors are widespread in 34 countries all over the world:* Armenia, Belarus, Belgium, Bulgaria, Canada, Czech Republic, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Latvia, Lithuania, Mexico, Moldavia, Netherlands, Poland, Portugal, Romania, Russia, Scotland, Senegal, Spain, Sultanate of Oman, Turkey, UK, Ukraine, USA.

## International Prize "ITHEA"


*International Prize "ITHEA" has been established in 1995.*
*It is aimed to mark achievements in the field of*
*the information theories and applications.*

**Awarded Scientists till 2003:**

| | | |
|---|---|---|
| 1995 | *Sandansky* | *K. Bankov, P. Barnev, G. Gargov, V. Gladun, R. Kirkova, S. Lazarov, S. Pironkov, V. Tomov* |
| 1996 | *Sofia* | *T. Hinova, K. Ivanova, I. Mitov, D. Shishkov, N. Vashchenko* |
| 1997 | *Yalta* | *Z. Rabinovich, V. Sgurev, A. Timofeev, A. Voloshin* |
| 1998 | *Sofia* | *V. Jotsov* |
| 1999 | *Sofia* | *L. Zainutdinova* |
| 2000 | *Varna* | *I. Arefiev, A. Palagin* |
| 2001 | *St.Peterburg* | *N. Ivanova, V. Koval* |
| 2002 | *Primorsko* | *A. Milani, M. Mintchev* |
| 2003 | *Varna* | *T. Gavrilova, A. Eskenazi, V. Lozovskiy, P. Stanchev* |

## IJ ITA major topics of interest include, but are not limited to:

### INFORMATION THEORIES

*Artificial intelligence*
- *Computer Intellectualisation*
- *Intelligent Networks and Agents*
- *Intelligent Technologies*
- *Knowledge discovery and Engineering*
- *Knowledge acquisition and formation*
- *Distributed artificial intelligence*
- *Models of plausible reasoning*
- *AI Planning and Scheduling*

*Business Informatics*
*Cognitive science*
*Decision making*
*Education informatics*
*General Information Theory*
*Hyper technologies*
*Information models*
*Intellectualisation of data processing*
*Knowledge-based society*
*Logical Inference*
*Natural language processing*
*Neuroinformatics*
*Philosophy and Methodology of Informatics*
*Software engineering and Quality of the programs*
*Theory of Computation*

### APPLICATIONS

*Business information systems*
*Communication systems*
*Intelligent information systems*
*Multimedia systems*
*Program systems with artificial intelligence*
*Pyramidal information systems*
*Special information systems*
*System for computer art and music*
*Very large information systems*

# THE INFORMATION

*K. Markov, K. Ivanova, I. Mitov, E. Velikova-Bandova*

*Abstract: The current formal as well as not formal definitions of the concept "Information" are presented in the paper.*

*Keywords: Information, General Information Theory, and Philosophy of Informatics*

## 1. Introduction

The fundamental notion of the General Information Theory is the concept "Information". All other concepts are defined based on this definition. The first not formal definition of the concept of Information was published in [Markov, 1988]. The main philosophical explanations were published in [Markov et al, 1993]. The first variant of the formal definitions were introduced in [Markov et al, 2003]. This work refines the philosophical basis and represents the more precise variant of the formal definition of the concept “information”.

### 1.1. Entity

In our examination, we consider *the real world* as a space of *entities.* The entities are built by other entities, connected with *relationships*. The entities and relationships between them form the internal *structure* of the entity they build. To create the entity of a certain structural level of the world, it is necessary to have:

- the entities of the lower structural level;
- establishing of the forming relationship.

*The entity* can dialectically be considered as a relationship between its entities of all internal structural levels.

*The forming relationship* has a representative significance for the entity. The destruction of this essential relationship causes its disintegration. The establishment of forming relationship between already existing entities has a determine significance for the emerging of the new entity.

The forming relationship is the reason for *the emergence* of individual properties, which distinguish the new entity from the forming ones.

*The relationships form and present the entity.*

### 1.2. Impact, Interaction, Reflection

Building the relationship between the entities is a result of the *contact* among them. During the contact, one entity *impacts* on the other entity and vice versa. In some cases the opposite impact may not exist, but, in general, the contact may be considered as two mutually opposite impacts which occur in the same time.

The set of contacts between entities forms their *interaction*. The interaction is a specific *interactive relationship* between entities which take part in it.

*The contacts of the given structural level are processes of interaction of the entities on the lower levels.*

During the establishing of the contact, the impact of an entity changes temporally or permanently the internal structure of the impacted entity. In other words, the realisation of the relationships between entities changes, temporary or permanently, their internal structure at one or at few levels.

The internal change in the entity, which is due to impact of the other entity we denote with the notion "*direct reflection*".

Every entity has its own level of sensibility. This means that the internal changes occur when the external influence is over the boundary of the sensibility of the entity.

The *"reflection impulse"* for given entity is the amount of the external influence needed for transition from one state to the reflection one.

The entities of the world interact continuously. It is possible, after one interaction may be realised another. In this case, the changes received by any entity, during the first interaction, may be reflected by the new entity.
This means the *secondary (transitive external) reflection* exists.
The chain of the transitive reflections is not limited. In general, the concept "transitive impact" (respectively "transitive reflection") of the first entity on the third entity through the second one will denote every chain of impacts (reflections) which start from first entity and ends in the third entity, and include the second entity in any internal place of the chain.
One special case is the *external transitive self-reflection* where the entity reflects its own relationships as a secondary reflection during any external interaction.
Some entities have an opportunity of *internal self-reflection*. The internal self-reflection is possible only for very high levels of organisation of the entities, i.e. for entities with very large and complicated structure. The self-reflection (self-change) of the entity leads to the creating of new relationships and corresponding entities in it. Of course, the internal self-reflection is a result of the interaction provided between entities in the low levels of the structure of the entity. Such kind of entities has relatively free sub-entities with own behaviour in the frame of self-preservation of the whole entity. As a result of the self-reflection some relationships and corresponding sub-entities are created or changed in the entity.
The combination of the internal and external self-reflection is possible.

### 1.3. Information

The reflection could not be detected by the entity that contains it. This is dialectical behaviour of the reflection - it is only an internal change caused by the interaction.
During as well as after the interaction between two entities, they may interact with other entities from the environment. If any third entity contains reflections of given entity received by two different ways:

1. by transitive impact of the first entity on the third one through the second entity,
2. by impact of the first entity on the third one which is different from the transitive one, i.e. it can be direct impact or transitive impact through another entity

then the third entity became as an external relationship between entities and their reflections - it became as *"reflection evidence"*.
We may say that the *reflection* of the first entity in the second one is "*information"* for the first entity if there is corresponded reflection evidence.
The generalisation of this idea leads to assertion that *every reflection can be considered as information, if there is corresponding reflection evidence.*

## 2. Formal Definitions

### 2.1. Entity

*Definition 1.* The entity $A$ is the couple $A=(E_A,R_A)$ where:
$E_A$ is a collection of sub-sets of a set $M_A$;
$R_A=\{r_i|i\in I,\ I \text{ is a set}\}$ is a nonempty set of relations in $E_A$, i.e.
$r_i\subset E_A\times E_A=\{(X,Y)|X,Y\in E_A\}$ is a relation and $\check{r}_i=r_i\cup\{(X,Y)|(Y,X)\in r_i)\},\ \forall i\in I$;
and:

1. $\varnothing\in E_A$;
2. $M_A=\cup X,\ X\in E_A$;
3. $\forall r\in R_A$ and $\forall X,Y\in E_A \Rightarrow ((\exists(X,Y)\in\check{r})$ or
$(\exists Z_1,..,Z_p\in E_A,\ Z_k\neq\varnothing,k=1,..,p : (X,Z_1)\in\check{r},(Z_1,Z_2)\in\check{r},\ldots,(Z_p,Y)\in\check{r})$ ■

The condition 3 means that $E_A$ is internally connected in respect to every relation $r\in R_A$.
$M_A$ is called forming set of $A$. The intersection $\rho_A$ of all relations in $R_A$ is called forming relation of $A$, i.e. $\rho_A=\rho(A)=\cap r_i,\ i\in I,\ r_i\in R_A$.

When relationship $(X,Y)$ belongs to the relation $r$ we will denote by $(X\to Y)_r$.
Every element $X$ can be considered as an entity $(E_X,R_X)$, for which $M_X=\{X\}$, $E_X=\{\varnothing,M_X\}$ and $R_X=\{\rho\}$, where $\rho=\{(\varnothing\to M_X)\}$. Because of this, the elements of every forming set $M$ can be assumed as entities.
*Definition 2.* Let $A=(E_A,R_A)$ is an entity. $B$ is a sub-entity of $A$ if $B=(E_B,R_B)$, where $M_B\subset M_A$, $\forall X\in E_B\ \exists Y\in E_A:\ X\subset Y$ and $\forall r_B\in R_B\ \exists r_A\in R_A: r_B\subset r_A$ ■

## 2.2. Impact and Reflection

*Definition 3.* Let $A=(E_A,R_A)$ and $B=(E_B,R_B)$, $A\neq B$ and $f\in R_A$.
A direct impact $\psi$ of $A$ on $B$ concerning the relation $f$ is a nonempty subset of $E_A\times E_B$ such that
$\psi=\psi_f=(A\to B)_\psi\subset E_A\times E_B=\{(X,Y)|X\in E_A,Y\in E_B\}$, $\psi_f\neq\varnothing$
for which:
if $X_i,X_j\in E_A$ and $Y_k,Y_l\in E_B$, such that $(X_i\to X_j)_f,(X_i,Y_k)\in\psi_f,\ (X_j,Y_l)\in\psi_f$
then exists $g\in R_B:(Y_k\to Y_l)_g$ ■

If $(X,Y)\in\psi_f$ we will denote $(X\to Y)_\psi$.
Let remember that $\varnothing\in E_A$ as well as $\varnothing\in E_B$. This means that for every $f\in R_A$ and $g\in R_B$ there exists zero direct impact $o_f=(A\to B)_o=\{(\varnothing,\varnothing)\}$.

*Definition 4.* Let $A=(E_A,R_A)$ and $B=(E_B,R_B)$ and $\psi_f=(A\to B)_\psi$ is an direct impact of $A$ on $B$.
A reflection of $A$ into $B$ realised by the direct impact $\psi$ is the couple
$F_\psi=F=(E_F,R_F)$ for which $E_F=\{Y\in E_B|\exists X\in E_A:(X,Y)\in(A\to B)_\psi\}$ and $R_F=\{r_\psi\}$,
$r_\psi=r_{\psi,f}=\{(Y_1,Y_2)|\exists X_1,X_2\in E_A:(X_1\to X_2)_f;\ (X_1\to Y_1)_\psi;\ (X_2\to Y_2)_\psi\}$ ■

$r_{\psi,f}$ is a reflection of the relationship $f$ by direct impact $\psi$.
The reflection $F_\psi$ will be assumed as a sub-entity of $B$.

## 2.3. Transitive Impact and Transitive Reflection

*Definition 5.* Let:
$A=(E_A,R_A)$, $B=(E_B,R_B)$, $C=(E_C,R_C)$;
$f\in R_A$, $g\in R_B$;
$\varphi_f=(A\to B)_\varphi$ is a direct impact of $A$ on $B$ concerning the relation $f$;
$\psi_g=(B\to C)_\psi$ is a direct impact of $B$ on $C$ concerning the relation $g$.
We will say that the direct impacts $\varphi_f$ and $\psi_g$ can be composed if the reflection of the relationship $f$ by direct impact $\varphi$ is a subset of $g$, i.e. $r_{\varphi,f}\subset g$.

*Definition 6.* If $\varphi_f$ and $\psi_g$ can be composed than the couple $\{\varphi_f,\psi_g\}$ is a transitive impact
$$\xi=\xi_f=\psi_g\circ\varphi_f=(A\to B\to C)_\xi$$ of $A$ on $C$ through $B$ ■

It is clear, that every transitive impact is a chain of at least two composed direct impacts. In general, this chain can contain more than two composed direct and/or transitive impacts. Such chain is transitive impact, too.

*Definition 7.* Let $A=(E_A,R_A)$, $B=(E_B,R_B)$, $C=(E_C,R_C)$.
Let $\xi=\{\varphi_f,\psi_g\}=\xi_{f,g}=(A\to B\to C)_\xi$ is transitive impact of $A$ on $C$ through $B$.

The transitive reflection of $A$ into $C$ through $B$ realised by the impact $\xi$ is the couple $G_\xi=G=(E_G,R_G)$ for which

$E_G=\{Z\in E_C \mid \exists X\in E_A$ and $\exists Y\in E_B:(X,Y)\in(A\to B)_\varphi$ and $(Y,Z)\in(B\to C)_\psi\}$ and

$R_G=\{r_\xi\}$,

$r_\xi=r_{\xi,f}=\{(Z_1,Z_2)\mid \exists X_1,X_2\in E_A:$

$(X_1\to X_2)_f;\ (X_1\to Y_1)_\varphi;\ (X_2\to Y_2)_\varphi;\ (Y_1\to Z_1)_\psi;\ (Y_2\to Z_2)_\psi\}$ ■

If $A$ and $C$ are equal than:

- the transitive impact $\xi$ will be called self-impact;
- the transitive reflection $G_\xi=G=(E_G,R_G)$ for transitive self-impact $\xi(A\to B\to A)$ will be called self-reflection.

## 2.4. Interaction and Interactive Reflections

Let denote by $\Omega_{AB}$ the set of all direct or transitive impacts of $A$ on $B$.

*Definition 8.* Let $A=(E_A,R_A)$ and $B=(E_B,R_B)$. An interaction between $A$ and $B$ is a set

$\Delta_{AB}=\{\Delta_i \mid \Delta_i\in\Omega_{AB}\cup\Omega_{BA},\ i=1,..,h\}$ ■

*Definition 9.* Let:

$A=(E_A,R_A)$ and $B=(E_B,R_B)$;

$\Delta=\Delta_{AB}=\{\Delta_i \mid i=1,..,h\}$ is an interaction between $A$ and $B$;

$F_{\Delta i}=F_i=(E_{Fi},R_{Fi})$ is reflection realised by the impact $\Delta_i$;

An interactive reflection $V_{AB}$ between $A$ and $B$ realised by the interaction $\Delta_{AB}$ is the set of all reflections $F_{\Delta i}$; i.e. $V_{AB}=\{F_{\Delta i}\mid i=1,..,h\}$ ■

The self-interaction $\Delta_{AA}$ is the interaction from $A$ to $A$ where all impacts are self-impacts. In such case the corresponding interactive reflection is called interactive self-reflection.

## 2.5. Information

*Definition 10.* Let:

$A=(E_A,R_A)$ and $B=(E_B,R_B)$;

$\tau$ is an impact of $A$ on $B$, i.e. $\tau=(A\to B)_\tau$. , $\tau\in\Omega_{AB;}$

$\exists$ entity $C=(E_C,R_C):C\neq A,\ C\neq B$;

$\exists\ \psi=(B\to C)_\psi$ which can be composed with $\tau=(A\to B)_\tau$;

$\exists$ transitive impact $\xi=\{\tau,\psi\}=(A\to B\to C)_\xi$ ;

$\exists$ impact $\varphi=(A\to C)_\varphi$; , $\varphi\in\Omega_{AC}$ and $\varphi\neq\xi$;

$F_\varphi$ is a reflection of the impact $\varphi$ and $F_\xi$ is a reflection of the impact $\xi$.

$F_\tau$ is information for $A$ in $B$ if $\exists r\in R_C:(F_\varphi\to F_\xi)_r$ ■

The entity $A$ is called source, the entity $B$ is called recipient. The relation $r\in R_C$ for which $(F_\varphi\to F_\xi)_r$ is called reflection evidence and the entity $C$ is called information evidence.

If $V_{AB}$ is an interactive reflection of between entities $A$ and $B$, and entity $C$ contains reflection evidences for all reflections of $V_{AB}$ than $C$ is called information witness.

*Every reflection may be considered as information iff there is corresponded information evidence or information witness.*

For practical needs, it is more convenient to follow the next consideration.

The reflection in the recipient represents both the relationships and the sub-entities of the source. From other point of view, the relationships build up and present the entities. Because of this, the reflected relationships are the essence of the reflection. In other words, if a reflection evidence exists then the reflection of the forming relationship may be considered as "information" for reflected entity.

Therefore, in the sense that the evidence exists to point what relationship (between what entities) is reflected and where it is done, we may say that "*the information is reflected relationship".*

## Conclusion

The translation of the philosophical theory into the formal one is a good approach for verification of the scientific ideas. The concept “Information” of the GIT was presented formally in this paper. The definition given above is a first step for building the formal part of the GIT. Together with the philosophical explanations, it gives us a useful tool for investigation of the information phenomena in the real world.

## Acknowledgements

This paper is based on the ideas considered during very creative discussions at the Int. Conf. "KDS 1997", September 1997, Yalta, Ukraine, and at the Int. Conf. "ITA 2000", September 2000, Varna, Bulgaria, as well as at the others scientific meetings organised by the International Workgroup on Data Base Intellectualisation (IWGDBI). Special discussion was provided at the Int. Conf. “Knowledge-Dialog-Solution”, Varna, Bulgaria, 2003.

Authors are very grateful to all participants in these fruitful discussions.

This work is partially financed by project ITHEA-XXI of the Institute of Information Theories and Applications FOI ITHEA.

## Bibliography

[Markov, 1988] Kr.Markov. From the past to the future of the definition of the concept of Information. Proceedings "PROGRAMMING '88", BAS, Varna 1988, p.150. (In Bulgarian).

[Markov et al, 1993] Kr.Markov, Kr.Ivanova, I.Mitov. Basic Concepts of a General Information Theory. IJ "Information Theories and Applications". FOI ITHEA, Sofia, 1993, Vol.1,No.10, pp.3-10

[Markov et al, 2003] Kr.Markov, Kr.Ivanova, I.Mitov. General Information Theory. Basic Formulations. FOI ITHEA, Sofia, 2003.

## Author information

Krassimir Markov - ITHEA - FOI Institute of Information Theories and Applications; Institute of Mathematics and Informatics, BAS; P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: foi@nlcv.net

Krassimira Ivanova - Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: foi@nlcv.net

Ilia Mitov - ITHEA - FOI Institute of Information Theories and Applications, P.O.Box: 775, Sofia-1090, Bulgaria;
e-mail: foi@nlcv.net

Evgeniya Velikova-Bandova – Sofia University, Faculty of Mathematics and Informatics. e-mail: velikova@fmi.uni-sofia.bg

# INTELLIGENT SYSTEMS MEMORY STRUCTURING

*V. Gladun*

*Abstract: The requirements for the memory structuring of intelligent systems are discussed. Simultaneously with the introduction of information into memory there should take place the processes of association links (bonds) formation, hierarchy systematizing, classification and concept formation. The growing pyramidal networks (GPN) meet these requirements. Many years of experience of GPN application for data analyses in chemistry and material studies proves their sufficiently high potential.*

*Keywords: intelligent systems, growing pyramidal networks, data analysis.*

## 1. Perception, presentation and analysis of information in intelligent systems

1.1. Competition of a computer with a man in resolving of intelligent problems more and more often ends up in the victory of a computer. But there arises the evident contradiction: computer genius victoriously solves the most complicated multivariate problems in artificial, relatively poor media (chess, for example) and “stumbles” over solving simple (for a man) life problems, requiring still quick understanding and assessment of multi-component situation.

What compensates the evident advantage of a computer over a man in quickness? The answer is unanimous that the reason lies mainly in memory structure. We will try to identify the structural peculiarities of memory, which, to our mind, are necessary for formation of clear explanations of a human phenomenon of information processing.

1.2. The peculiarity of intelligent systems, which causes no doubts, is the ability to analyze the perceived information. The process of analysis consists in picking up the integral parts and characteristic attributes of the analyzed whole. The product of analysis is organized (for example, with the help of logic links) sum of object attributes.

Thus, irrespective of the type of the information perceived (continuous or discrete) at a certain stage of analysis there appears a discrete representation of objects in the form of arranged aggregate of information blocks – attributes. The attributes serve as building blocks for the following analytical processes, leading in the long run, towards formation of generalized information models of the objects perceived. So, discretization of the perceived information, which consists in demonstrating attributes of objects, is an important peculiarity of intelligent systems.

1.3. The prevailing tendency in developing of intelligent systems is the improvement of man – machine interaction until the achievement of partner level of man-machine relations. That is why it is important to use natural, pertaining to a man principles of problems, situations and media modeling in computers. Partner model types (a man and a computer) should be similar. In life activity of a man a very important role is played by logic - linguistic information models, i.e. such models where the main elements are not numbers and calculations but names and logical bonds. Logic - linguistic models are adequately described with natural language constructions, and it is one of their decisive merits for designing of a man – machine interface. In computers to come there should be created conditions for man – machine solving of problems in partner mode providing switching over from a computer to a man and vice versa within the process of solving of problem. Such mode could be set up only by means of adjustment of information model types, used by partners. Logic – linguistic models are the most acceptable model types for such an adjustment.

1.4. Formation of memory structure is done simultaneously with perception of information and under the impact of the information perceived and already stocked. The memory structure reflects the information perceived. Information structuring is an indispensable function of memory.

The main processes of structuring include formation of associative links by means of identifying the intersections of attributive representations of objects, hierarchic regulation, classification, forming up generalized logical attributive models of classes, i.e. concepts.

Under real conditions of information perception there is often no possibility to get whole information about an object at once (for example, because of faulty foreshortening or lighting during the reception of visual

information). That is why the processes of memory formation should allow for the possibility of "portioned" construction of objects models and class models by parts.

1.5. In different processes of information processing objects are represented by one of the two means: by a name (convergent representation) or by a set of meanings of attributes (displayed representation). The structure of memory should provide convenient transition from one representation to another. Mechanisms, providing such transition in neuro system of a man at recognition or recollection are considered in the works of S.G. Voronkov and Z.L.Rabinovich [1].

Let us sum up the above mentioned theses in the form of requirements to memory structuring in intelligent systems.

- In intelligent systems knowledge of different types should be united into net-like structure, designed according to principles common for all types of knowledge.
- The network should reflect hierarchic character of real media and in this connection should be convenient for representation of gender-type bonds and structures of composite objects.
- Obligatory functions of the memory should be formation of association bonds by revealing intersections of attributive object representations, hierarchic structuring, classification, concept formation.
- Within the network there should be provided a two-way transition between convergent and displayed presentations of objects.

## 2. Growing Pyramidal Network

The above mentioned requirements are met by growing pyramidal network. The theory and practical application of growing pyramidal networks are represented in many publications [2-5]. In this paper we present somewhat changed rules of formation of growing pyramidal network, ensuring their construction at the introduction of object attributive descriptions by parts. The example of a growing pyramidal network is presented in Fig.1.

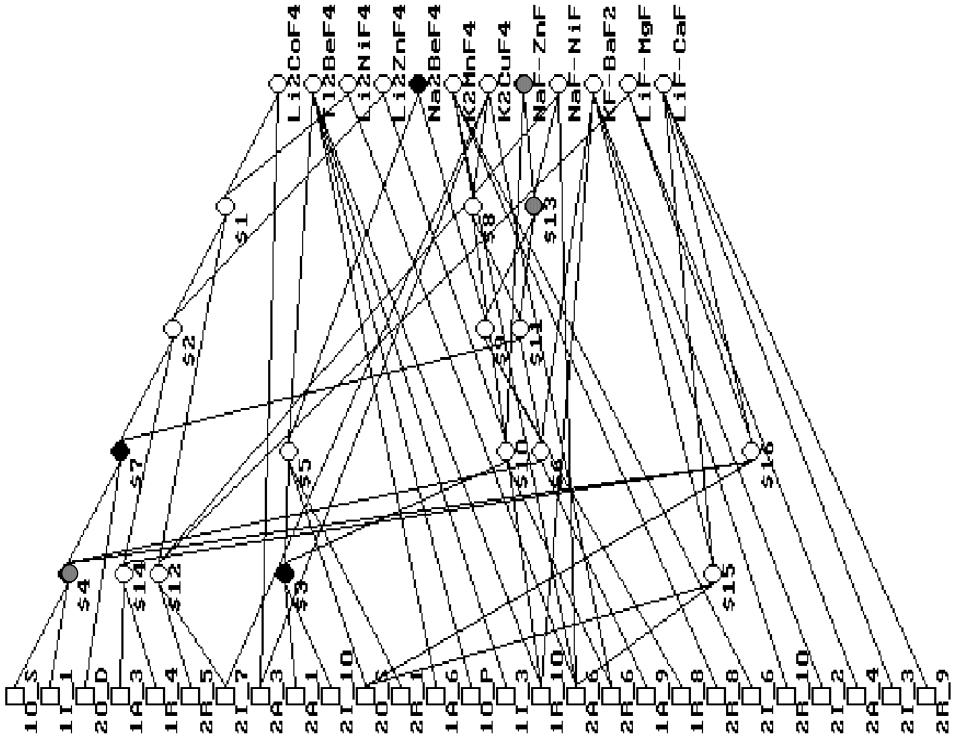



Fig. 1.

*The pyramidal network* is called a cycled oriented graph, where there are no vertices having one incoming arch. The vertices that have no incoming arch are called *receptors,* other vertices are called *conceptors*. The

subgraph of the pyramidal network which includes *a* vertex and all vertices, from which there are paths to *a* vertex, is called *pyramid* of *a* vertex. The vertices pertaining to *a* vertex pyramid form up its *subset*. The set of vertices towards which there are paths from *a* vertex is called *superset*.

In a subset and a superset of a vertex there are 0-subset and 0-superset that consist of those vertices that are immediately connected with it. While constructing a network the sets of meanings describing some objects (materials, aggregate states, situations, illnesses etc.) serve as incoming information. Receptors correspond to the meanings of attributes. In different problems they could be names of properties, relationships, states, actions, objects and objects classes. Conceptors correspond to the descriptions of objects as a whole and intersections of descriptions. The network shown in the Fig.1 is built on the basis of the Table where the objects are pairs of chemical elements, forming and not forming compounds.

Table

| Object | Class | 1O | 1A | 1R | 1I | 2O | 2A | 2R | 2I |
|---|---|---|---|---|---|---|---|---|---|
| Li2CoF4 | A | S | 3 | 4 | 1 | D | 3 | 5 | 7 |
| Ti2BeF4 | A | P | 6 | 10 | 3 | S | 1 | 1 | 10 |
| Li2NiF4 | A | S | 3 | 4 | 1 | D | 6 | 5 | 7 |
| Li2ZnF4 | A | S | 3 | 4 | 1 | D | 1 | 6 | 10 |
| Na2BeF4 | A | S | 9 | 8 | 1 | S | 1 | 1 | 10 |
| K2MnF4 | A | S | 9 | 10 | 1 | D | 6 | 8 | 6 |
| K2CuF4 | A | S | 9 | 10 | 1 | D | 3 | 6 | 7 |
| NaF-ZnF2 | B | S | 9 | 8 | 1 | D | 1 | 6 | 10 |
| NaF-NiF2 | B | S | 9 | 8 | 1 | D | 6 | 5 | 7 |
| KF-BaF2 | B | S | 9 | 10 | 1 | S | 6 | 10 | 2 |
| LiF-MgF2 | B | S | 3 | 4 | 1 | S | 4 | 5 | 7 |
| LiF-CaF2 | B | S | 3 | 4 | 1 | S | 6 | 9 | 3 |

In the Table he object descriptions are given, where 1O, 1A, 1R, 1I are the names of attributes, describing the first element of the compound; 2O, 2A, 2R, 2I are the names of attributes, describing the second element of the compound, and the letters and figures in cells are the meanings of the corresponding attributes.

In the initial state the network consists only of receptors. Conceptors are formed as a result of the work of algorithm of network construction. The algorithm described in a number of publications [2-5] is meant for the work in situations, where the attributive description of each object is fully known and is introduced as a whole. With appearing of new attributes, which characterize the object, it is necessary to form a new complete description of the object and to replace the pyramid that represents it with another one, which corresponds to the new description. But as it was mentioned in real situations of functioning of an intelligent agent simultaneous perception of all characteristics of an object is far from possible. In such cases the information about objects comes in parts. Then there arises the necessity to change a bit the algorithm of constructing a network to provide the possibility to include into the existing object pyramids new attributes according to their appearance without replacement of pyramids as a whole. Now we will present the description of the changed algorithm.

At the introduction of an attributive description of an object receptors corresponding to the meanings of attributes coming into the description are transferred into the state of *excitation*.

The excitation is propagated through the network. The conceptor is switched to the state of excitation if all vertices of its 0-subset are excited. Receptors and conceptors preserve the state of excitation within the period of performing all the operations of constructing of the network.

Let at the introduction of the description of some object $F_a$ be the subset of the excited vertices of 0-subset of *a*- vertex; *G* is the set of the excited vertices of the network having no other excited vertices in their supersets.

Introduction of new vertices and arcs is done according to the following rules.

Rule 1.

If *a* vertex is not excited and $F_a$ set contains more than one element, then arcs, connecting vertices from $F_a$ set with *a* vertex are annulled and a new conceptor is introduced, which is connected by incoming arcs with vertices of $F_a$ set and by an outcoming arc with *a* vertex. The new vertex is in the state of excitation.

Fulfillment of the rule 1 is illustrated by Fig.2(I, II). The network II appears after excitation in the network I of 2,3,4,5 receptors. As it follows from the rule 1, the condition of introduction of a new vertex into the network is the situation when a certain vertex is not completely excited (not all vertices but not less than two of its 0-subset are excited). New vertices are introduced into 0-subsets of not completely excited vertices.

After introduction of new vertices into all areas where the condition 1 is met, the rules 2 or 3 are applied.

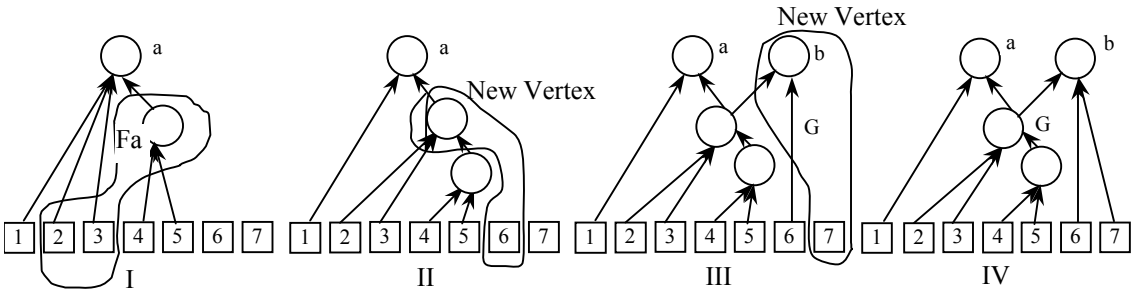



Fig. 2.

Rule 2.

If *G* set contains more than one element and does not include the vertex marked with the name of the object introduced, a new conceptor joins the network and is connected by incoming arcs with all vertices of *G* set. The new vertex is in the state of excitation.

Fulfillment of the rule 2 is illustrated in Fig.2 (II, III). The network III appears after excitation of receptors 2, 3, 4, 5, 6 in the network II.

Rule 3.

If *G* set contains the vertex, marked with the name of the object introduced, this vertex is connected by incoming arcs with other vertices of *G* set.

Fulfillment of the rule 3 is illustrated in Fig.2 (III, IV). The network IV appears after excitation of receptors 2, 3, 4, 5, 6, 7 in the network III under condition that this set of receptors corresponds to the description of *b* object.

In the changed algorithm the possibility of introduction of new attributes into the existing pyramids is provided by the rule 3.

Pyramidal networks are convenient for performing different operations of associative search. For example, one could choose all objects, containing a given combination of attribute meanings, following the paths coming out the vertex, which corresponds to this combination. For the access of all objects the descriptions of which intercross with the description of the given object, it is enough to trace the paths coming out of the vertices which form up its pyramid. All processes, connected with construction of the network at the processing of one description are localized in a relatively small part of the network, i.e. in the pyramid, corresponding to this description.

Hierarchical structure of the networks, which allows them to reflect the structure of composing objects and gender-species bonds naturally, is an important property of pyramidal networks.

Conceptors of the network correspond to the combinations of attribute meanings, defining conjunctive classes of objects. With insertion of excited vertices into the pyramid of the object there takes place linking of the object with the classes, the definitions of which are represented by these vertices. Thus, while building a network, there form up conjunctive classes of objects, i.e. classification without a teacher takes place. Classifying properties of a pyramidal network are very important for automation of media and situation modeling.

The transfer from convergent representations of objects (conceptors) to displayed ones (sets of receptors) is fulfilled by a survey of pyramids in different directions.

In growing pyramidal networks there are realized the processes of forming generalized logical models of object classes, i.e. concepts.

The formed up concept of any complexity is represented in the network by an ensemble of specially picked out vertices. On the basis of network analysis a special procedure builds up a concept in the form of a logical expression.

Logical expressions defining classes of objects are united in Cluster Data Base (CDB). CDB contained information about object groups (clusters) that are specific for the domain under investigation. CDB are used for classification, diagnostics and prognostication.

When the concept for a certain class is formed, the problems of forecasting and diagnostics are reduced to the problem of classification. Classification of new objects is done by comparing their attributive descriptions with the concept, determining the class of objects to be forecasted or diagnosed. The objects could be classified calculating the meaning of logical expressions, representing the corresponding concepts.

In a pyramidal network the information is stocked by its reflection in the network structure. Information about objects and object classes is represented by ensembles of (pyramid) vertices, distributed along the whole network. Introduction of new information causes redistribution of bonds between the vertices of the network, i.e. change of its structure.

Of course, the benefits of pyramid networks are fully demonstrated with their physical realization, which allows parallel spreading of signals through the network.

There is the analogy between the main processes taking place in growing pyramidal networks and neuron networks. The decisive advantage of a growing pyramidal network is the fact that its structure is formed completely automatically depending on the introduced data. As a result there is achieved optimization of information presentation due to adaptation of the network structure to the structural peculiarities of the data. Unlike neuron networks the adaptation effect is achieved without introduction of a priori excess of the network. The learning process does not depend on predetermined network configuration. The drawback of neuron networks if compared to the growing pyramidal network is also the fact that generalized knowledge in them cannot be represented in the form of rules or logical expressions. It makes their interpretation and understanding by a man difficult.

The program system CONFOR (CONcept FORmation) that implements methods of data analysis on the basis of growing pyramid networks has been tested by time. The typical applied problems, for solving of which this system was used are: forecasting new chemical compounds and materials with the indicated properties, forecasting in genetics, geology, medical and technical diagnostics, forecasting malfunction of complex machines and sun activity.

## 3.Summary

Pyramidal network is a network memory, automatically tuned into the structure of incoming information. Unlike the neuron networks, the adaptation effect is attained without introduction of a priori network excess.

The research done on complex data of great scope showed high effectiveness of application of growing pyramidal networks for solving analytical problems. Such qualities as simplicity of change introduction, combining processes of information introduction with processes of classification and generalization, high associativity makes growing pyramid networks an important component of forecasting and diagnosing systems.

## Bibliography

1. Voronkov G.S. Rabinovich Z.L. Natural media of memory and thinking: model // Intern.conf.: Knowledge-Dialogue-Solution-2001.-St.-Pb.-2001(in Russian).
2. Gladun V.P. Partnership with computers. Man-computer task-oriented systems.-Kiev: ”Port-Royal”,2000(in Russian).
3. Gladun V.P. Solution planning.-Kiev: Naukova dumka, 1987(in Russian).
4. Gladun V.P. Processes of new knowledge formations.-Sofia: ”Pedagog 6”,1994(in Russian).
5. Gladun V. Vaschenko N. Analytical processes in pyramidal networks. // International Journal: Information Theories and Applications.-2000, V.7, №3.

## Author information

Victor Gladun – V. M. Glushkov Institute of Cybernetics, National Academy of Science of Ukraine, Prospect akad. Glushkova 40, 03680, Kiev, Ukraine; e-mail: glad@aduis.kiev.ua

(The work has been done within the framework of the project INTAS #00-397)

# DISTANCES BETWEEN PREDICATES IN BY-ANALOGY REASONING SYSTEMS

*V. Koval, Yu. Kuk*

*Abstract: The purpose is to develop expert systems where by-analogy reasoning is used. Knowledge "closeness" problems are known to frequently emerge in such systems if knowledge is represented by different production rules. To determine a degree of closeness for production rules a distance between predicates is introduced. Different types of distances between two predicate value distribution functions are considered when predicates are "true". Asymptotic features and interrelations of distances are studied. Predicate value distribution functions are found by empirical distribution functions, and a procedure is proposed for this purpose. An adequacy of obtained distribution functions is tested on the basis of the statistical $\chi^2$ –criterion and a testing mechanism is discussed. A theorem, by which a simple procedure of measurement of Euclidean distances between distribution function parameters is substituted for a predicate closeness determination one, is proved for parametric distribution function families. The proposed distance measurement apparatus may be applied in expert systems when reasoning is created by analogy.*

*Keywords: expert systems, production rules, predicates, distances between predicates, by-analogy reasoning.*

## Introduction

Partnership systems are known to be the ones [1] able not only to use experts' knowledge, but also to derive themselves new knowledge from data accumulated in memory. They have means used to derive knowledge from data represented as statistical or empirical "object-feature-time"-type tables [2]. While inferences are obtained in traditional expert systems only deductively, partnership systems use additionally inductive inference features, by-analogy reasoning construction facilities and non-monotone reasonings [1]. The by-analogy reasoning creation basis is the rule that resembling conditions entail resembling effects in immediate proximity to known productions. Therefore, to construct a by-analogy reasoning mechanism, one should be able to compare a condition and an effect resemblance degree. A knowledge in expert systems is usually represented as "if $X_1 \,\&\, X_2 \,\&\, \ldots \,\&\, X_m$, then $A$"–type productions. Compare two productions, for instance, by some PROLOG language features, and left and right sides of both productions are compared. Productions coincide if compared predicates fully coincide. If productions do not coincide, then partnership systems take a non-coincidence degree into account. For this purpose, a distance between predicates is introduced in such systems, and it becomes possible to measure a degree to which one production resembles another. Thus, it is also possible to construct a by-analogy reasoning inference mechanism. By-analogy reasonings may be illustrated by the following example. Assume that it is necessary to check whether conditions $X_1 \,\&\, X_2 \,\&\, \ldots \,\&\, X_m$ lead to an effect $A$. An inference system detects that a knowledge base (KB) contains a resembling knowledge, i.e. "if $Y_1 \,\&\, Y_2 \,\&\, X_3 \,\&\, \ldots \,\&\, X_m$, then $A$ ", a truth of which is equal to $P$. The conditions $Y_1$ and $Y_2$ do not coincide with $X_1$ and $X_2$ in this knowledge. Their non-coincidence degree is calculated. Hence, find a distance $d(X,Y)$ between the predicates $X = X_1 \,\&\, X_2$ and $Y = Y_1 \,\&\, Y_2$, and, if it does not exceed a threshold $\eta$, the conclusion is that $A$ is probable. The truth of this inference is $P' < P$. A truth lowering value depends on a length of a distance between $X$ and $Y$. The by–analogy inference rule scheme may be represented as

$$\frac{B', B \to A, d(B', B) < \eta}{A}. \qquad (1)$$

Example 1. Let a predicate subject domain be a set of real functions $f(x)$ with one variable. Consider three predicates: 1) predicate $B'$, i.e. "to be function $\frac{\sin(x)}{x}$"; 2) predicate $B$, i.e. "to be polynomial $f_n(x)$ with power exponent *n=2m*", where

$$f_n(x) = 1 - \frac{x^2}{2 \cdot 3} + \frac{x^4}{2 \cdot 3 \cdot 4 \cdot 5} - \frac{x^6}{2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7} + \cdots + (-1)^m \frac{x^{2m}}{2 \cdot 3 \cdot 4 \cdot \cdots \cdot (2m+1)} \text{»;} \quad (2)$$

and 3) predicate A, i.e. "to be represented as product of linear co-factors $f(x) = \prod_{i=1}^{n} (x - \alpha_i)$, where $\alpha_i, i = 1, \ldots, n$ are roots of equation $f(x) = 0$". The expression $B \to A$ is known [3]. Calculate the distance between $B$ and $B'$ by the following formula: $d(B', B) = \sup_x |g(x) - f_n(x)|$.

When $n$ is chosen, this distance can be made shorter than any number $\eta$ that is as low as possible: $d(B', B) \le \eta$. This fact be proved, if the function $g(x) = \frac{\sin(x)}{x}$ is expanded into Taylor series:

$$g(x) = \frac{\sin(x)}{x} = 1 - \frac{x^2}{2 \cdot 3} + \frac{x^4}{2 \cdot 3 \cdot 4 \cdot 5} - \frac{x^6}{2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7} + \cdots + (-1)^m \frac{x^{2m}}{2 \cdot 3 \cdot 4 \cdot \cdots \cdot (2m+1)} + \ldots . \quad (3)$$

Since $d(B', B) \le \eta$, then $B' \to A$ is the by-analogy inference (expression (1)), i.e. the function $g(x) = \frac{\sin(x)}{x}$ can also be expanded into linear co-factors. Since the roots of the equation $g(x) = \frac{\sin(x)}{x} = 0$ are $\pi, -\pi, 2\pi, -2\pi, \ldots$, then, when obtained by analogy, expansion (3) has the following form: $\frac{\sin x}{x} = (1 - \frac{x^2}{\pi^2})(1 - \frac{x^2}{4\pi^2}) \cdots (1 - \frac{x^2}{n^2\pi^2}) \cdots$. Pursuant to this formula, it is possible to determine the factor under $x^2$, i.e. $-(\frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \cdots)$, and to make the latter equal to the factor under $x^2$, i.e. to $-\frac{1}{2 \cdot 3}$, in expansion (3). And $\frac{1}{2 \cdot 3} = \frac{1}{\pi^2} + \frac{1}{4\pi^2} + \frac{1}{9\pi^2} + \cdots$ is the result, from which the famous Euler formula follows:

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} \cdots = \frac{\pi^2}{6} .$$

## 1. Distances between Predicates

*1.1 Empirical Predicate Distribution Functions.* An $m$-ary predicate $X = X(y_1, \ldots, y_m)$ is understood as a function, values of which are statements about *m* objects. Such objects are predicate argument values. A predicate is an object "feature" under $m = 1$ and it is a "relation" between *m* objects under $m > 1$.

Introduce the notions of empirical frequencies and of predicate value distribution functions needed in order to compare two "resembling" predicates $X$ and $Y$. Consider the following cases.

1) $m = 1$ and a number of different true statements about an object feature is finite and equal to $K$. Bring an integer number, respectively, $1, 2, \ldots K$ in correspondence with each such statement. Let there be $n$ objects from some subject domain and, respectively, $n$ true statements about a single feature of every such object. Define an empirical frequency for an $i$-th statement as $p_i = \frac{k_i}{n}$, where $k_i$ is a number of $i$-th statements from among a whole number of $n$ true statements. Pursuant to these frequencies, define an empirical distribution function $F_n^*(x)$ as a step function of a real variable $x$. This function is equal to zero under $x \le 1$, to $p_1$ under $1 \le x < 2$, to $p_1 + p_2$ under $2 \le x < 3$, … , and is to 1 under $x \ge K$. The derived empirical frequencies $p_i$, $i = 1, \ldots, K$, $\sum_{i=1}^{n} p_i = 1$, and $F_n^*(x)$ characterize this predicate well enough.

2) And now here is the general case when $m > 1$ and a number of different true statements has a power of a continuum. Construct an empirical distribution function $F_n^*(x)$. Let there be $n$ selections that have *m* objects

from some subject domain and $n$ true statements about relations between $m$ objects from each selection. Bring a real number from the space $R_1$ in correspondence with each such statement. The result is that there are $n$ numbers $x'_1,\ldots,x'_n$ on the straight line $R_1$. Arrange these numbers in the ascending order, i.e. the variational series $x_{(1)} \le \ldots \le x_{(n)}$ is formed. Define $F_n^*(x)$ as a step function with the steps equal to $1/n$. It is the function of a real variable $x$, and it is equal to zero under $x \le x_{(1)}$, to $k/n$ under $x_{(k)} \le x < x_{(k+1)}$, $k = 1,\ldots,n-1$, and to 1 under $x \ge x_{(n)}$.

3) Consider the formula $X = X_1 \,\&\, X_2 \,\&\ldots\&\, X_K$. Construct an empirical distribution function $F_n^*(x)$ for the formula $X$. Let a predicate $X_i$ be $m_i$-ary, $m_i \ge 1$ and a number of different true statements about $m_i$ objects has a power of a continuum. Then, bring the values of $X_i$ in correspondence with the $i$-coordinates of the points from the $K$-dimensional space $R_K$. Let there be $n$ selections that have $m$ objects ($m = \sum_{i=1}^{u} m_i$) from some subject domain and $n$ true statements about relations between $m$ objects from each selection. To reflect relations between $m$ objects, bring the $K$-dimensional vector from the space $R_K$ in correspondence with each such statement. The result is that there are $n$ vectors $x'_1,\ldots,x'_n$ from $R_K$, where $x'_i = (x'_{i,1},\ldots,x'_{i,K})$. Define $F_n^*(x)$, where $x = (x_1,\ldots,x_K) \in R_K$, as follows. Consider a set $B_x = \{y \in R_K : y_i < x_i,\ i = 1,\ldots,K\}$. Denote a number of $x'_1,\ldots,x'_n$ by $\nu(B_x)$ as for the vectors that got into $B_x$. Assume the following: $F_n^*(x) = \nu(B_x)/n$, $x \in R_K$.

It can be shown by analogy with Glivenko-Cantelli theorem [4] that the following assertion is valid for empirical distribution functions: $F_n^*(x)$ converges under $n \to \infty$ to some single limited predicate distribution function $G(x)$.

*1.2. Calculating a Distance between Predicates.* Differences in functions of distribution of two predicates or formulas can be used in by-analogy reasoning systems in order to compare two "resembling" predicates or formulas. Let $G(x)$ and $Q(x)$ be predicate value probability distribution functions or formula value probability distribution functions, respectively, for $X$, the first predicate or formula, and for $Y$, the second predicate or formula. In practice, empirical distribution functions or distribution function estimates are used as the former ones. They are selected from appropriate standard parametric distribution function families and tested for adequacy. The distribution function estimate derivation methodology is considered below.

Definition 1. *A distance $d(X,Y)$ between predicates or* formulas *$X$ and $Y$ is a distance $d(G,Q)$ between two value distribution functions $G(x)$ and $Q(x)$ when these predicates or* formulas *are true under their values.*

Consider the distance $d$ between two formulas $X = X_1 \,\&\, X_2 \,\&\ldots\&\, X_u$ and $Y = Y_1 \,\&\, Y_2 \,\&\ldots\&\, Y_w$ for the case when a "feature" or a "relation", described by each separate predicate, are by no means associated with "features" or "relations" described by other predicates. Let $G_{X_1}, G_{X_2},\ldots,G_{X_u}$ and $Q_{Y_1}, Q_{Y_2},\ldots,Q_{Y_w}$ be the distribution functions, respectively, for $X_1, X_2,\ldots,X_u$ and $Y_1, Y_2,\ldots,Y_w$. Then, $d$ between $X = X_1 \,\&\, X_2 \,\&\ldots\&\, X_u$ and $Y = Y_1 \,\&\, Y_2 \,\&\ldots\&\, Y_w$ is equal to the distance between two products of the respective distribution functions $G_X = G_{X_1} \cdot G_{X_2} \cdot\ldots\cdot G_{X_u}$ and $Q_Y = Q_{Y_1} \cdot Q_{Y_2} \cdot\ldots\cdot Q_{Y_w}$.

If an expression for $d$ between predicates or formulas is chosen correctly, it is possible to use further on "good" features of this distance, for instance, the distance calculation procedure itself may be simplified. Consider various expressions used to calculate distances between $X$ and $Y$. The distance

$$d(X,Y) = d(G,Q) = \sup_x |G(x) - Q(x)| \qquad (4)$$

*means an absolute deviation of values for one distribution function with respect to another distribution function at each point and the distance*

$$d(X,Y) = d(G,Q) = \int (G(x) - Q(x))^2 \, dQ(x) \qquad (5)$$

takes a root mean square deviation of these values into account.

Example 2. Calculate the distance $d$ between the predicates $B'$ and $B$ from Example 1. For every $x$, the real value for $B'$ is equal to $g(x)=\frac{\sin(x)}{x}$. Therefore, the distribution function $G(y)$ for this predicate is equal to zero under $y<g(x)$ and to 1 under $y\geq g(x)$. The values of $B$ correspond to the values of the polynomial $f_n(x)$: $y_1=f_1(x)$, $y_2=f_2(x)$, ..., $y_n=f_n(x)$ that, under different and sufficiently large $n$, $n>n_1$, are arranged in a certain way within the interval $\Delta$ of the following form: $\Delta=[g(x)-\eta, g(x)+\eta]$. However, when $n$ increases, the points $f_n(x)$ approach the point $g(x)$ because of $f_n(x)\to g(x)$. The distribution functions $Q(y)$ for these points are not found, since only the upper estimate for $d$ between $B'$ and $B$ is important. The following is made: move each of these points away from $g(x)$ in such a way that they fill in the interval $\Delta$ uniformly. The result is that the distribution function $\widetilde{Q}(y)$ in its new position becomes uniform, but the distance between $G(y)$ and the new $\widetilde{Q}(y)$ increases here in comparison with the previous one between $G(y)$ and $Q(y)$ Therefore: $d(B,B')=d(G,Q)<d(G,\widetilde{Q})$. Since

$$\widetilde{Q}(y)=\begin{cases} n_1/n, & when\ y=g(x)-\eta \\ \frac{n-n_1}{2n\eta}y+\frac{n_1}{n}-\frac{n-n_1}{2n\eta}(g(x)-\eta), & when\ g(x)-\eta\leq y\leq g(x)+\eta \\ 1, & when\ y\geq g(x)+\eta \end{cases}$$

then, if formula (5) is used, the following expression takes place: $d(G,\widetilde{Q})=\frac{1}{8}(1-\frac{n_1}{n})^3\eta<\frac{\eta}{8}$. Hence, the upper estimate is derived for $d(B,B')$. Thus, if $\sup_x|g(x)-f_n(x)|\leq\eta$, then $d(B,B')<\eta/6$. Therefore, these formulas for the distances are equivalent.

*1.3 Kulbak–Leibler Distance, $\chi^2$–Distance, Hellinger Distance.* Consider now different types of distances between two predicates $X$ and $Y$ for the case when their distribution functions $Q$ and $G$ have, respectively, the densities $q(x)$ and $g(x)$ as for a measure $\mu$. The Lesbegue measure may be used for one group of distribution functions (absolutely continuous distributions) and a counting measure may be taken for another group (discrete distributions) as $\mu$. Let $N_Q$ be a carrier of $Q$ ($N_Q=\{x: q(x)>0\}$), and let $N_G$ be a carrier of $G$ ($N_G=\{x: g(x)>0\}$). The Kulbak–Leibler distance between $X$ and $Y$ is calculated in the following way:

$$r_1(X,Y)=r_1(\mathrm{G},Q)=\int_{N_G}\ln\frac{g(x)}{q(x)}g(x)\mu(dx).$$

The $\chi^2$–distance between $X$ and $Y$ is

$$r_2(X,Y)=r_2(\mathrm{G},Q)=\int_{N_Q\cup N_G}\frac{(q(x)-g(x))^2}{g(x)}\mu(dx).$$

The values of $r_1(X,Y)$ and $r_2(X,Y)$ are more than or equal to zero. However, the equalities $r_1(X,Y)=0$ and $r_2(X,Y)=0$ are possible only under $Q=G$. Since $r_1(X,Y)$ and $r_2(X,Y)$ are not the symmetric functions of $Q$ and $G$, then $r_1(X,Y)$ and $r_2(X,Y)$ are not the distances in the general case because of $r_1(X,Y)\neq r_1(Y,X)$ and $r_2(X,Y)\neq r_2(Y,X)$. Nevertheless, essentially speaking and from the statistical point of view, $r_1(X,Y)$ and $r_2(X,Y)$ characterize a deviation of $Q$ from $G$.

The Hellinger distance between $X$ and $Y$ is

$$r_3(X,Y)=r_3(\mathrm{G},Q)=\int_{N_Q\cup N_G}\left(\sqrt{g(x)}-\sqrt{q(x)}^2\right)\mu(dx).$$

and it is already the symmetric function for $X$ and $Y$. The value $\sqrt{r_3(Q,G)}$ possesses all the metric characteristics between the functions $\sqrt{q(x)}$ and $\sqrt{g(x)}$ in the metrical space $L_2$.

Consider the features of these distances, important when a predicate resemblance threshold is chosen. If a predicate closeness degree is characterized by such distances when $q(x)/g(x)$ is close to 1, then the following result turns out to take place:

$$r_1(Q,G) \approx \frac{1}{2} r_2(Q,G) \approx 2r_3(Q,G).$$

Asymptotically, all the distances behave in the same way. To study this asymptotic feature, assume that $G$ and $Q$ for $X$ and $Y$ are taken from one and the same parametric family and defined, respectively, by the parameters $\theta$ and $\theta+\Delta$. Then, the rate of the convergence to zero for the distance between $X$ and $Y$ is equal to $O(\Delta^2)$ under $\Delta \to 0$. This fact follows from the asymptotic equality

$$r_3(\Delta) \approx \frac{I(\theta)}{4}\Delta^2,$$

where $I(\theta)$ is the Fisher information found by the formula

$$I(\theta) = \int \frac{(g'_\theta(x))^2}{g_\theta(x)} \mu(dx).$$

*1.4 Predicate Comparison Procedure Simplification Theorem.* The predicate resemblance determination procedure falls into two stages: 1) calculate a distance between predicates; and 2) compare a calculated distance with a threshold $\eta$. Let $G$ and $Q$ be distribution functions for predicates $X$ and $Y$ that belong to the same parametric family $\Psi = (G_\theta | \theta \in \Theta)$ and differ only in their parameters. Assume that $G_{\theta_1}$ and $G_{\theta_2}$ are, respectively, the predicate value distribution functions for $X$ and $Y$. Consider the Kulbak-Leibler, $\chi^2$- and Hellinger distances as the ones between predicates: $\rho_i(\theta_1,\theta_2)$, $i=1,2,3$. The following theorem is true.

Theorem 1. *Assume that value distribution functions for predicates $X$ and $Y$ belong to a parametric distribution function family $\Psi = (G_\theta / \theta \in \Theta)$. Let the following conditions be met: 1) a parametric set $\Theta$ is compact; 2) $G_{\theta_1} \neq G_{\theta_2}$ under $\theta_1 \neq \theta_2$; 3) for every $\theta \in \Theta$, Fisher information is restricted: $0 < I(\theta) \le 4b < \infty$. Then, $\rho_i(\theta_1,\theta_2) \le \delta$, $i=1,2,3$ is equivalent to $(\theta_1-\theta_2)^2 \le \delta / b_i$, where $b_i$, $i=1,2,3$ are constant, $b_1 = 2b$, $b_2 = 4b$, $b_3 = b$.*

This theorem reduces the predicate resemblance determination procedure to the simple procedure by which a Euclidean distance between distribution function parameters is determined. The $\Theta$-set compactness condition is not assumed to be restricting and it means that $\Theta$ is restricted. The second condition means that $\rho_i(\theta_1,\theta_2) > 0$ takes place under $\theta_1 \neq \theta_2$.

*1.5. Distribution Function Estimates.* As a rule, a predicate distribution function is not known. It is not very convenient to deal with empirical distribution functions. Therefore, the already known classes of distributions are used and estimates for $G(x)$ are created. Assume that an unknown estimate of $G(x)$ for a predicate $X$ belongs to $\Psi = (G_\theta / \theta \in \Theta)$. Construct an empirical function $G_n^*$ for $X$. Let $G*$ be a function from $\Psi$ that is closest to $G_n^*$ as for a distance $d$, i.e.

$$d(G^*, G_n^*) = \min_{\Pi \in \Psi} d(\Pi, G_n^*).$$

$G*$ with the parameter $\theta*$ is an estimate for $G(x)$ as for a minimum of $d$.

Consider the practical methods used to create the estimates for distribution functions. First of all, describe the $\chi^2$–procedure that helps to find estimates. In this case, the distance

$$d(G,Q)=\sum_{i=1}^{r}\frac{(P_G(\Delta_i)-P_Q(\Delta_i))^2}{P_G(\Delta_i)}$$

is used as $d$ ; $\Delta_1,\ldots,\Delta_r$ are non-intersecting sets of a predicate value space $R$ and their union is equal to $R$;

$$P_G(\Delta_i)=\int_{\Delta_i} dG(x),\ P_Q(\Delta_i)=\int_{\Delta_i} dQ(x),\ i=1,\ldots,r.$$

Take $G_n^*(x)$ as $Q(x)$. The estimate $\theta*$ as for the given minimum distance is a value of $\theta$, and

$$d(G_\theta,G_n^*)=n\sum_{i=1}^{r}\frac{\left(P_\theta(\Delta_i)-\frac{\nu_i}{n}\right)^2}{P_\theta(\Delta_i)}=\sum_{i=1}^{r}\frac{(nP_\theta(\Delta_i)-\nu_i)^2}{nP_\theta(\Delta_i)}. \quad (6)$$

is minimized under this distance; in the present case, $\nu_i=nG_n^*(\Delta_i)$ is a number of predicate values that got into the set $\Delta_i$ and under which a predicate is "true". Differentiate expression (6) with respect to the parameters, the components of which make up the vector $\theta$, make the derivatives equal to zero, and the equation system is derived relative to unknown parameters. Solve this system and find the estimates for the parameters. The obtained $G*(x)$ is then tested for adequacy. If a test result shows that $G*(x)$ is not adequate to the data, then an initial distribution function family should be changed.

Consider the practically important maximum likelihood method also used to derive the estimates. To create a maximally likely estimate means to define one more important distance between an arbitrary $Q$ and $G_\theta$ from $\Psi=(G_\theta\ /\ \theta\in\Theta)$. It is assumed that $G_\theta$ possesses a density $g_\theta(x)$ with respect to a measure $\mu$. Such a distance is expressed by the formula

$$\rho(G_\theta,Q)=-\int \ln g_\theta(x)Q(dx).$$

If an empirical $G_n^*$ is taken as $Q$, then the estimate for $\theta$ is called the maximum likelihood estimate and it minimizes the distance $\rho(G_\theta,G_n^*)$. The yielded function estimates have the "good" features, i.e. they are efficient and asymptotically not biased.

*1.6. Testing for Adequacy.* Obtained distribution function estimates are tested for adequacy before a distance between predicates is found by means of them. An adequacy of a found function is tested for by the $\chi^2$–statistics. The testing mechanism is as follows. Consider the hypothesis that, when a predicate is "true", probable predicate values are distributed by $G*(x)$. Divide a predicate value space into a finite number of sets $\Delta_1,\ldots,\Delta_r$ without common points. Calculate the values of $p_i=P_G(\Delta_i)$. Determine the frequencies $\nu_i$, i.e. a number of predicate values under which it is "true" and that got into a set $\Delta_i$. Calculate the statistics

$$\chi^2=\sum_{i=1}^{r}\frac{(np_i-\nu_i)^2}{np_i}.$$

It is possible to show by analogy with [4] that the $\chi^2$-statistics distribution function does not depend on an initial predicate value distribution function at all under $n\to\infty$. The former function is expressed by the formula

$$w_{r-1}(x)=2^{\frac{1-r}{2}}\Gamma^{-1}\left(\frac{r-1}{2}\right)x^{\frac{r-3}{2}}e^{-\frac{x}{2}},\quad x>0$$

and helps to find the point $x_{0.05}$ for which the expression

$$\int_{x_{0.05}}^{\infty} w_{r-1}(x)dx=0.05$$

takes place. If $\chi^2>x_{0.05}$, then the choice of a distribution function is wrong.

## 2. A By-Analogy Reasoning System Flowchart

Figure 1 depicts a by-analogy reasoning system flowchart. Let the following request be received by the system: "Is effect $A$ possible when conditions $X_1 \& X_2 \& \ldots \& X_m$ are met?" The production rule "if $X_1 \& X_2 \& \ldots \& X_m$ then $A$" is sought for in the KB. If it is found, the answer is positive. If it is absent, the production rule "if $Y_1 \& X_2 \& \ldots \& X_m$, then $A$" is sought for, the conditions of which contain the same predicate names as in the request conditions, but the first predicate differs from first predicate in the request. If this rule is not found in the KB, then such a production rule is sought for, the conditions of which differ from the request conditions already in the second predicate: "if $X_1 \& Y_2 \& \ldots \& X_m$ then $A$". And so on. For more certainty, let the procedure result be that the desired production rule "if $X_1 \& X_2 \& \ldots \ldots \& X_{i-1} \& Y_i \& X_{i+1} \& \ldots \& X_m$ then $A$" is found in the KB at the $i$-th step. However, the predicate $Y_i$ does not coincide in this rule with the predicate $X_i$. Therefore, the procedure is started up that determines a closeness of predicates that do not coincide.

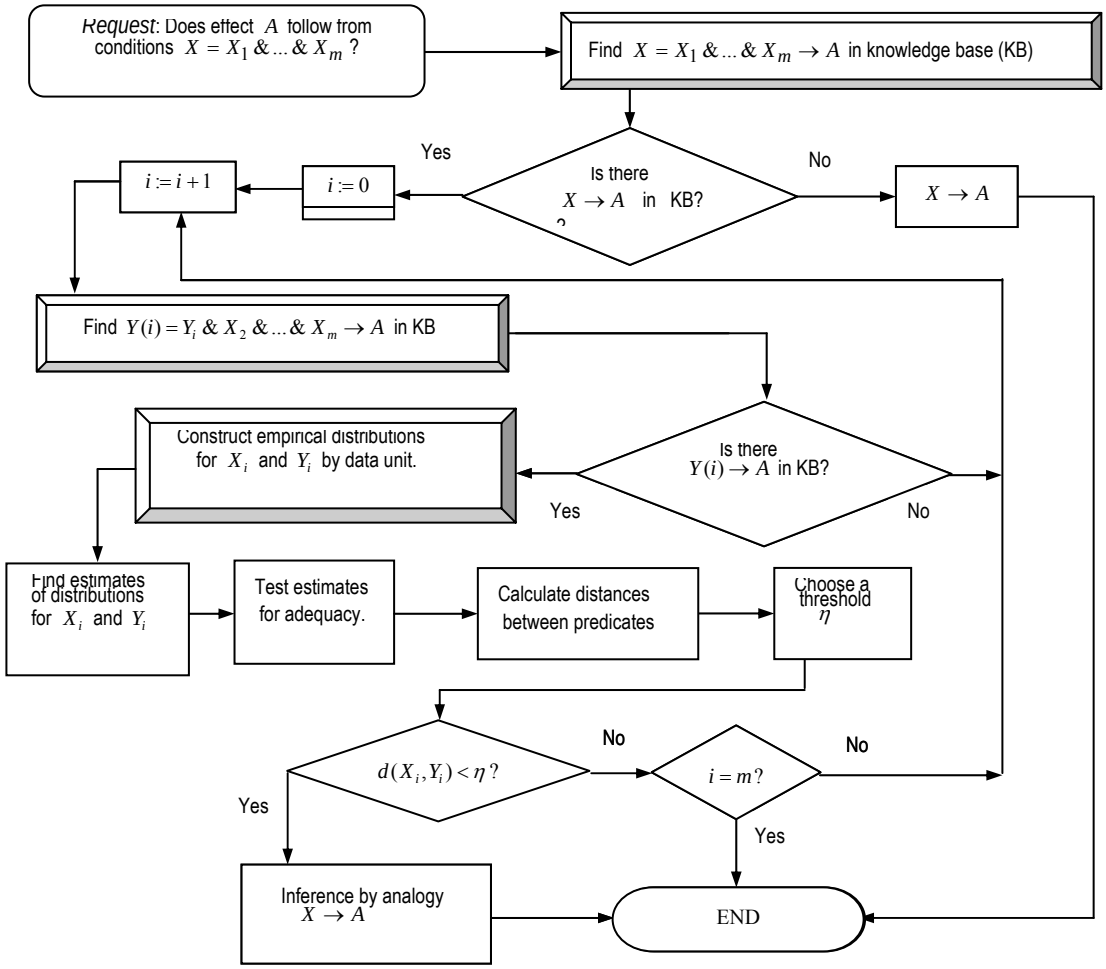



Figure 1. By-Analogy Reasoning System: A Flowchart

The data about distribution of the values of $X_i$ and $Y_i$, under which they are "true", are extracted from the database. Pursuant to these data, the empirical $F_n^*(x)$ and $G_n^*(x)$ are constructed for the values under which they are "true". In accordance with $F_n^*(x)$ and $G_n^*(x)$, the estimates of $F^*(x)$ and $G^*(x)$ are found as for $X_i$ and $Y_i$. To find such estimates, introduce a distance between the distribution functions. To derive

the estimates, find the distribution functions from the specified families that are closest to the found empirical functions in the sense of an introduced distance metrics. The obtained estimates for $F^*(x)$ and $G^*(x)$ are then tested for their adequacy as for the available empirical data by the $\chi^2$–criterion. If the estimates for $F^*(x)$ and $G^*(x)$ do not fit available empirical data, choose another family where the same estimates are sought for again. The adequate estimates of $F^*(x)$ and $G^*(x)$ are yielded, and a distance $d$ between the considered predicates is calculated by means of them. This distance determines a degree of "resemblance" or "closeness" for $X_i$ and $Y_i$. Predicates are close if a distance between them does not exceed some threshold. As a threshold, a sufficiently small positive number $\eta$ is chosen, and a value of this number states a by-analogy inference truth. Under $d(X_i, Y_i) \le \eta$, there is the following by-analogy inference: "if $X_1 \,\&\, X_2 \,\&\, \ldots \,\&\, X_m$ then $A$". If $d(X_i, Y_i) > \eta$ takes place, then a found production rule is rejected, and a new production rule is sought for that differs from a required one in a next-coming $(i+1)$-th predicate.

## Conclusion

The paper considers different-type distances between predicates. They are the distances between predicate value distribution functions under which predicates are "true". The asymptotic features of such distances and the interrelation between the latter are studied. The paper proposes the procedure used to find distributions of predicate values for the case when predicates are true. The distribution functions are found by the empirical distribution ones. The paper also deals with the mechanism that tests an adequacy of a yielded distribution function on the basis of the $\chi^2$–criterion. The predicate resemblance determination procedure is replaced by the simple procedure that determines Euclidean distances between distribution function parameters. The replacement theorem is proved for the parametric families. The proposed distances can be used in expert systems in order to construct by-analogy reasonings.

## Bibliography

[1] N.G. Zagoruyko. Application Methods for Data and Knowledge Analysis. Novosibirsk, 1999, 269p.
[2] V.N. Koval, Yu.V. Kuk. Finding Unknown Rules of an Environment by Intelligent Goal-Oriented Systems, "Information Theories and Applications", International Journal, vol. 17, N 3, p. 127-138, Sofia, 2001.
[3] G. Polya. Mathematics and Plausible Reasoning. Prinston, New Jersey. 1954, 464p.
[4] A.A. Borovkov. Mathematical Statistics. Moscow, Nauka, 1984, 472p.

## Author information

Valeriy Koval – Institute of Cybernetics, Head of Department, address: 03680, Kiev, Prospect Glushkova, 40, Ukraine; e-mail: icdepval@ln.ua
Yuriy Kuk - Institute of Cybernetics, senior scientific researcher, Ukraine; Kiev, e-mail: vkyk@svitonline.com .

# ON NEURON MECHANISMS USED TO RESOLVE MENTAL PROBLEMS OF IDENTIFICATION AND LEARNING IN SENSORIUM

*G.S. Voronkov, Z.L. Rabinovich*

*Abstract: The paper considers some possible neuron mechanisms that do not contradict biological data. They are represented in terms of the notion of an elementary sensorium discussed in the previous authors' works. Such mechanisms resolve problems of two large classes: when identification mechanisms are used and when sensory learning mechanisms are applied along with identification.*

*Keywords: sensory learning, identification, task solution, elementary sensorium.*

## Introduction

The main component of thinking quite probably consists in the problem (task) resolution. The question of how to simulate the brain function of thinking becomes more and more urgent.

The paper represents some viewpoints related to arrangement of neural mechanisms implementing two large mental problem classes, viz. how the mental identification and learning tasks are solved.

The above viewpoints follow from the general description of sensory systems and of the sensorium in all. This general description is based on the model paradigm [1]. According to this paradigm, the sensorium is the neuron model of the familiar sensory environment. The model paradigm notion is actually the projects of the ideas, mostly known and taken from various spheres of knowledge about the brain and information processes, onto the unique hierarchical neural network. This network comprises the elementary sensory system [2], the memory medium [4] and the intelligent medium [4]. It is proposed as the generalized result obtained when data about structural and functional arrangement of sensory systems are analyzed [2, 5]. The paper collects and develops the fragments of the model paradigm notion, related to thinking.

## Elementary sensorium: some basic notions

1. The long-term memory (LTM) is the very model from the neurons (the very sensorium; see the right part of Fig. 1), it reflects the potential familiar sensory environment (about unfamiliar environment, see below, # # 1 and 2). In the typical structure (TS; see the left part of Fig. 1), the complex familiar potential object is submitted as a whole by symbolic neuron and its elementary properties – by the receptors and quasi-symbolic neurons. In the sensorium, the complex familiar potential object is submitted as a whole by the neuron on the upper most synaptic level; here the different modal key properties of this object are submitted too. At the lower levels, its complex subproperties and their traits are submitted. The most elementary indecomposable properties are submitted in the very first sensory systems TS of the sensorium.

Being influencing, stimulating, the object is submitted in the model by the same neurons, but already in the excited state. The characteristics specific only to the stimulating object (force, duration, importance) are submitted in functional parameters of neurons (force and duration of excitation), the memory of them is kept for some time as a changed excitability of the formerly activated neurons and the changed conductivity of the ways to them, in other words – in the changed synaptic weight. It is the short-term memory - STM. Its mechanisms are rather investigated neuro-physiological mechanisms of the plasticity.

The presentation of unfamiliar object activates the neurons of low synaptic levels, only that of the neurons matching the familiar properties and subproperties of this unfamiliar object. They are remembered for some time in STM with the help of mechanisms described above. As it is possible to see, the storing in STM is not accompanied by formation of the new connections. The storing in LTM is the arrangement of the symbolic neurone to match the new object (as a whole), the new symbolic neuron and formation of its TS connections.

The mutual positive connections inside TS (between symbolic and quasi-symbolic neurons) provide rhythmization of activity, its quantumization, amplification and contrasting. The latter is provided by the lateral inhibition characteristic for biological neuro-networks.

The rhythmic process, in the separate TS, develops, lasts for some time and attenuates. The attenuation occurs due to the collecting recurrent inhibition. All these processes occur in the olfactory bulb, TS prototype, and the bulb computer model [6]. The descending connections in the elementary sensorium provide synchronization of rhythmic activity of appropriate TSs. The activity of the uppermost TS plays a leading role in this.

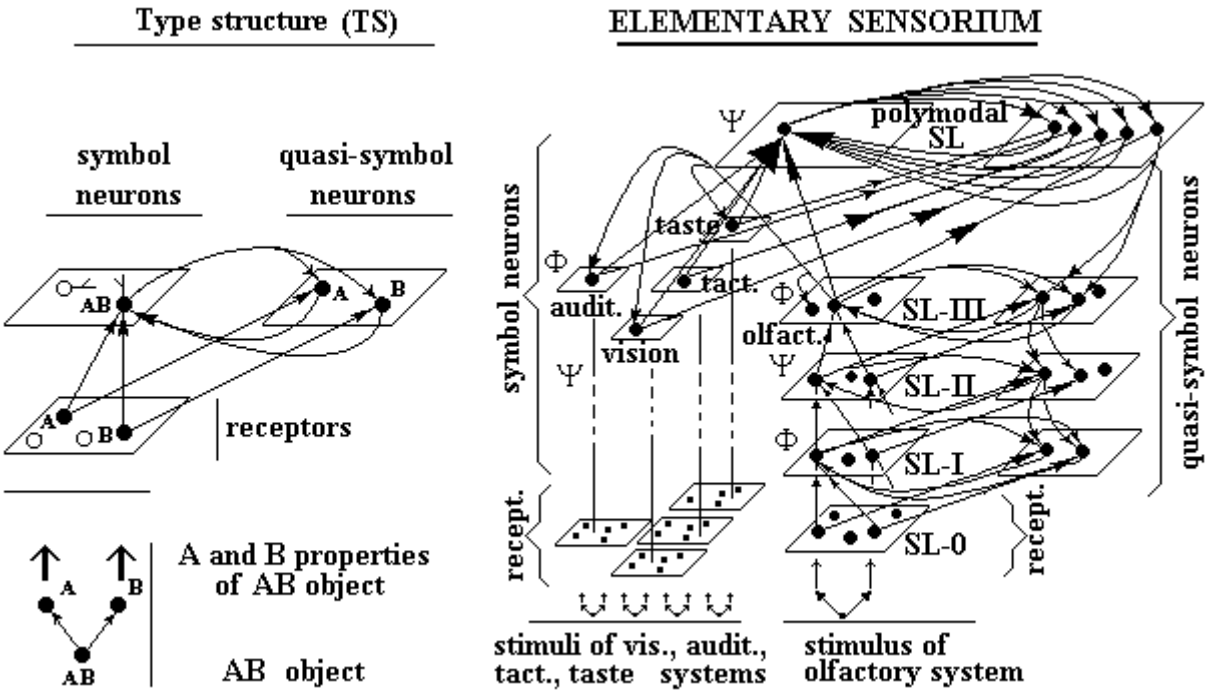



Fig. 1.

## Task and solution: general notions

2. The task here is a complex object (situation, picture, phenomenon) of the sensory environment, shown to the elementary sensorium for identification; the solution is the identification of this object. The absence of identification is absence of solution. The familiar object can be unidentified due to several circumstances: or due to it is insufficiently distinct among the environmental objects, or it is shown insufficiently full (fragmentary), or due to combination of these circumstances. We also believe, that any presentation (actualization) of the environment is "presentation with the purpose of identification" since identification is the basic function of the sensory systems. The identification process is initiated by influence of adequate stimulus on the receptors, then the process is directed by the network architecture, the neuron connections, their present excitability state and others.

The stated question can also be a part of the task condition (if the task is also shown in the verbal form), it is the important component of the task condition, in fact, the major "help" making the object distinct in the background.

Note, that the examined elementary sensorium (Fig. 1) is isolated from motivation, emotional and other systems of the brain. So, by its example, we actually analyze only the own possibilities of this simple circuit (in terms of the tasks solution) and only occasionally we take account of its communications with other brain structures. In this analysis we imagine that this circuit works by the known neuro-physiologic mechanisms. This analysis is limited by the use of only biological mechanisms and model paradigm framework.

The shown unfamiliar (new) object cannot be identified in principle, since it is not submitted in the sensorium as a whole. Identification of this object in the sensorium is possible only after formation of its neuron model (see below, # 6).
This understanding of the "task" and "task solution" is followed by classification of the tasks, being the large share of all mental tasks in all, to two classes. The first class involves the tasks, solved by the identification mechanisms only. This is the class of "identification tasks ". The second class involves the tasks, solved, by the mechanisms forming new "concepts" in addition to the identification mechanisms. The concept-forming mechanisms are similar to or are the sensory learning mechanisms. It is the class of "learning tasks". Note here, that the identification and sensory learning differ from the comparative process; the latter is not considered in the work.

3. The presented identification task, if it is not solvable at once, is submitted in the brain by the activated sensorium part. This part consists of the activated neurons, matching the background objects and their properties including the task object and its properties. However, the latter (viz. the neurons matching the task object) do not reach the condition of activity, appropriate to identification (see below # 7). Thus activated part of the sensorium is the initial situation.
To the solution of identification task there corresponds the final situation (identification state). It is a steady, for some time, special dynamic activity state of neurons matching the task object only. These activated neurons represent simultaneously both the task object as a whole and its major properties. The feature of the dynamic state consists in the fact that the activity of these neurons is amplified and rhythmical, and their rhythms are synchronized (see below, # 8). In the period, when the final situation takes place, the neuron rhythmic activity dominates the activity of others neurons of the elementary sensorium. The neuron domination is expressed in priority influence of their activity on the effector and other brain systems. In the stated representation, the domination state is correlated with the phenomenon called by the standard term "attention".

4. The process of transition from the initial situation to the final situation is the process of task solution.
By definition, the identification task assumes, that the final situation neurons are submitted in the sensorium a priori; they are the part of the neuron model (long-term memory), representing the potential environment in the brain. Thus, the process of the solution consists in selective activation of the final situation neurons. In addition to "the matching realization mechanism" in the direct sensory pathway, during selective activation, other elementary sensorium mechanisms (see below, # # 7, 8) act; the set of the selective activation mechanisms is possible to designate as mechanisms of identification.
The moment, when the final situation is reached, correlates, according to the stated representation, to the subjective emotional sensation, insight (eureka). Here, it is possible to note, that, since the problem of sensation remains a principle “blank space" in the brain problem, this subjective display of biological sensorium work, used here in interpretation of the elementary sensorium work, is only formal.
If the task is not solved (i.e. the final situation is not reached, the insight is not shown), for continuation of the solution process, it is necessary to expand the task conditions by entering the additional fragments or different-sort helps making object distinct in the background.

5. There may be the following identification tasks, for instance. 1) A game for children. It is necessary here to look for and identify a known contour within a web of lines (branches). Task statements are: a represented picture proper (e.g. branches of a bush); a question asked in some form; other prompts. The solution is vision and identification of a desired (known) contour. 2) Examination school tasks. The task solution means here to identify a task type (scheme) since the solution of each task type (description or model in the most generalized form at upper levels) is already present in the sensorium (or the linguistic system) of a person being examined on the basis of a passed program. 3) Tasks solved by analogy. As in the case of school tasks, gist of solution consists in identification of the general scheme characterizing both an initial situation

and another but the known one, i.e, it is necessary to identify a scheme already represented in the sensorium.
4) In essence, solution syllogisms (i.e. deduction process) also possibly come to identification.
It is obvious, that being similar in principal, the task solution processes in the given examples 1) -4) should differ among themselves with TS activation sequence.

6. For the learning task, the final situation cannot, in principal, be reached with the help of only the identification mechanisms since the absence of the uppermost (uniting) TS and other TSs, which potentially would represent the task object (and its properties), does not allow the solution process to reach the identification state (see # 2).
To solve this task it is necessary to add the missing TSs into the initial situation. This can be carried out only by formation of new TSs. At the first synaptic level, the new TS formation is initiated by influence of a new stimulus of on the receptors (sensory learning). To initiate TS formation at a high level, it is necessary to activate TS symbolic neurons of the previous level. The latter can proceed in two ways: through the sensory learning and/or through the influence of activating brain systems on the sensorium symbolic neuron fields. Evidently, the neuron model formation in the second way largely proceeds from the "internal resources" of the sensorium, from LTM. Note here, that formation of TS symbolic neurons in the elementary sensorium has some formal similarity with formation of new tops in pyramidal networks [7].
The model formation proper (new TSs formation) can be considered as a really creative process since it singles out a new object (or property) in sensory environment (or in its neuron model) and models this object (viz., matches the new TS symbolic neuron with the new object). In linguistic terms, the singling-out of the new object in the environment (and the formation of the new "object" on the basis of knowledge) is formation (birth) of the new concept.

## Neuron mechanisms of identification and sensory learning in the elementary sensorium

7. In the elementary sensorium (Fig. 1), matching the potential sensory environment, the initial situation is represented by the activated neurons of typical structures (TS), paired with the presented task (see # 3).
Thus if any of the key stimulus appears presented," the matching realization mechanism" of the sensory direct pathway selectively activates the symbolic neuron of the uppermost TS (the neuron input is organized as disjunction). This neuron matches the task object as a whole.
It activates the matching quasi-symbolic neurons. The latter match the single key properties of the task object. By descending projections, these quasi-symbolic neurons activate the lower TS quasi-symbolic neurons matching the subproperties of the key properties.
The neuron activation of all lower TSs can be facilitated by the activities of the ascending pathways. By associative (signification) projections, the TS symbolic neurons of the initial situation sub-activate (or activate) the symbolic neurons of other TSs.These neurons play the role of the context. As a result, the quantity of the activated neurons (and TSs) involved increases. The quantity of the involved TSs determines the depth and detail of identification.

8. Every TS is rhythmically active because of an arrangement of neuron bonds in TSs (Fig. 1, [2], [5]). *Initial situation* TSs are not yet united by the uppermost TS and, therefore, their operation is autonomous to a significant extent. Their rhythms may also be not synchronized at this moment. If the uppermost TS is activated, it synchronizes an operation of every final situation TS according to the above bonds. The synchronization in neuron rhythmic activity of all TSs is reached for some time period. And here is the “completed pyramid” making up and corresponds the final situation, viz. task solution. This period of the most intensive neuron rhythmic activity of the completed pyramid runs along with the highest neuron inhibition of all the neighbouring TSs not matching the final situation. The above activity domination is thus provided. The above inhibition occurs due to lateral inhibition bonds. This period runs for some time. It is limited by a recurrent summed inhibition accumulated inside the TS (see # 3). Therefore, the *task solution* is the synchronized rhythmic neuron activity of the complete TSs pyramid, stable for some time. The synchronized rhythmic activity is, so to say, the dynamic attractor where the rhythmic activity of the pyramid vertex TS acts as the order parameter. According to the above idea, this very state is correlated with sensation of

identification (understanding) of the complete picture as whole and simultaneously with modal perception (vision, hearing) of the picture details.

9. What is a possible mechanism of new TSs generation (their symbol and quasi-symbol neurons, bonds between these neurons and their bonds with other TSs) when the learning task is being solved? One can make the assumptions close to the truth so far, proceeding, in particular, from the following. 1). The data, derived from the sensory learning in ontogenesis, show that these mechanisms really exist. This can be exemplified by generation of detectors of vertical lines for kittens grown up in the environment with vertical lines and the absence of the detectors when kittens are grown up in the "horizontal" environment (Hubel, Wisel, 1962). 2). So called stem cells are shown in nervous system. They are generated into new neurons. 3). The constant "searching" growth of the neuron processes is illustrated. The growth is terminated when a contact with a target has been established. 4). Some hypotheses about the targeted growth mechanisms (e.g., it may be the chemical affinity principle) and about the selective bond formation principles (the Hebb principle) are advanced. 5). A host of neuron mechanisms have been studied that provide the selectivity of neurons and plasticity of the neuron inputs. 6). The targeted axon projections, including convergence, on neurons of sequence synapse levels is shown.

10. Consider the hypothetical elementary process of generation of the very first sensorium TS (Fig. 1). A set of various-type receptors is formed according to the genetic program. The sensory learning for an object supposes a reiterated activation of a certain set of different-type receptors. This activation initiates the growth of axons to the next level neurons. The first contact of any axon out of a group of growing axons with one of the neurons makes this neuron a target for other simultaneously growing axons due to the Hebb principle, for instance. This is a possible way of the symbol neuron generation. The symbol neuron corresponds to a joint array of certain properties when a neuron input is generated according to conjunction or to a class of certain properties when the input is generated according to disjunction. Possibly, the targeted 1:1-projection of receptors onto quasi-symbol neurons is provided by the chemical affinity factor. The generation of positive mutual communications between a symbol neuron and corresponding quasi-symbol neurons may likewise be based either on chemical affinity or on the Hebb principle. Supposedly, the latter is also the basis for generation of significative communications between any symbol neuron and other particular symbol neurons. The same principles may also provide the generation of descending bonds of quasi-symbol neurons. Since the symbol neuron fields act as receptor fields for TSs of the next synaptic levels, the above process can provide the generation of the second-, third- and higher-level TSs. The TSs generation process is initiated by activation stimuli: the sensory stimuli occur at sensory learning, and, the stimuli from activation brain systems occur at learning based on remembering. The process is directed by genetically provided growth, the target selection mechanisms and by the network architecture.

It is clear that the new TS generation in the neuron model, proceeding from only inner sensorium resources, calls for additional mechanisms to involved, in particular, the short memory mechanisms for identified fragments of the initial situation as well as their periodic recurrence (withdrawal from the memory). Likewise, the tasks may differ by the significance of the initial situation fragments: the higher the level of TSs, being the fragments of the initial situation, and the more fragments are involved; the greater is the possibility for the sensorium to solve the problem. This argument is true for the problems of both classes.

It is evident from the above description that the learning problem solution process needs more energy and more time than the solution of identification problems.

## Conclusion: the role of tutor and language system in identification and sensory learning

11. The described way, where the concepts are derived in the elementary sensorium, is possibly basic providing the emergence of the concepts in fauna evolution and in the creative human thinking. Likewise, the sensorium concepts in ontogenesis in animals are generated more efficiently, namely, via learning by tutor. In fact, the tutor-involving method consists of the sensory learning supplemented with the important prompt into the initial situation. The prompt helps single out the object as a whole from the background. When higher animals bring up their progeny, the behaviour of mature animals acts as the prompts. Humans also have this

element, but prompts are implemented largely verbally through the language system associated with the sensorium [2,8]. The language system matures in ontogenesis faster than the sensorium. Thus, in addition to the genetic factor, the volume of the formed sensorium (LTM) is incomparably larger than that in animals; the human sensorium is actually of another quality. It may well be, the role of the language system was as much important in development of the sensorium (genetically fixed) due to their communication function in the unusually rapid evolution of humans.

Likewise, there occurred the evolution of the language system proper. In addition to the communication and simulation functions, the intelligent function (task solution) was also developed in it. There is a ground to assume that the above description fragments of thinking neuron mechanism in the sensorium are also applicable for the language system.

In addition, the interaction between the sensorium and the language system is a new factor in thinking, including the creative one, and adds new additional mechanisms there. This makes the qualitative level of thinking even higher.

It is evident that, the working-up of the aspects, in addition to the whole number of other ones not considered in this paper can make an essential contribution into the development of the model paradigm notion about thinking mechanisms.

## Bibliography

[1] Voronkov G.S. Brain and Information. Problems of Neuro-Cybernetics - Proc. Jubilee International Conference on Neuro-Cybernetics, Rostov-on-Don, 2002, vol. 1, p. 15-18. (In Russian);
http://www.biolog.ru/vnd

[2] Voronkov G.S. Sensory System - A neuron Semiotic Model of an Adequate Environment. In: Comparative Physiology of a Higher Nervous Activity of Men and Animals. Moscow, Nauka, 1990, p. 9-21. (In Russian)

[3] Voronkov G.S., Rabinovich Z.L. A Natural Medium of Memory and Thinking and Its Model Representation. KDS-2001 - Knowledge-Dialog-Solution: Proc. Int. Scientific and Practical Conference, St. Petersburg, 2001, vol. 1, p. 110-115. (In Russian)

[4] Voronkov G.S., Chechkin A.V. Neuron Semiotic Systems as Intelligent Media. KII-96 - Artificial Intelligence - 96: Proc. 5th National Conference with International Participation, Kazan, 1996, vol. 1, p. 26-35. (In Russian)

[5] Voronkov G.S. Analyzing Principles of Coding in Sensory Systems // Biological Sciences, Dep. VINITI, No. 2144 - V88, 1998. 39p. (In Russian)

[6] Voronkov G.S., Izotov V.A. A computer Model for Neuron Mechanisms of Information Processing at the First Synaptic Level of Olfactory System // Intellectual Systems, 1998, vol. 3, No. 1-2, p. 87-108 (In Russian)

[7] Gladun V.P. Partnership with Computer. Kiev, Port-Royal, 2000, 119p. (In Russian)

[8] Voronkov G.S., Rabinovich Z.L. Sensory and Language Systems as two Forms of Knowledge Representation // News of Artificial Intelligence, 1993, No. 2, p. 116-124. (In Russian).

## Author information

Voronkov Gennadi Sergeevich – The M.V. Lomonosov Moscow State University, Leninskie Gory, Moscow, 119899, Russia; e-mail: gsv@gvoronkov.home.bio.msu.ru

Rabinovich Zinovi Lvovich – V.M. Glushkov Institute of Cybernetics, Prospect Acad. Glushkova, 40 03680, Kiev-187, Ukraine; e-mail: eco@public.icyb.kiev.ua

# TOWARDS THE SEMIOTICS OF NOOSPHERE

*V.Lozovskiy*

*Abstract: Civilization has brought us into the noosphere world. Besides physical, around (and inside of) us exist and function also mental and cultural entities. It is impossible to perform now knowledge acquisition, knowledge base creation and organizational systems management without adequate consideration of object's noosphere statuses. I tried here to clarify basic viewpoints concerning this issue, hoping that elaboration of common methodological foundations of semiotic modeling will be useful for developers and also for users of new generation automation systems.*

*Keywords: noosphere, antroposphere, semiotics, modeling.*

> "We are faced with a harmonized collectivity of consciousnesses to a sort of superconciousness. The earth not only becoming covered by myriads of grains of thought, but becoming enclosed in a single thinking envelope, a single unanimous reflection."
>
> Père Pierre Teilhard de Chardin, 1881-1955

## Introduction

Organizational systems are extremely difficult to manage. As a rule, we have almost no adequate mathematical models to apply. In this situation step forth ergonomics, naturalness of models from the human viewpoint, easiness of support and permanent modification. It becomes self evident, that one should resort to the methods of applied semiotics, which draw together human conceptualizations and computer implementations. From the other side, we should carefully reconsider the statuses of objects participating in organizational systems, taking into consideration, that we are dealing with organizational systems management in the World, that has entered the noosphere era.

The material, presented here, reflects our initial advances in this field. That is why much attention is paid to philosophical and methodological issues, to formulation of basic definitions. We have to acquire and elaborate the new conceptual system and develop the new language for the new types of problems.

## Noosphere Approach to Organizational Systems Management

Intrinsic limitations of rigorous mathematical models for «soft» problem domains stems from fundamental discrepancies between formal languages and real expert knowledge of application specialists in these domains. This knowledge is extremely anthropomorphic: indefiniteness, vagueness, uncertainty, ambiguity, inaccuracy, abundance of qualitative and linguistic descriptions and appraisals (large, substantial, dangerous, promising, sustainable, sound, «has a negative effect on…», «undesirable consequences», etc.). Determining role play so called «human elements»: subjectivity, emotionality, tiredness, illogicality, laziness, intricate dynamic structure of interpersonal relations – sympathies, antipathies, trust, prejudices, offence, gratitude, revenge, family relations, sense of justice, duty, humor, «sense of deep internal satisfaction», team spirit, presence of own goal creation and goal achievement mechanisms. Existing modeling methods almost completely ignore these human elements.

The valid scientific approach to the problem of management and control in such problem domains, requires, first of all, evolving the paradigm of real world knowledge representation, together with methods for its handling and manipulation: acquisition, integration, verification, information access. The world had entered the Information Society era, but we still are in acute need of languages, adequate to this problem. Serious steps in this direction were done by V.I.Vernadsky [Vernadsky, 1943] and French geologist/paleontologist Jesuit theologian and philosopher Père Pierre Teilhard de Chardin [Teilhard, 1947]. They proposed to use the term «noosphere» (Ionian Greek "noos" = mind) for integrative designation of physical world realities together with the whole mankind knowledge. Teilhard advanced the notion of «Omega point» - the ultimate stage of the evolution on the road of cultural and knowledge integration development. But, at that time, were absent computers, information technologies, applied semiotics, knowledge representation and engineering.

## The Noosphere Doctrine from the Positions of Epistemology

The noosphere genesis very metaphorically and romantically was described by M.Prishvin (Prishvin, 1957]: «Somewhere, on the invisible sky of the whole mankind, are wandering the great thoughts, accrued during all centuries, which passed by; they cast shadows, as clouds, and looking on them, the most sensitive humans comprehend the thoughts».

The history of the Earth and, generally speaking of the whole Universe comprises several stages. Classifications, given by different researchers, closely resemble one another. As for example, one can mention [Turchin, 1977] and [Korsuns'kyy, 2000]. For our needs, it seems reasonable to fix the following five stages:

*cosmogenesis* – processes of Universe transformations, which have lead to our Galaxy, solar system, and Earth creation;

*geogenesis* – inorganic processes in solid, liquid and gaseous media;

*biogenesis* – uprising of metabolic processes and life on Earth, humankind in its initial stages;

*sociogenesis (civilization)* – further development of humankind activity: natural resources development, intentional or unintentional interference with geological and biological processes, environmental state and processes; in the course of time civilization was increasing its influence on the state and transformation of the World;

*noogenesis* – the epoch, where the total mass of humankind World knowledge explosively increases, anthropomorphic influence on environment becomes heavy on planetary scale, and human mental power could be used to fundamentally change behavior patterns, taking into consideration all planetary factors, aiming at thorough harmonization of civilization development on the whole Earth and beyond.

This list gives the ordering of the starting points for the corresponding stages. They all are still active now and keep influencing one another.

Let us give more detailed definition of the concept: «*noogenesis*». It is the evolutionary stage of World development, when certain conditions– «*nooconditions*» - determining the transition from *homo sapiens* to «*intelligent society*» hold.

Scientific, technological, social, ethical, moral and cultural potential of the society achieve such crucial level, at which the following conditions become feasible.

As a most general goal for humankind is adopted freedom of pursuit for happiness for each live being as far, and to such extent, which does not interfere with the interests of other live beings.

As the basic ethical principle is adopted the reciprocity principle: treat others as you would like to be treated by them.

*The Global distributed knowledge and data Base (GB) is formed, which explicate knowledge of humankind in all spheres of activity: theories, applications, factual information, skills, arts, beliefs, know-hows, etc. – in any modality: mass media, printed matter, computer data, words of mouth... Computer modality is strongly preferred (ease of creation, support and dissemination).*

Advanced information telecommunication infrastructure is created, which supports efficient and convenient communication between society members and access to the GB for its creation, support and use for decisions support in all spheres and levels of societal activities.

Juridical, social, administrative, educative and political societal functions are oriented towards the most efficient fulfillment of all previously formulated nooconditions.

Fulfillment of condition 5, due the explosive progress of Internet technologies, may be considered as being solved, in principle. At the same time, preparedness of society to implementation of condition 4 is restrained by difficulties of public thinking paradigm change, cultural orientation, abundance of white spots in knowledge integration methods, methods of modeling for «soft» problem domains – to which systems of humanitarian knowledge and management in organizational systems belong.

## Applied Semiotics as a Paradigm for Information-Noosphere Models

Semiotic approach to the problem of modeling preserves high correlation level between the problem domain and its model. It helps considerably on the stages of model creation and support during its whole life cycle.

*Semiotics* – is the science, studying the structure, properties and dynamics of formal symbolic systems in their relation to physical and cultural world – from the position of world cognizing system, development and use of symbolic models of reality, behavior of such models, rules of interpretation and manipulation.

Semiotics stems from works of ancient philosophers. More detailed presentation of its basic concepts was done by G.Frege [Frege, 1892]. The notion of signification was introduced by A.Church [Church, 1956]. Fundamental investigation of the basic semiotic triad – things, properties and relations – was done by A.I.Uyomov [Uyomov, 1963]. E.F.Skorohod'ko [Skorohod'ko, 1962] was among the first, who proposed to build computer models using relations – «rx-codes». This approach was further elaborated in works on situation management and control [Pospelov, 1975, 1986]. Further development of applied semiotics approach for semantic nets and relational knowledge representation systems were presented in [Lozovskiy, 1979 – 1999].

*Applied semiotics* differs from its philosophical sibling in its object and approach: here are studied not just pure formal, but quite real symbolic models, which are implemented by means of computers for modeling and control of real applied objects and systems.

Computer semiotic models are marked with intrinsic semantic wealth and depth, it gives the possibility to create on this basis procedures for assimilation of new knowledge, checking it for consistency and completeness, create plans of goal-directed behavior, support behavior monitoring and control, diagnostics, decision support. Applied specialist can interact with such model in terms of customary notions: actions, states, goals, scripts, procedures, functions. Compiled down to low computer level program solutions are devoid of these abilities. Semantics on these levels is already lost, one can have only preprogrammed scenarios and receipts– without dynamics and the ability to analyze and explain unforeseen situations. Let us formulate the basic definitions.

*Universum (Universe)* – the set of all entities (objects) in the Universe (noosphere).

*Entity, or object* – something, cut out from Universe by observer, guided by certain pragmatic considerations; at which their attention is directed, about which they speak or think, which is referenced somehow or other. For example, «Black Sea», «is to the West of…» , «Christianity», «gravitation», «socialist emulation», «sustained development of the regional economy» …

Any research, analysis, modeling presumes cutting out of the Universe certain restricted domain. Of course, this action is very informal, specific and depends upon the problem to be solved, viewpoint of the researcher and on many other circumstances. We will call it *problem domain* (PD). Usually, PD includes the object of investigation, modeling and/or control per se, and also, its environment. Their interaction and interference should be considered in complex.

One more delicate issue concerns the interrelation between the model and the object being modeled. Should the model be included in PD? And, at last, what to do with researcher, who studies the given PD, builds the model, produces certain conclusions on this base, and then uses the knowledge obtained in actual work with the prototype object? The model, in its turn, can have direct links with the object of modeling, receiving from it raw measurement data. Results of modeling, in their turn, can directly or indirectly influence the object of modeling, researcher or end users, their administration.

*Problem domain* (PD) thus is the set of all entities (objects), which have material effect on the problem of analysis, modeling or control being solved.

Giving so broad definition to the PD notion, I am aware, that it goes against the classic philosophical canons. Within the sphere of our consideration become included, together with «material», also «ideal» entities. From the viewpoint of «pure» science, it is the deadly sin, conceptual farrago… I have only two answers.

Firstly, life encourages us to adopt the noosphere approach to management and control of organizational systems (OS) in contemporary world. As in this world practically on equal terms act material and «ideal» objects, our picture of world should adequately reflect it.

Secondly, what was beyond the power of the classical philosophy becomes feasible today, on the basis of semiotics and knowledge representation researches. One needs to change the paradigm, and we have the means to do it.

PD encompasses the set of entities – *d-entities* – domain entities. The same meaning I assign to the term «object» - something at which our attention is focused, which is under consideration. It should be emphasized, that «entity» - could be really *everything*, even something ephemeral, which participates in our picture of the world – not only material things, physical objects.

From the viewpoint of noosphere approach, in our interpretation, d-entities can have one of the three possible noosphere statuses.

*Physical entities* (P-entities) – objects, which exist or undoubtedly existed in physical world, in «reality». The issue of «physical existence» - is the ancient philosophical stumbling block, one of the main battlefields between materialism and idealism. In spite of many fallacious, disputable and boundary situations (phlogiston, UFO, telekinesis, God Almighty), we are forced to use this loose definition, because it nevertheless brings us closer to reality and can play the role of common ground in PD languages development and usage for academicians and application specialists. It is quite discomfortingly, really, but we should eat this crow, accepting the fact, that this issue cannot be completely formalized. Simply, we should only provide for the facilities to modify the knowledge base correspondingly, if cognitive positions of knowledge engineers or application specialists somehow change. It is the routine general requirement for lifetime support for all bases, and this case is nothing more restrictive or outstanding than others.

*Mental entities* (M-entities) – objects (thoughts, ideas, representations), which are formed within the thinking systems of Intellectual Subjects (IS), and used by them in the processes of cognition, analysis, modeling, forecasting, planning, control.

We can think about IS as of biological subjects (humans, animals), or as of artifacts – robots, expert, decision support, artificial intelligence systems. The physical nature of IS and of their internal processes are of no relevance. Important are their cognitive – semiotic and functional properties, joint interactive dynamics of m-entities of the given subject, interaction with the outer world, information transformation processes.

Besides images of specific p-entitites (Ukraine, A.S.Pushkin, Duke's Monument in Odessa), in the brain of IS can exist pure abstract entities (mathematical theories, concept of dwelling, «what is good», etc.) and also fictional, folklore, mythological, religious conceptions, which, in principle, have no p-prototypes.

We arrive at the conclusion, that m-objects are quite specific entities, which dwell in unique media – thinking brain of IS. From one viewpoint, they are undoubtedly real, objective, because they are formed within the material media and with the help of material absolutely real biological, chemical, electric processes, have material carriers, but, from the other viewpoint, they are subjective, belong to specific individual, depend upon their ability to interprete them, obtain their *meaning*. They are extremely hard to be explicated by individuals themselves, in the form of introspection, and even more harder – by collocutor, psychoanalyst.

*Intellectual subject* (IS) – is the subject, which:

- possesses the unique ability: to build the model of him/her/its-self and of corresponding environment; this model should adequately, *objectively reflect* essential properties of pertinent p-objects;
- possesses the ability to act purposefully;
- can build and modify its own knowledge representation system, including abstract m-entities, use axiomatic method, definitions, methods of deduction, induction, abduction, algorithms;
- possesses learning capability, ability to plan own activity, implement plans, can actively interact with environment in order to obtain information needed and implement own goal directed activity.

*Objectivity of reflection* is settled a posteriori – depending on the results of goal directed activity of the given IS: if it is successful, one may assert that their model was objectively reflecting reality.

Civilization, integration of knowledge and skills, communication between IS, upbringing and education of young people would be impossible, if results of individual intellectual achievements would remain on the level of separate individuals' m-objects. With necessity, we arrive at the conclusion about existence of the third type of entities – cultural objects (c-entities).

*Cultural entities* (C-entities) – are the objects of *«cultural world»*, created by evolution and civilization within the frame of world human community – on the most general scale: sciences, arts, customs, religions, rituals, laws, regulations, plans, purely «human» relations and feelings (fear, hatred, love, trust, admiration, amazement, irony, etc.).

C-entities are far not simple objects - they are centaurs, integrating objective and subjective features. They are sort of objectified analogs of m-entities – as far as they are alienated from individual subjects. C-entities do not exist «in nature». They «materialize» only during the process of interpretation by some IS.

Distinction between m- and c-entities can be elucidated as follows. Conception of the «War and Peace» conceived by Leo Tolstoy, - is the system of m-entities. When it was written, printed, distributed read and understood by different people, we can say, that the corresponding c-entity sprang into existence. It can be an object of discussions; there appear more or less canonical interpretations of its characters, one can use them metaphorically, and such metaphors will be perceived by educated people in similar way.

Sometimes arise the temptation to abandon cultural gnoseological category at all, and handle c-objects as ordinary p-objects, with which they really have much in common. Frequently information on certain c-object is much more complete and systematized, than on some p-object. Thus, Cinderella or Red Hat are much more familiar and similarly understood by many people, than, say, hydrogeologic characteristics of North-West Black Sea shelf.

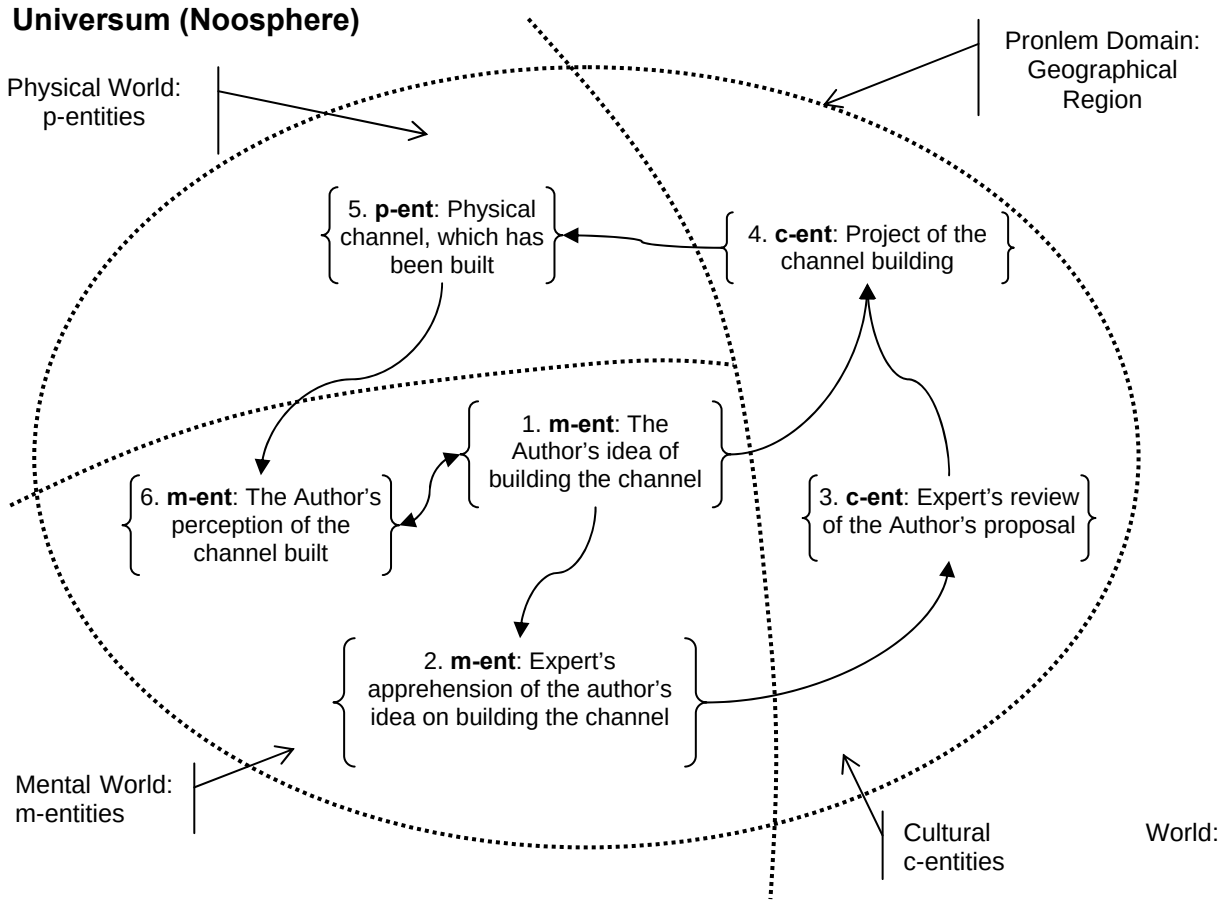



Fig. 1 An example of noosphere semiotic structure of a problem domain

Some c-objects have p-objects as their material correlates. Other c-objects go without material correlates at all – verbal folklore, «common laws». To exemplify «reality» of c-objects let us consider juristic term: «common-law marriage» - typical c-entity - a marriage without a civil or ecclesiastical ceremony, generally resulting from an agreement to marry followed by the couple's living together as husband and wife, performing joint housekeeping. If the court acknowledges two people being in the state of common-law marriage, certain cultural attitudes, rights and obligations arise between them and towards their community property, children, etc., leading to corresponding physical attitudes and actions with certain p-objects.

IS interact with physical environment in two ways: directly (your car bumped – God forbid! – into the road post; in this case neither you, nor arrived police officer have any doubts about reality of the car, road post, damages) and cognitively – through the reflection, modelling in the IS's brain. In this case, m-objects formed in participants' and witnesses' knowledge representation systems, of course, are different. Creation of the accident report – c-entity - is thus aimed at working out single official viewpoint on this accident. But even after it is completed and even signed by participants, one cannot be sure that the synthesis of unique, adequate to reality, c-object has taken place. Here you are! We have arrived at idealistic, in fact solipsistic, interpretation: for each human participant of the current incident «objectively» exist only their own (subjective!) feelings and mental constructs, m-objects… Of course, it is only until the guilty person will have to pay for the property damage and car repair with quite real, «material» money (p-object with its corresponding c-object, as a correlate, by the way!).

The border between physical and cultural world appears, first of all, as a result of natural distinction between p-entities and their cultural correlates, interpretations, which create World cognizing systems.

Let us take the c-object: «Project of the channel building» (4, Fig. 1). The corresponding p-object – the set of printed lists – is omitted, for simplicity, from the figure. With c-object (4) is associated the certain semantic and pragmatic interpretation. Information, acquired by specific person on making acquaintance with this object, differs with professional orientation, theoretical baggage and personal experience, their goals, accompanying circumstances, relations with authorities, adjoining organizations, etc. Complexity and ambiguity in cultural objects' evaluation – is among the main difficulties in creation of management and control systems for OS.

Sometimes, m-objects can play the role of p-objects – entities beyond the comprehension of specific World cognizing system. Thus, lecturer's efforts are aimed at creation within the student's minds some specific system of m-objects – knowledge and skills in the PD of interest. The following examinations are aimed at explication of «objective» c-situation in students' understanding. This tendency towards m-objects' explication – is the task permanently being solved by humanity. Science, its cutting edge, is done predominantly in the sphere of subjective, m-objects. In the course of time, the definite paradigm is created, then – specific scientific school. Along with this, primary subjective conceptions (m-concepts), are becoming the common property of given scientific community, supported by published works, information exchange on conferences, implemented applications, start to acquire more and more objectivity… At this stage, we can say, corresponding c-objects flourish, becoming explicated from individual m-objects in the process of human thinking and communication. And at last, the new theories, well developed branch of science find their way into textbooks and are presented to the broad students' communities, obtaining a kind of objectivity. This «a kind of» is becoming evident when different scientific schools, evolving different paradigms, create different empirical theories. Some c-objects can, at last, obtain their physical, «real» p-incarnations.

Let us illustrate these considerations. Look at the schematic example – fragment of noosphere – semiotic object structure for chosen PD (Fig. 1). PD is a part of the Universe. Before IS appearance and noosphere emergence the Universe comprised exclusively p-objects of natural origin (the Earth, gravitation, electromagnetic waves, chemical elements, minerals, vegetable and animal kingdoms, etc.). IS perceive all these with their organs of sense, measuring instruments, supplementing the picture being constructed with their own conceptualizations, considerations, hypotheses, logical inferences.

This argumentation supports materialistic hypothesis on primariness of substance. M-entities' existence is possible only in the conscience of thinking subjects, their material bearers, which should be extant *before*, the first thought could arise. Thought, «word» without material substratum – is senseless concept. It is impossible to create a painting without brush, paint, canvas. C-entities formation in certain orders takes place on the basis of corresponding subjective m-entities. Reverse process also takes place: transformation of c-entities into m-entities of concrete individuals during their upbringing and education. These processes are inseparable from appearance of natural languages, which play the dual role – of thought and communication.

Let in the brain of some subject – the author – appeared the idea of building some channel: (1), Fig. 1. It is subjective m-entity, available only to its author. Then some other individual – expert – makes acquaintance with it thus creating in their brain the corresponding m-entity (2). Just like the source author's plot, it is completely subjective notion: we have no guarantees of its adequacy to the subject of investigation. Its formation is liable to be influenced by abundance of uncontrolled and implicit factors (educational qualification, practical experience, acquaintedness with the field, their pertaining to some or other professional group, paradigm adhered to, etc.).

The expert's work results in creation of c-entity (3) – review of the author's proposal on channel building. From this moment on – expert's opinion becomes available to other humans – the transition of given entity from subjective mental category into cultural – objectivization, or explication – has taken place.

Assume that the idea of channel creation was supported by authorities, and, as a result, the new c-object appears: «Project of the channel building» (4). It comes as a result of many people efforts, analysis of many documents, reference manuals, etc.

Let then the designed channel is physically built. It is designated in our scheme with the p-entity (5). The author of the initial idea, making acquaintance with its implementation in reality, can form in their conscience the corresponding notion (6).

What have we arrived at as a result of this consideration?

Firstly. We should decidedly agree, that the three worlds allocated by us: physical, mental and cultural are closely adjoined and are interrelated.

Secondly. We should become convinced, that goal directed activity within the frame of contemporary human society should be considered from the positions supporting natural trinity of designated noosphere semiotic

categories. This conclusion falls into certain contradiction with «pure» scientific approach, warning us against mixing «material» and «ideal», building unpenetrable watershed between these worlds.

Thirdly. Creating modern computer expert and control systems, pretending to be highly adequate to reality and be competent on decision support level, one should be certain to deal correctly with objects of all three worlds. Particularly, if traditionally management and control sciences consider one stage reflection of physical world realities by computer knowledge representation system, the more adequate noosphere approach to the process of IS interaction with environment and especially with other IS leads to importance of taking into consideration the phenomenon of m-entities reflection. Roughly speaking – while interacting with intellectual beings, we should take into account, how do they perceive our ideas and representations. Faults, commited already at this – first – reflection level may lead to unwanted results. For successful performance of communicatory acts in the vein of goal directed activity, due interference of human elements, one should take into account not less than two, and sometimes even more levels of reflection [Lefebvre, 1973].

In the process of evolution, the role of human elements was increasing, and nowadays, especially when we deal with organizational systems management, cannot be ignored. That is why one should adequately take into account peculiarities of «ideal» - mental and cultural objects, efficiently coordinating these processes and activities dealing with p-entities. And all this – under the pressure of complicating factors: incomplete, invalid, unreliable, imprecise information, errors of all possible kinds, dealing with incompletely observable and incompletely controllable systems.

M- and c-objects both can possess different level of objectivity, reflection precision, modeling the properties of prototype d-object and are situated at different points on abstraction axis, originating in PD. Its other extremity corresponds to pure abstract objects. Role of the prototype can be played by any (p-, m- or c-) object of the PD.

We arrive at the conclusion that in any organizational control and management system, and also in any knowledge representation system, one should accurately discern noosphere status of entities, with which we deal.

## Conclusion

Automation of management and control in organizational systems required revision and more precise consideration of semiotic definitions and approaches to problem domain model building in the frame of noosphere representations. The current research was aimed at elicitation of noosphere statuses for the entities which participate in the domain. It was proposed to consider three problem subdomains: physical, mental and cultural spheres followed by explicit consideration of these statuses in semiotic models. The subsequent research will be concentrated on developing the methods of goal directed activity in organizational systems management representations basing on explicit accounting of noosphere statuses of participating entities.

## References

[Vernadsky, 1943] Владимир Иванович Вернадский, Несколько слов о ноосфере, 1943-1944, Из Архива В. И. Вернадского: http://vernadsky.lib.ru/ .

[Prishvin, 1957] М.Пришвин, Собрание сочинений, т. 5, М., 1957, 683 с.

[Teilhard, 1947] Pierre Teilhard de Chardin, The Formation of the Noosphere, 1947, http://technoetic.com/noosphere/

[Kuhn, 1962] Thomas Kuhn, The Structure of Scientific Revolutions, University of Chicago Press, 1962

[Turchin, 1977] Valentin F.Turchin, The Phenomenon of Science. A Cybernetic Approach to Human Evolution, NY, Columbia University Press, ISBN 0-231-03983-2, 1977

[Korsuns'kyy, 2000], Володимир Корсунський, Засади сучасного наукового світогляду, Київ, 2000

[Uyomov, 1963] А.И.Уемов, Вещи, свойства и отношения, АН ССР, Институт философии, Издательство АН ССР, М., 1963, 184 с.

[Frege, 1892] Gottlob Frege, Über Sinn und Bedeutung, «Zeitschrift für Philosophie und philosophische Kritik» No. 100, 1892, pp. 25-50

[Church, 1956] Alonzo Church, Introduction to Mathematical Logic, Vol. 1, Princeton University Press, 1956

[Skorohod'ko, 1962] Э.Ф.Скороходько, Информационный язык для технических наук, «Математическая и структурная лингвистика», № 1, Киев, ИК АН УССР, 1962, 50 с.

[Pospelov, 1975] Д.А.Поспелов, Большие системы. Ситуационное управление, М., «Знание», 1975, 64 с.

[Pospelov, 1986] Д.А.Поспелов, Ситуационное управление, Теория и практика, Сер.: Проблемы искусственного интеллекта, «Наука», ГРФМН, М., 1986, 284 с.

[Lozovskiy, 1979] В.С.Лозовский, Ситуационная и дефиниторная семантика системы представления знаний, "Кибернетика", No. 2, 1979, стр. 98 – 101

[Lozovskiy, 1990] В.С.Лозовский, Сетевые модели, разд. 1.3 в кн.: Искусственный интеллект, в 3-х кн., Кн. 2: Модели и методы. Справочник, п/р Д.А.Поспелова, М., "Радио и связь", 1990, стр. 28 - 49

[Lozovskiy, 1998] Lozovskiy Vitaliy (UA): Common Sense Semiotics, Conference Proceedings: Knowledge-Based Software Engineering (Smolenice, Slovakia), P. Navrat and H. Ueno (Eds.), IOS Press, Amsterdam, Berlin, Oxford, Tokyo, Washington, DC, ISSN: 0922-6389, ISBN: 90 5199 417 6 (IOS Press), 1998, pp.232-240

[Lozovskiy, 1999] Vitaliy Lozovskiy, On the Road to Parasemiotics, ASC/IC'99 - Труды 4-го международного семинара по прикладной семиотике, семиотическому и интеллектуальному управлению, Институт программных систем РАН, Российский университет дружбы народов, Российская ассоциация искусственного интеллекта, Москва, октябрь 1999, ISBN5-89574-064-2, с. 158-166

[Lefebvre, 1973] В.А.Лефевр, Конфликтующие структуры, «Советское Радио», М., 1973.

## Author information

Vitaliy Lozovskiy – Institute of Market Problems, Economic and Ecological Researches, National Acad. Sci. Ukraine, 29, Francuzskiy Boulevard, Odessa, 65044, Ukraine, Sen. Sci. Res,; e-mail: loz@loz.intes.odessa.ua

# PLANNING TECHNOLOGIES FOR THE WEB ENVIRONMENT: PERSPECTIVES AND RESEARCH ISSUES

*A. Milani, S. Marcugini*

***Abstract**: This work will explore and motivate perspectives and research issues related with the applications of automated planning technologies in order to support innovative web applications. The target for the technology transfer, i.e. the web, and, in a broader sense, the new Information Technologies (IT) is one of the most changing, evolving and hottest areas of current computer science. Nevertheless many sub-area in this field could have potential benefits from Planning and Scheduling (P&S) technologies, and, in some cases, technology transfer has already started. This paper will consider and explore a set of topics, guidelines and objectives in order to implement the technology transfer a new challenges, requirements and research issues for planning which emerge from the web and IT industry.*

*Sample scenarios will be depicted to clarify the potential applications and limits of current planning technology. Finally we will point out some new P&S research challenge issues which are required to meet more advanced applicative goals.*

***Keywords**: Planning, Web, IT, technology transfer*

## Introduction

The Information Society (IS) is announced by a set of interrelated emerging technologies where the web it is certainly one of the most apparent and popular elements. These technologies envision new relationships of the individuals between his/her own tasks and the new tools.

Individuals are forced to develop new methods of work in order to exploit the ITs at their best, new tools and application should reflect and model these new methods in order to be effective.

Despite of the successful buzzword “web” (and popular e-*something* terms such as: "e-commerce", “e-business” etc.), it is important to focus on a wider vision of the potential role of planning and scheduling

technologies (P&S) in the Information Society not limited to the web applications. An exponential number of internet based services, low cost mobile devices over GSM and UMTS networks, tools which integrate traditional and knowledge based systems, represent promising application fields where P&S can have a primary or a supporting role.

P&S is a research area which dates back to [Fikes1971], scientific and technological results are due to a wide community of researchers and scientific programs on the topic.

P&S can convey the flexibility typical of AI technologies in order to answer to expectations and requirements of new methods of work and for the exploitation of the new I

There are some important elements which facilitate the application of P&S technologies to the web:

- *Digital content*, the web is machine readable and it provides a large quantity of machine readable data and software entities to interact with; automatic knowledge acquisition can be potentially realised in the web environment;
- *Virtual is real*, the web offers the chance having "real" planning application bypassing the complex robot machinery needed in traditional planning application fields such as aerospace or robotics, for example the actions "browse a list of available books" and "buy-a-book" can be easily executed in the digitalised world thus producing a real and useful effect;
- *Scalable complexity*, web applications can have different degree of complexity, simple applications can exists for simple existing planning models. As an example it is worth citing machine translation as a case of a technology which is not mature (i.e. for literature translation) but it could be worthwhile in niches applications (i.e. web pages automated translation)

Planning and scheduling technologies represent a key factor in the framework of the service technologies for the global new IS especially with respect to the current scenario of web applications, they provide:

- *Knowledge based flexibility*, P&S convey the flexibility typical of knowledge based technologies in support of new modalities of work, production, commerce, entertainment, education etc.;*Models of change and interaction*, P&S provides models of change and interaction, this research area has developed systems which provide automated generation of plans, tasks monitoring, plans checking etc.

Another key factor which motivates the application of P&S technologies to the web is its dimension: web is *large*. In other words, a massive audience of personal end-users and business end-users is developing the need of customised versions of services, and consequently the need for web industry of tools and technologies to support them. This scenario is made more complex by the increasing diffusion of personal mobile devices which fosters the development of new modalities of work and interaction between the individuals and the business organisations, and it tends also to modify the traditional modality of B2B interaction.

For this purpose the relevant issues to be supported and managed by services and applications over the information infrastructure are in general: *autonomy, adaptation, distribution, mobility, agent interaction, automatic collaborative support* etc.; which are also typical issues deeply related with P&S research.

Another important element has been the proposal of Semantic Web [Berners-Lee 2001], which has the ambitious goals of allow reasoning on a web of knowledge and meanings, with respect to the initial web made of presentation tags (i.e. HTML) and a web of syntactic structures and terms (i.e. XML). The data on the semantic web are defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. Semantic Web is of great important in the future of knowledge bases applications for the web, because of its success among the research community, the growing number of available tools and application and the standardisation factor, since it is supported and coordinated by W3C Consortium. Despite of the initial focus on ontologies and relationships, research started under the Darpha Agent Markup Language (the DAML program [DAML 2000] has lead to models of more dynamical aspects of the web. For example, DAML-S is oriented DAML-based Web Service Ontologies in order to facilitate the automation of Web service tasks including automated Web service discovery, execution, interoperation. It is worth mentioning the DAML-PDDL translators [PDDL2.1][DAML-PDDL] [McIlrait 2002] which apply P&S to the semantic web.

## Defining the Scenario: The Web as an Environment

In this view the *web is a planning environment*, i.e. web entities are the object of P&S systems.
Web entities exists ( a simple *web planning domain* is made of elements such as web pages, emails, files etc.) on which typical *actions* can be executed by users (e.g. pay, subscribe, supply, order, browse, look for, find, download etc.) or provided by service systems (e.g. web page servers, search engines, mail servers, messenger servers etc.). Moreover user and systems are pursuing *web tasks*, i.e. tasks which involve web entities (e.g. find a book, buy it and download), and in general user transactions and user activities over the web (e.g. informative tasks, educational tasks, distance work tasks etc.). Sometimes these tasks are not given automatic support, i.e. they rely on the user decisions, or, if automatic support is given, then it uses a too rigid and procedural approach, which does not satisfy the wide variety of user/business needs.
Planners can develop plans to act on the web virtual world in order to reach goals on *web entities*, a key point is that web entities do not necessarily have a "real" counterpart, and they are not necessarily designed as a part of a single distributed system.
Consider for example the following completely *virtual* plan for the *virtual* goal of promoting a web site: *buy disk space, transfer your old site to a popular web portal in order to obtain a better click rate*. The synthesis of this plan would require the ability of describing a model of change (e.g. defining the meaning of acting on the web), goals (defining concepts such "a better visibility"), and a model of the actions to be taken to execute it (i.e. pay, subscribe, download etc.), i.e. web entities should be represented as part as part of a planning domain.
In addition there are some traditional planning phases that need to be reconsidered in the web domain:

- *domain knowledge acquisition* can be realised by activities of information gathering, information discovery and comparison (about existing portals, rates and prices), with respect to more traditional planning domain in which domain knowledge has not frequent variations;
- *action execution* in the web could imply to take into account of a dynamical scenario, in which not only actions can have unpredictable failure, but the domain can change during execution.

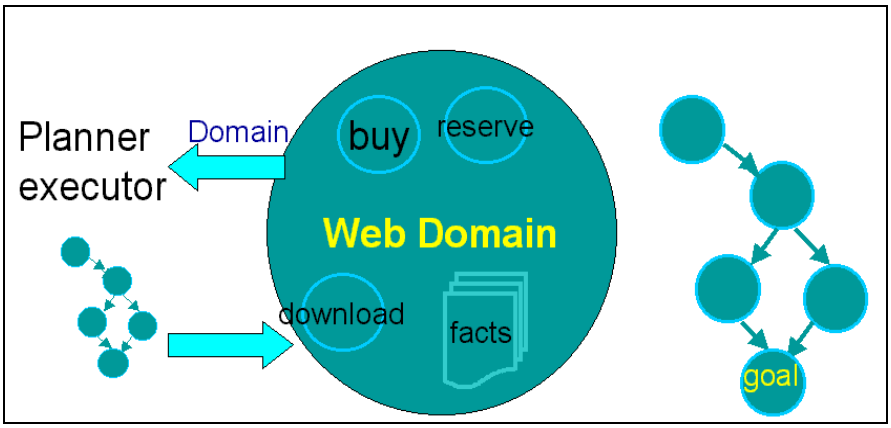



Fig.2 The Web as a Planning Environment

A more general scenario is characterised by activities which take place only *partially* on (or through) the web.

*A simple example is that of a personal planner assistant, which suggests the user to buy a textbook online as* part of *the plan "successfully preparing an exam", which contains some other* not web *steps such as* going to lessons *and* doing exercises *. User tasks and goals are in general related with real world activities and should interact and be coordinated with actions and plans which act in the web domain.*

When more production related activities become available on the web (for example: *suppliers chains, customers, markets, delivery, payments etc.*) the manufactures planning activity will need to model the web as part of the production plans.

## A Glossary of Terms for Mapping Planning on the Web

In this section it will be shown how the basic features of planning domain model can be mapped into corresponding web entities, and contribute to solve and support web tasks and goal oriented activities on the web. On the other hand the mapping allows pointing out the limits of the current planning models and the research issues which are required to be solved in order to give a complete account of operating on the web domain.

Glossary Terms: State, Initial State, Goal State.

The notion of state is central to most planning models. In the web the concept of state is represented by the sum of the states of the various component of the web domain:

- *user internal state*, the set of features and facts describing the user state (e.g. maintenance conditions, general constraints, preferences, but also info about users such as identities and passwords);
- *web-state facts*, the state of info as available on the web, in the simplest form web-state is represented by the set of current web pages, existing files etc., more likely the state description would address the content and the semantic of available information (e.g. stock quotations, available item on e-markets etc.)
- *web actors state,* this is represented by the internal state of interactive services, consider, for example, activities which require multiple steps in order be carried out, such as online reservation of a flight (it is necessary to be able to represent the current step of a given transaction), or consider services which require access authorisation (it is necessary to represent that certain existing resources are/are not available for use).

The concept of *state that is used in planning to model initial states, goals state, and to model change as modification of state* will need reformulation in the framework of web applications, because of the inadequacy of current planning model to fully describe the state of a web domain in all its aspects.

Research Issue: the main drawback of the current planning models of states derive from the fact that the web is a *vast and dynamical environment of active resources*, on the other hand planning domains are usually characterised by domain states which are fully knowledgeable and somewhat static, i.e. the planner is the only one (or one of the few) agent in the domain.

Summarising, the main P&S research issues that should be addressed in order to fully capture the main features of a *web state* are:

- *managing uncompletness*, web state is vast and not completely knowledgeable, the web domain is inherently incomplete there is no hope to model inside the planner *a complete* description of all the web, then mechanisms and strategies are required in order to do knowledge acquisition and to *circumscribe* the domain description for the planning problem at hand ;
- *managing events*, web state changes independently of the planning actor, this is usually addressed in planning models by the concept of *events*, i.e. state changes which are out of the control of the planner [Ghallab 1998]; in the web domain the amount of *planning time* appear to be a crucial factor, since dynamical changes can occur during planning;
- *managing inconsinstency*, web state can contain contradictory information coming from different sources, consider for example info about the weather conditions used by a travel planner; concepts such as trust and believes and should be modelled [Ambite 1998];
- *managing richer knowledge representation,* P&S models usually focus more on actions and plans than on the adequacy of the knowledge representation formalism to capture the *static* aspect of the domain; since the web offers information in different but equivalent forms, the planners should be aware of ontologies and mechanisms for relating information from different sources, consider for example the problem of representing the concept of *price*, in the sense of amount of money to be payed for buying a good, an effective representation requires that *stock exchange market quotations*, *monthly renting rates*, *price of a book* and *currency exchange rates* would represent instances of the same concept of price [Berners-Lee 2001],

Glossary Terms: Actions and Operators

The other fundamental concept in planning models is the notion of *action* and *operators* , which represents the basic elements to model change in the domain.

*Actions* (i.e. istances of *operators*) represent the state transiction which occurs during the execution. Operator and action are usually modelled in term of precondition/effects.

In a web domain actions are represented by the available services, which change the user/web state according to a preconditions/effects model. For example available services on the web such as *buying a book, reserving a meeting room, downloading a satellite image of a town, moving a remote webcam* , *sending an email*, are actions in the sense that are allowed to take place only under certain preconditions on the web state (e.g. *having enough money on the account, availability of the tow sat picture etc.*) and produce effects on it (e.g. *changing book ownership, meeting room reservation state, a copy of the picture locally available etc.*).

Research Issues: it is worth noticing that an effective model of *web actions* should take into account the elements formerly pointed out about web states (uncompletness, inconsistency in particular), but it requires, in addition, to consider some *web specific* P&S research issues about the actions/operators model:

- *operators can change, appear and disappear*, new services become available and changes over time, issue: dynamical discovery, monitoring and maintaining of domain models;
- *actions execution time,* actions on the web take time (depending on bandwith and servers overhead factors) , issue: timing constraints and failure recovery;

Glossary Term: Plan

"A *plan* is a set of ordered actions, that, if executed in the initial state will transform it in the goal state". This definition of solution plan is common to most planning models since [Fikes 1971], and it is easy to see that the definition easily applies also to the web domain. Currently most web tasks and goal oriented activities which take place on the web can be described basically as sequences of actions, i.e. on the web, *plans are sequences of interrelated services requests made on web entities*, (consider for example a typical user driven web-plan such as*: look for items sellers on search engine, browse and compare prices, order one and pay for it*).

Research Issues: although plan as sequences or partially ordered sequences are a formalism that is sufficient for planning in the web at a basic level, it is worth to investigate on planning models which

- *provide expressive and flexible models of plans*, few planning models have a satisfactory management of actions with duration, loops, conditionals [Lin 1995];
- *combine plan knowledge with task oriented languages*, planning models based on a hierarchical approach [Erol 1995], as well as of workflow management models [WorkFlow-Roadmap 2003] offer examples of task oriented formalisms which should be included into generative planning models.

The objective should be to reflect also in the structure of plans, some typical element of the web domain, we have already pointed out, such as contingency, non-determinism, and dynamical aspects of the domain.

Glossary Terms: Plan Monitoring, Execution, Sensing, Re-planning

Planning terms such as *plan monitoring, execution, sensing* and *re-planning* regard the phase that follow plan synthesis and aimed at actually reaching the goals by executing the plan in the real world. In the web domain (where *virtual and real* are somewhat overlapping) can characterised in a "web specific" way traditional planning concepts like *execution* and *sensing*:

•*Execution*, since the web is *large*, execution can involve complex decisions about choosing the service to invoke among a plurality of available and equivalent ones; moreover *execution takes time* also for processing and data transfer, then criteria are needed to combine the issue of execution time with the issue of bandwidth; finally note as the implementation of *actuators* would be greatly favoured by the diffusion of web services, but the use of *wrappers* or similar mechanism should also be considered as an intermediate solution for interacting with services initially designed for human users;

•*Sensing*, in the web environment sensing mainly means: *actively* looking for info, i.e. gathering info [Knoblock 1995] [Golden 1996b] [Naveen 1997] and results from web sources (for example *web pages, searching databases, streaming video, results from called services*)*;* main open problems in this area are related with the integration of information from different non-homogeneous sources, and with decisions to be

taken about *what to sense* (consider for example criteria like: *time for sensing and processing vs bandwidth vs timing goals and execution constraints*)
It is also worth noticing that the dynamical nature of the web suggests the investigation of theoretical models [Haigh1996] [Friedman 1997], where the plan synthesis phase is interleaved with execution and sensing.

## Conclusion

It is not easy to envision the future of technology transfer in an area where the target industry, i.e. the web and the broader IT society, is moving and developing at an high unprecedented pace toward hard to predict directions. For these reason, previsions and objectives in the long term are not very realistic, and should be limited only to short term and medium term scales.
On the basis of the current achievements and potentiality of P&S technologies depicted in the previous sections some conclusive guidelines for future perspectives and goals can be drawn along two main dimensions: End-user requirements and Research goals.
The first examples of the technology transfer between P&S and the Web have been in the area of:

- automation and autonomy for machine to machine services, (P&S, softbot and agents technologies Etzioni 1994])task support, Travel Assistant [Ambite 2001], i.e. P&S based dynamic integration of web sources for travel organisation (managing and combining multiple sources for airfares, parking fares, timetables, weather conditions etc )automatic maintenance of web applications, web info and web services [Smith 2001] [McIlraith 2001] [Marcugini 2002][Thakkar 2002]; (the P&S knowledge based representation of change, constraint checking and failure repairing techniques has been applied to support automatic maintenance in not web domains [Chien, 1998]);

### End-user requirements,

End-users are intended to be in most cases personal users, but in the following we also means by end-user a company, an organisation as a whole, (e.g. in organisational workflow management) or an industry using the technology as a component of more complex systems.
The application issues that are expected to be asreached in a short term are mainly in the area of

- support/automation of personalised services and user adaptive services on the web (educational, e-learning, recreation, information; automatic synthesis/monitoring of online courses and educational plans based on the personal skills/advancements; automatic synthesis of personalized newspaper based on user models);
- supporting mobility and distribution in organisation: scheduling activities & information workflow in distributed organisation through the web and new devices/communication media, supply chain management
- *supporting user web tasks*, (e.g. cooperative recognition, automation, monitoring and repairing of user web tasks, consider, for example, a planning based systems which support the user in task such as “organize a vacation” or “organize a meeting”);
- automatic integration of Services/Web Services, goals directed synthesis and maintenance in specific domains

There are potential applications of P&S as components for the new IT systems:

- *components for web servers and tools*, (e.g. applications to goal guided synthesis and maintenance of web sites/pages)
- *components for supporting online services*, (e.g. scheduling load distribution to online assistance operators, activation of assistance/emergency chains, online configuration of products);
- *component in systems for online supply chains management*, (issues: logistics, resources management, workflow management, tasks scheduling and assignment)

On a medium/long term horizon, we expect increasingly to met higher requirements on tools and techniques for automation of Web Services, and increasing requests for personal support services, either on the web or onboard of personal assistant devices, such as:

- robust automatic integration of Web Services, in broader application domains, support for automatic discovery of services and failure recovery

- autonomous web agents, goals directed software agents which operate over the web on user's behalf (e.g. auctioning systems, stockbot etc.);
- support to massive adaptative services on small personal devices, planning and scheduling support for daily activities (e.g. home devices);
- cooperative and distributed planning which integrates personal devices/web services/web agents for global user goals*online P&S services*, general or specialized tools which plan or schedule on demand (e.g. for checking/rescheduling activities) and are available on the web,

Research goalsThe following research goals are prerequisites for supporting the initial and next steps of the technology transfer, on the other hand it seems to exist an increasing autonomous interest of the research community on topics such as modelling dynamic domains, time management, interleaving planning/execution and others, which can be exploited for modelling the web environment, the main topics related to this purpose include:

- modelling domain discovery during planning
- contingency and sensing in web environment
- time duration, failure recognition, and repair
- expressive and robust model of execution plans, (loop iterations)
- mixed initiative planning models, task support models
- incorporating web oriented extension in PDDL-like language (see [KE-Roadmaps] and [WorkFlow-Roadmap])
- wrappers and planning operator wrappers

in a second phase advancements in the previous topics are needed, as well as an increasing the research efforts toward more technology oriented topics such as *portability*, *interoperability, scalability* and advanced planning models, these objectives include:

- portability for planning algorithms, strategies, heuristics, preprocessing
- interoperability: mapping between planning models
- models and measures of planning scalability (small devices/ small domains)
- models for cooperative distributed planning, (domains/problem partitioning/integration)

It is generally expected that knowledge based technologies (the semantic Web proposal for instance) can give effective contribution to web applications. We have pointed out that planning technologies, in particular, can positively exploit on the web their ability of modelling action and changes. We also point out the potential scalability of the technology transfer of P&S to the web in a short/medium term perspective from simple applications which use ready-off-the-shelf planning technologies to more complex applications which need appropriate innovative models. On the other hand web and IT applications represent a source of interesting open research issues such as managing of uncomplete and dynamical domains; interoperability among domains, planners and planning tools.

## Bibliography

[Ambite 2001] J.L. Ambite, *Heracles: Building Planning and Information Assistants* , Proceedings of the ECP/PLANET Workshop on Automated P&S in New Methods of Electronic, Mobile, and Collaborative Work, Toledo, Spain, September 2001,

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

[Chien,1998] S. Chien, F. Fisher, H. Mortensen, E. Lo, R. Greeley, A. Govindjee, T. Estlin, X. Wang, *Using Artificial Intelligence Planning Techniques to Automatically Reconfigure Software Modules*, Artificial Intelligence and Knowledge-based Systems, Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998

[DAML 2000] Burke M. manager, *The Darpha Agent Markup Language Project DAML* , www.daml.org

[DAML-S 2002]The DAML Services Coalition (Anupriya Ankolenkar et al.), *DAML-S: Web Service Description for the Semantic Web*, The First International Semantic Web Conference (ISWC), Sardinia (Italy), June, 2002.

[DAML-PDDL] Mc Dermott D., Dou D., Qi P., *PDDAML:An Automatic Translator Between PDDL and DAML* http://www.cs.yale.edu/homes/dvm/daml/pddl_daml_translator1.html

[Erol 1995] K. Erol, D. Nau, J. Hendler, *HTN Planning: Complexity and Expressivity,* in AAAI-94, Seattle, July, 1994

[Etzioni 1994] Etzioni O., Weld D. *A Softbot-Based Interface to the Internet*, Communications of the ACM, 1994

[Fikes 1971] Fikes, R. and Nillson, N., *STRIPS: A new approach to the Application of Theorem Proving to Problem Solving*, Artificial Intelligence 2, 1971, pp. 189-208.

[Friedman 1997] Friedman M., Weld D.S., *Efficiently Executing Information Gathering Plans* Proceeding of International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 1997

[Ghallab 1998] Ghallab M. *Chronicles as a practical representation for dealing with time, events and actions*, Proceeding of the 6th Italian Conference on Artificial Intelligence (AIIA'98), Padova (Italy), Lecture Notes on AI, pp.6-10, September 1998

[Golden 1996a] Golden K., Etzioni O., Weld D.S., *Planning with Execution and Incomplete Information*, Technical Report, University of Washington, n.UW-CSE-96-01-09, 1996

[Golden 1996b] Golden K., *Leap before you look: information gathering in the PUCCINI Planner*, Proceedings of AIPS 1998

[Haigh 1996] Haigh K.Z., Veloso M., *Interleaving Planning and Robot Execution for Asynchronous User Requests,* AAAI Spring Symposium on Planning with Incomplete Information for Robot Problems, 1996

[KE-RoadMap 2003] T.L. McCluskey editor, *Knowledge Engineering for Planning Roadma*, PLANET-II Technical Coordination Unit on Knowledge Engineering, March 2003, available on line at http://www.planet-noe.org/TCUs

[Knoblock 1995] Knoblock G.A., *Planning Executing, Sensing and Replanning for Information Gathering*, Proceeding of the International Joint Conference onAI, IJCAI 1995

[Laithwaite 2001] Laithwaite B., *BritishTelecom Workforce Management System*, Proceedings of the ECP/PLANET Workshop on Automated P&S in New Methods of Electronic, Mobile, and Collaborative Work, Toledo, Spain, September 2001,

[Lin 1995] Lin, S.-H., Dean, T. (1995). *Generating optimal policies for high-level plans with conditional branches and loops*. In Proceedings of the Third European Workshop on Planning, pp. 205--218.

[Marcugini 2002] Marcugini S., Milani A., *Automated Planning of Web Services,* Proceedings of the 3rd International Conference on Electronic Commerce, ICEC 2002, Hong Kong, October 2002

[McDermott 2002] McDermott D., *Estimated-Regression Planning for Interactions with Web Services*, Proceedings of the AI Planning Systems Conference (AIPS'02), June 2002.

[McIlraith 2001] McIlraith, S., Son, T.C. and Zeng, H., *Semantic Web Services* , IEEE Intelligent Systems. Special Issue on the Semantic Web. 16(2):46--53, March/April, 2001

[McIlrait 2002] McIlrait S., Fadel R., *Planning with Complex Actions,* Proceedings of AIPS 2002 Workshop on Exploring Real World Planning, Tolouse, France, April 2002

[Naveen 1997] Naveen A., Knoblock G. A., Levy A.Y., *Information Gathering Plans with Sensing Actions,* in Recent Advances in AI Planning: Proceedings of European Conference on Planning ECP97, Springer-Verlag, New York, 1997

[PDDL2.1] Fox M., Long D. editors, *PDDL 2.1: Planning Domain Description Language* http://www.dur.ac.uk/d.p.long/IPC/pddl.html

[Sahuguet 1999] Sahuguet A., Azavant.F. *WysiWyg Web Wrapper Factory*. In WWW8, 1999.

[Smith 2001] Stephen F. Smith S.F., Hildum D., Crim D. *Toward the Design of Web-based Planning and Scheduling Services*, Proceedings of the ECP/PLANET Workshop on Automated P&S in New Methods of Electronic, Mobile, and Collaborative Work, Toledo, Spain, September 2001,

[Thakkar 2002] Thakkar S., Knoblock G.A., Ambite J.L., and Shahabi C., *Dynamically Composing Web Services from On-line Source,* Workshop on Intelligent Service Integration, AAAI02, Edmonton, Alberta, Canada, 2002 , *http://www.isi.edu/info-agents/dotnet/aaaiworkshop2002.pdf*

[WebTCU] A.Milani editor, *Website of the PLANET-II Technical Coordination Unit on Planning and Scheduling for the Web*, http://www.planet-noe.org/TCUs

[WorkFlow-Roadmap 2003] D. Borrajo Editor, *Planet Workflow Management R&D Roadmap*, PLANET-II Technical Coordination Unit on Workflow Management, February 2003, available on line at http://www.planet-noe.org/TCUs

## Author information

Alfredo Milani - Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: milani@unipg.it

Stefano Marcugini - Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: gino@dipmat.unipg.it

# THE KNOWLEDGE: ITS PRESENTATION AND ROLE IN RECOGNITION SYSTEMS

*A. D. Zakrevskij*

*Abstract: The concept of knowledge is the central one used when solving the various problems of data mining and pattern recognition in finite spaces of Boolean or multi-valued attributes. A special form of knowledge representation, called implicative regularities, is proposed for applying in two powerful tools of modern logic: the inductive inference and the deductive inference. The first one is used for extracting the knowledge from the data. The second is applied when the knowledge is used for calculation of the goal attribute values. A set of efficient algorithms was developed for that, dealing with Boolean functions and finite predicates represented by logical vectors and matrices.*

## INTRODUCTION

Knowledge is the central concept in a wide variety of investigations dedicated to pattern recognition problems [1, 3, 6]. Solving them begins with the choice of a proper *world model* – an abstract artificial world reflecting some important qualities of real subject areas – important from the point of view of the problems to be solved.

In this paper, we use a very simple model which defines this abstract world as a finite set of objects possessing some combinations of qualities. More formally, this model is defined as a set $W$ of logical $n$-vectors presenting values of attributes composing the set $X = \{x_1, x_2, ..., x_n\}$. The attributes could be binary (Boolean) or multi-valued. In the latter case each of the attributes $x_i$ is characterized by a corresponding finite set $V_i$ of alternative values. The Cartesian product of these sets $V_1 \times V_2 \times ... \times V_n$ (or $\{0, 1\}^n$ in the binary case) constitutes the space of attributes $M$. Elements from $W$ can be regarded as abstract models of real objects of a natural subject area. The world as a whole is represented by a subset $W \subseteq M$ or by the corresponding finite predicate $\varphi(x_1, x_2, ..., x_n)$ which takes value 1 on the elements of the set $W$. In case of two-valued attributes this predicate is a Boolean function $f(x_1, x_2, ..., x_n)$. That approach is rather simple, inasmuch the world is considered only as the set of its elements (without regarding any relations between them), but it is sufficient for solving many practical problems.

Usually, only partial information about the world $W$ is known represented in terms of data and knowledge. Suppose that the *data* present information concerning some separate elements from $W$, describing these elements by corresponding logical vectors. Taken together, these vectors represent a so called sampling population (a reliable selection from the subject area) and constitute the set $F$ serving further as a *database*. As a rule, $|W| << |M|$ and $|F| << |W|$. The *knowledge*, on the contrary, presents information about qualities of the whole subject area, expressed by some inherent in the regarded subject area regularities which establish some relationships between attributes.

Both data and knowledge present information about the regarded world, but they are of different type. Comparing them, we could say that data consist of affirmations about the *existence* of some objects with definite combinations of qualities. Unlike that, the knowledge puts some restrictions on what is possible, affirming the *non-existence* of objects with some other combinations of qualities.

The pattern recognition process, taken as a whole, may be roughly divided into two main stages: obtaining some knowledge by data mining and predicting values of goal attributes by using this knowledge. The methods of inductive and deductive inference are applied at these stages, accordingly. Their efficiency depends greatly on the form in which the knowledge is presented. A special attention is paid below to this point.

## THE KNOWLEDGE – CONCEPT AND FORMAT

Within the framework of our world model, the knowledge is defined as a set of regularities. The key question is to choose a proper model for them. Starting from general assumptions it is accepted that any regarded regularity defines a logical connection between some attributes: it means that some combinations of attribute values are declared impossible (prohibited).

Evidently, the less attributes are connected by some regularity, the stronger is the latter. That is confirmed, for instance, by the long history of investigations in physics and other nature sciences. On the other hand, if we choose several attributes and decide to connect them by a regularity, it will be the weakest when it forbids only one combination of those attributes values.

In the Boolean case such a regularity can be expressed by the logical equation $k_i = 0$ or by the equivalent to it equation $d_i = 1$, where $k_i$ is a conjunct formed of some attributes (in direct or inverse mode) from the set $X$, and $d_i$ is the corresponding disjunct, satisfying the relation $d_i = \neg k_i$. For instance, equations $ab'c = 0$ and $a' \vee b \vee c' = 1$ ( ' is the symbol of inversion, equivalent to $\neg$) represent the same regularity, which prohibits the combination of values $a = 1$, $b = 0$, $c = 1$ .

A regularity of this kind is called *implicative* (more general than functional one) [5]. It prohibits a set of attribute value combinations forming an interval in the Boolean space $M$ over $X$ – the characteristic set of the conjunct $k_i$. In other words, the regularity affirms that this interval is empty, not containing any elements of the world $W$. The size of that interval (the number of its elements) equals $2^{n-r}$, where $n$ is the number of all attributes and $r$ is the rank of the implicative regularity – the number of attributes coming into it. It becomes clear now how the strength of the regularity is defined by its rank.

Suppose that $X = \{a, b, c, d, e, f\}$ and consider the implicative regularity $ab'e = 0$ forbidding the combination 101 of values of the attributes $a, b, e,$ accordingly. The corresponding empty interval of the space $M$ contains eight elements: 100010, 100011, 100110, 100111, 101010, 101011, 101110 and 101111. All these elements ($\alpha$) are "between" two elements 100010 (minimal) and 101111 (maximal), satisfying the inequality $100010 \leq \alpha \leq 101111$, and hence justifying the term *interval*. The equation $ab'e = 0$ may be changed for the equivalent equation $ab'e \rightarrow 0$ with the implication operator $\rightarrow$ (if… then…), known as a *sequent* (its left part is always a conjunction, and the right part – a disjunction). The latter equation may be subjected to equivalence transformations consisting of transferring arbitrary literals between the left part (conjunction) and the right one (disjunction), changing each time their type (positive for negative or *vice versa*). In such a way we could obtain the following set of the equivalent equations

$ae \rightarrow b$ (if $a = 1$ and $e = 1$, then $b = 1$), $ab' \rightarrow e'$, $a \rightarrow b \vee e'$, …, $1 \rightarrow a' \vee b \vee e'$.

The last one could be changed for the disjunctive equation $a' \vee b \vee e' = 1$, which may be coded by the ternary vector 01 - - 0 - .

A set of regularities given in such a form can be presented by a ternary disjunctive matrix $D$, called below a *knowledge matrix*, which columns are marked with attribute symbols. For example, the knowledge matrix

$$D = \begin{array}{cccccccc} a & b & c & d & e & f & g & h \\ 1 & - & - & 0 & - & - & 0 & - \\ - & - & - & 1 & - & 1 & - & - \\ 0 & 1 & - & - & - & - & - & - \end{array}$$

affirms that every object of the regarded area must satisfy the equations

$$a \vee d' \vee g' = 1, \; d \vee f = 1, \text{ and } a' \vee b = 1.$$

In other words, in the considered Boolean space there exists no object which has any of the following combinations of values of some attributes:

$(a = 0, d = 1, g = 1)$, $(d = 0, f = 0)$ and $(a = 1, b = 0)$.

The set of these equations can be reduced to one equation $D = 1$ where $D$ is a CNF (conjunctive normal form) represented by the matrix $D$.

$$D = (a \vee d' \vee g')\,(d \vee f)\,(a' \vee b) = 1.$$

By inverting both left and right parts of the equation $D = 1$ we get the equivalent equation $K = 0$ with the left part $K = \neg D$ – a DNF (disjunctive normal form) presenting a so called *veto function* $V$, which defines the prohibition area.

For the regarded example

$$K = a'dg \vee d'f' \vee ab' = 0.$$

The suggested form of implicative regularities turned out to be extremely convenient at the stage of deductive inference, where the methods developed for theorem proving automation are successfully applied [2]. As it is shown below, regularities of the considered type could be rather easily discovered in the database, and it is not difficult to evaluate their strength and plausibility, which is very important for their further application.

In the case of finite predicates generalized conjuncts and disjuncts could be used to present the knowledge [7, 8]. Any interval in the space of multi-valued attributes is defined as a direct product of non-empty subsets $\alpha_i$ taken by one from each set $V_i$. Its characteristic function is defined as a conjunct, and the negation of the latter is a disjunct.

Suppose $X = \{x, y, z\}$, and the attributes $x$, $y$, $z$ select their values from the corresponding sets $V_1 = \{a, b, c\}$, $V_2 = \{a, e, f, g\}$, $V_3 = \{h, i\}$ (note that these sets may intersect). Let $\alpha_1 = \{a\}$, $\alpha_2 = \{a, e, g\}$, $\alpha_3 = \{h, i\}$. The interval $I = \alpha_1 \times \alpha_2 \times \alpha_3$ has the characteristic function (conjunct)

$$k = (x = a) \wedge ((y = a) \vee (y = e) \vee (y = g)) \wedge ((z = h) \vee (z = i)),$$

which could be simplified to

$$k = (x = a) \wedge ((y = a) \vee (y = e) \vee (y = g)),$$

inasmuch as $(z = h) \vee (z = i) = 1$. If this product enters the equation $k = 0$ which reflects a regular connection between $x$ and $y$, then $I \cap W = \varnothing$, i. e. the interval $I$ turns out to be empty.

As it can be seen from the given example, the structure of a conjunctive term in the finite predicate algebra is more intricate compared with that of the binary case – the two-stage form of the type $\wedge \vee$ is inherent in it. One can avoid that complexity changing the equation k = 0 for the equivalent equation $\neg k = 1$ and transforming $\neg k$ into a one-stage disjunctive term $d$. Such transformation is based on the de-Morgan rule and changes expressions $\neg (x_i \in \alpha_i)$ for equivalent expressions xi $\in V_i \backslash \alpha_i$ . This is possible since all sets Vi are finite.

For the considered example

$$d = \neg k = (x \neq a) \vee ((y \neq a) \wedge (y \neq e) \wedge (y \neq g)) = (x = b) \vee (x = c) \vee (y = f).$$

Adhering to the tradition, let us call similar expressions as disjuncts. Suppose that the knowledge obtained either from experts or by induction from the data is represented by a set of disjuncts $d_1, d_2, ..., d_m$ . Generated by them, equations $d_i = 1$ are interpreted as conditions which should be satisfied for any objects of the world, and it is possible to reduce them (equations) to a single equation $D = 1$ the left part of which is presented in the conjunctive normal form – CNF $D = d_1 \wedge d_2 \wedge ... \wedge d_m$ . It follows from here that in the finite predicate algebra the CNF has some advantage over the disjunctive normal form – DNF $K = k_1 \vee k_2 \vee ... \vee k_m$ which is used in the equivalent equation $K = 0$. Indeed, DNF has three stages $(\vee \wedge \vee)$, whereas CNF – only two $(\wedge \vee)$.

In the case of multi-valued attributes, it is more convenient to use sectional Boolean vectors and matrices introduced for representation of finite predicates [7]. A sectional Boolean vector consists of some sections (domains) corresponding to attributes and each section has several binary digits corresponding to the attribute values indicating definite properties. For example, the section corresponding to the attribute *color*, which has the values *blue*, *red*, *green*, *yellow*, *brown*, *black* and *white*, should have 7 bits. For the example given above, the vector 010.1000.01 describes an object with the value *b* of the attribute *x*, the value *a* of the attribute *y* and the value *i* of the attribute *z*. Obviously, if a vector represents some element of the space *M* of multi-valued attributes, it has the only 1 in each section. The situation is different in the case of some fuzziness. The vector 011.1001.01 can be interpreted as presenting a partial information about the object, when we know only that $x \neq a$, $y \neq e$, $y \neq f$ and $z \neq h$. Note, that each of these inequalities serves as an *information quanta* and is marked by a zero in the corresponding component of the vector.

Giving an example of presenting the knowledge, suppose that $X = \{a, b, c\}$, $V_1 = \{1, 2, 3\}$, $V_2 = \{1, 2, 3, 4\}$ and $V_3 = \{1, 2\}$. Then the knowledge matrix

$$D = \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} 0\,0\,1\,.\,0\,0\,1\,0\,.\,0\,0 \\ 1\,1\,0\,.\,0\,0\,1\,1\,.\,0\,1 \\ 0\,1\,0\,.\,1\,1\,0\,0\,.\,1\,0 \end{array} \end{array}$$

0 0 1 . 0 1 0 0 . 0 1

may be interpreted as a set of disjunctive equations

$$(a = 3) \vee (b = 3) = 1,$$
$$(a = 1) \vee (a = 2) \vee (b = 3) \vee (b = 4) \vee (c = 2) = 1,$$
$$(a = 2) \vee (b = 1) \vee (b = 2) \vee (c = 1) = 1,$$
$$(a = 3) \vee (b = 2) \vee (c = 2) = 1$$

or as one equation with a CNF in the left part:

$$((a = 3)\vee(b = 3)) \wedge ((a = 1)\vee(a = 2)\vee(b = 3)\vee(b = 4)\vee(c = 2)) \wedge$$
$$\wedge ((a = 2)\vee(b = 1)\vee(b = 2)\vee(c = 1)) \wedge ((a = 3)\vee(b = 2)\vee(c = 2)) = 1.$$

## DATA MINING

A very important part of the pattern recognition problem is obtaining knowledge from data [3]. The data could be represented by a sampling population *F* – a set of some randomly selected elements from the regarded world *W*.

As it was formulated above, we solve that problem by analyzing the distribution of elements of the set *F* in the space *M* (suppose it is Boolean) and revealing implicative regularities which are reflected by empty intervals (not intersecting with *F*). That operation can be reduced to observing a Boolean data matrix *K* and looking for such combinations of attribute values which do not occur there.

The number of attributes coming into an implicative regularity is called its *rank*. It coincides with the rank of the corresponding interval. Remind that the less attributes are tied with a regularity, the stronger is the tie. So, it is worthwhile to look for regularities of smaller rank.

Consider, for example, the following data matrix *K*:

| *a* | *b* | *c* | *d* | *e* | *f* |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 1 | 0 |
| 0 | 0 | 1 | 0 | 1 | 0 |
| 1 | 1 | 1 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 1 |

There are no empty intervals of the rank 1, because each column contains 1s and 0s. So we look further for empty intervals of the rank 2 and find five of them, corresponding to the following combinations: ($a$ = 0, $f$ = 1), ($b$ = 1, $d$ = 0), ($b$ = 0, $e$ = 0), ($c$ = 1, $f$ = 1), ($d$ = 0, $e$ = 0). In a more compact form these intervals may be represented by conjuncts *a'f*, *bd', b'e'*, *cf*, *d'e'*. Can we consider that these found empty intervals reflect real regularities inherent in the world from which the data were extracted? Such conclusions could be accepted if only they are plausible enough.

Consider the general case of *n* binary attributes and *m* elements in the sampling population (selection) *F*. Suppose, we have found an empty interval of the rank *r* (comprising $2^{n-r}$ elements of the Boolean space *M* and put forward the corresponding hypothesis, affirming that this interval is free of any elements from the regarded world *W*. May we rely on it and make with its help some logical conclusions when recognizing an object with the unknown value of the goal attribute? The problem is to estimate the plausibility of that hypothesis.

We should take into account that the regarded interval could be empty quite accidentally, while in reality the selection *F* is taken by random from the whole space *M* – in that case there could be no regularities in the disposition of the elements from *F* in *M*.

It would be useful to express the probability *p* of such an event as a function $p(n, m, r)$ of the parameters *n, m, r*. The hypothesis can be accepted and used further in procedures of deductive inference if only this

probability is small enough. Its calculation is rather difficult, so it was proposed in [5] to approximate it by the mathematical expectation $E(n, m, r)$ of the number of empty intervals of the rank $r$.
That value can be calculated by the formula

$$E(n, m, r) = C_n^r 2^r (1-2^{-r})^m ,$$

where $C_n^r$ is the number of $r$-element subsets of an $n$-element set, $C_n^r 2^r$ is the number of intervals of the rank $r$ in the space $M$, and $(1-2^{-r})^m$ is the probability of some concrete interval of the rank $r$ to be empty, not containing any elements from $F$.
Some empty intervals could intersect, hence $E(n, m, r) \geq p(n, m, r)$. The question is how big could be the difference $E(n, m, r) - p(n, m, r)$? It was shown, that it becomes negligible small for small values of $E(n, m, r)$. But that is just the case of interest for us.
It turns out that the value of the function $E(n, m, r)$ grows very rapidly with rising $r$. That is evident from the Table 1 of the dependence $E$ on $r$ under fixed values of other parameters: $n = 100$ and $m = 200$.

Table 1. The dependence $E$ on $r$ under fixed $n$ and $m$

| $r$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $E(100, 200, r)$ | $1.24\times10^{-58}$ | $2.04\times10^{-21}$ | $3.26\times10^{-6}$ | $1.56\times10^{2}$ | $4.21\times10^{6}$ | $3.27\times10^{9}$ |

It is clear that the search for empty intervals and putting forward corresponding hypotheses can be restricted in this case by the relation $r < 4$. If some empty interval of the rank $r < 4$ is found, we can formulate the corresponding regularity with good reason, but there are no grounds for that if $r \geq 4$. So, when $n = 100$ and $m = 200$, there is no sense in looking for empty intervals of the ranks more then 3. The search for regularities could be strongly restricted in that case by checking for emptiness only intervals of the ranks 1, 2 and 3, which number is

$$C_{100}^1\times2^1 + C_{100}^2\times2^2 + C_{100}^3\times2^3 = 1{,}333{,}400.$$

Not much, compared with the number $3^{100}$ of all intervals in the Boolean space of 100 variables, approximately $5.15\times10^{47}$.
A threshold $\omega$ may be introduced to decide whether it is reasonable to regard an empty interval as presenting some regularity: the positive answer should be given when $E < \omega$. Its choice depends on the kind of problems to be solved on the base of the found regularities.
Suppose $\omega = 0.01$. Then the maximum rank $r_{max}$ of intervals which should be analyzed when looking for regularities could be found from Table 2, showing its dependence on $n$ and $m$.

Table 2. The dependence of the maximum rank $r_{max}$ on parameters $n$ and $m$

| $n \setminus m$ | 20 | 50 | 100 | 200 | 500 | 1000 |
|---|---|---|---|---|---|---|
| 10 | 1 | 2 | 3 | 4 | 5 | 6 |
| 30 | 1 | 2 | 2 | 3 | 4 | 5 |
| 100 | 1 | 1 | 2 | 3 | 4 | 5 |

Two conclusions, justified for the regarded range of parameters, could follow from this table. First, in order to increase $r_{max}$ by one it is necessary to double the size of the experiment – the number $m$ of elements in $F$. Second, approximately the same result could be achieved by reducing by a factor of 10 the number of attributes used for the description of the regarded objects.
Suppose $r_{max} = 2$ which is enough when the selection $F$ is rather small. In that case we have to pay attention only to pairs of attributes, looking for some forbidden combinations of their values. This task can be executed by an incremental algorithm. It analyzes the elements of the selection $F$ consecutively, one by one, and fix such two-element combinations which have occur, using a symmetrical square Boolean $2n\times2n$ matrix $S$ for that, with rows and columns corresponding to the values $x_1 = 0$, $x_1 = 1$, $x_2 = 0$, $x_2 = 1$, etc. Its elements corresponding to occurring combinations are marked with 1. The rest combinations (not occurring) are presented by zero (empty) elements and accepted as forbidden. The regularities presented by them connect some attributes in pairs and are called syllogistic [6]. For example, regarding the following selection $F$ (only to illustrate the algorithm, despite the fact that the selection is too small for $r_{max} = 2$):

| *a* | *b* | *c* | *d* | *e* | |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 1 | 1 |
| 1 | 1 | 0 | 1 | 1 | 2 |
| 1 | 0 | 0 | 1 | 1 | 3 |
| 0 | 1 | 1 | 0 | 0 | 4 |
| 1 | 0 | 0 | 1 | 1 | 5 |
| 0 | 1 | 1 | 0 | 0 | 6 |

we shall find in the end ten two-element combinations which do not occur in *F*, and consider them as syllogistic regularities. They can be presented by the following ternary knowledge matrix *D*:

| | *a* | *b* | *c* | *d* | *e* |
|---|---|---|---|---|---|
| | 0 | 0 | - | - | - |
| | 0 | - | - | 1 | - |
| | 1 | - | 1 | - | - |
| | 1 | - | - | - | 0 |
| *D* = | - | 0 | 1 | - | - |
| | - | 0 | - | 0 | - |
| | - | 0 | - | - | 0 |
| | - | - | 1 | 1 | - |
| | - | - | 1 | - | 1 |
| | - | - | - | 1 | 0 |

When the selection *F* is noticeably bigger compared with the number of attributes, the maximum rank $r_{max}$ of implicative regularities could be 3, 4 or even more. The run-time for their finding swiftly increases. Nevertheless it is restricted, because the number of intervals to be checked could be approximated by $C_n^3\, 2^3$, $C_n^4\, 2^4$, etc.

It is a little more difficult to extract knowledge from the space of *n* multi-valued attributes $x_1, x_2, \ldots, x_n$ [9, 11]. To begin with, define the probability *p* that some concrete disjunct will be satisfied by an accidentally chosen element of the space. It could be calculated by the formula

$$p = 1 - \prod_{i=1}^{n} (r_i / s_i),$$

where $s_i$ is the number of all values of the attribute $x_i$, and $r_i$ – the number of those of them which do not enter this disjunct. For instance, $p = 1 - 2/2 \times 3/4 \times 1/3 = 3/4$ for the disjunct 00.1000.101. Let us divide all disjuncts into classes $D_j$, forming them from disjuncts with the same value of *p*. And let us number these classes in order of increasing *p* and introduce the following conventional signs: $q_j$ – the number of disjuncts in the class $D_j$, $p_j$ – the value of *p* for elements from $D_j$.

Find now the mathematical expectation $E_j$ of the number of disjuncts from the class $D_j$, which do not contradict the random *m*-element selection from the regarded space:

$$E_j = q_j (p_j)^m,$$

and introduce the analogous quantity $E_k^+$ for the union of classes $D_1, D_2, \ldots, D_k$:

$$E_k^+ = \sum_{l=1}^{k} E_j.$$

Inductive inference is performed by consecutive regarding classes $D_j$ in order of their numbers and summarizing corresponding values $E_j$ until the sum surpasses a threshold $\omega$, which is introduced with taking into account the specific of the problems to be solved. All disjuncts belonging to these classes are accepted as regularities if they do not contradict the data, i. e. if they are satisfied by any element of the selection *F*.

The expert may fix several thresholds and assign accordingly different levels of plausibility to the found regularities. For example, regularities obtained by thresholds $10^{-10}$, $10^{-6}$, $10^{-3}$ could be estimated as *absolutely plausible*, *usually*, *most likely*. This differentiation gives some flexibility to recognition procedures. Choosing a proper level of plausibility one can use only some of regularities contained in the knowledge base and vary in such a way the plausibility of the logical conclusions obtained during recognition. For example, using only the most plausible regularities can result in obtaining a little number of logical conclusions, but more reliable ones, while extending the used part of the knowledge base extends also the set of obtained logical conclusions, at the expense of their plausibility.

We do not regard here the important problem (touched in [12]) of extracting knowledge from partial data – when values of some attributes of the elements from *F* remain unknown.

## SOLVING EQUATIONS OF DEDUCTIVE INFERENCE

The recognition problem can be regarded as the problem of a closer definition of qualities of some observed object not belonging to the experimental selection from the subject area [5, 14]. It is formulated in terms of logical equations, Boolean or predicate, and the tree searching technique of deductive inference is applied for their solution [4, 10, 13].

Suppose, we know the values of *s* from *n* attributes of this object. That is equivalent to location of the object in a certain interval of the Boolean space *M* presented by the corresponding elementary conjunction *k* of the rank *s*. The problem is to define by logical reasoning, as sure as possible, the values of the remaining $n - s$ attributes, using for that the information contained in the knowledge ternary matrix *D* and in the corresponding veto function *V*.

Let us regard the set $X_k$ of attributes with known values and the set of all forbidden combinations of values of the remaining attributes − for the considered object. The latter set can be described by a proper Boolean veto function $V(k)$ that could be easily obtained from *V*. Indeed, it is sufficient for that to transform the formula representing the function *V* by changing symbols of attributes presented in *k* for values (0 or 1) satisfying the equation $k = 1$. Denote this operation as $V(k) = V{:}k$.

Suppose that we want to know the value of an attribute $x_i$ which does not come into $X_k$. The necessary and sufficient condition for the prohibition of the value 1 of that attribute is presented by the formal implication $kx_i \Rightarrow V$, i. e. belonging of the interval presented by the conjunction $kx_i$ to the prohibition region described by the function *V*. Analogously, the necessary and sufficient condition for the prohibition of the value 0 is presented by $kx_i' \Rightarrow V$.

It is not difficult to deduce from here forecasting rules to define the value of the goal attribute $x_i$ of the object characterized by *k*. These rules are shown in a compressed form in Table 3 presenting the decision (a set of possible values of $x_i$ – the bottom row) as a function of predicates $kx_i \Rightarrow V$ and $kx_i' \Rightarrow V$.

Table 3. Forecasting the value of the attribute $x_i$

| | | | | |
|---|---|---|---|---|
| $kx_i \Rightarrow V$ | 0 | 0 | 1 | 1 |
| $kx_i' \Rightarrow V$ | 0 | 1 | 0 | 1 |
| $x_i$ | {0, 1} | {1} | {0} | $\varnothing$ |

Note that four outcomes could appear at this approach. On a level with finding the only value (0 or 1) for the attribute $x_i$, such situations could be met when both values are acceptable or neither of them satisfies the veto function *V*. At the last case the existence of the object $\alpha$ characterized by *k* contradicts the knowledge base, and that could stimulate some correction of the latter. However, the probability of such an event is low enough, taking into account the way of forming the knowledge base.

For example, if

$$V = acf' \vee be'f \vee a'd'e \vee b'df \vee b'c'd'$$

and $k = abf$, then $V(k) = V{:}abf = e'$. It could be concluded from this that the regarded object $\alpha$ has value 1 of attribute *e*, but there are no restrictions on other attributes (*c* and *d*). If by the same function *V* the object $\alpha$ is characterized by $k = c'e'f$, then

$$V(k) = b \vee b'd \vee b'd' = 1 \text{ (all is forbidden)},$$

and that means that the object contradicts the knowledge.

The predicates $kx_i \Rightarrow V$ and $kx_i' \Rightarrow V$ are accordingly equivalent to the predicates $V{:}kx_i = 1$ and $V{:}kx_i' = 1$, and that allows us to reduce their calculation to checking corresponding submatrices of the knowledge matrix $D$ for consistency. Fixing values of some attributes in the function $V$ is changed for selecting the corresponding minor of the matrix $D$ by deleting some rows and columns, which could be followed by further possible simplification.

Suppose, we regard the same (already minimized) knowledge matrix $D$ corresponding to the veto function $V = acf' \vee be'f \vee a'd'e \vee b'df \vee b'c'd'$ and know that for the observed object $a = 1$ and $c = 1$. Taking into account this new information we transform the matrix $D$. We delete from it the columns marked with $a$ and $c$ because these variables became constant, and delete also the rows 3 and 5 now satisfied by these constants. Further simplification is rather evident, using the following rule: $x\,(x' \vee H) = x\,H$, where $x$ is a Boolean variable and $H$ – an arbitrary Boolean formula.

$$
D^* = \begin{array}{cccccc|c} a & b & c & d & e & f & \\ \hline 0 & - & 0 & - & - & 1 & 1 \\ - & 0 & - & - & 1 & 0 & 2 \\ 1 & - & - & 1 & 0 & - & 3 \\ - & 1 & - & 0 & - & 0 & 4 \\ - & 1 & 1 & 1 & - & - & 5 \end{array}
\rightarrow
\begin{array}{cccc|c} b & d & e & f & \\ \hline - & - & - & 1 & \\ 0 & - & 1 & 0 & \\ 1 & 0 & - & 0 & \end{array}
\rightarrow
\begin{array}{cccc|c} b & d & e & f & \\ \hline - & - & - & 1 & \\ 0 & - & 1 & - & \\ 1 & 0 & - & - & \end{array}
$$

We can conclude now that $f = 1$, by necessity. As to the remaining attributes, their values cannot be forecasted uniquely. They obey the next two conditions: $b' \vee e = 1$ and $b \vee d' = 1$. This system of logical equations has two solutions. Either $b = d = 0$ (with an arbitrary value of $e$), or $b = e = 1$ (with an arbitrary value of $d$).

Solving the recognition problem in a special case, when the values of all attributes are known except the goal one, could be facilitated by preliminary partitioning the Boolean space of attributes into four regions. After that it would be sufficient only to find out to which of them the regarded object belongs and make the corresponding conclusion.

The characteristic Boolean functions of these regions are obtained on the base of the rules shown in Table 3. The operations $f : x_i$ and $f : x_i'$ changing the argument $x_i$ of the function $f$ for constant 1 or 0, accordingly, are used by that. The region where the value of the attribute $x_i$ remains unknown is described by the function $V^{-}(x_i) = (V : x_i)' \wedge (V : x_i')'$, the region where $x_i$ receives the value 1 is presented by the function $V^{1}(x_i) = (V : x_i)' \wedge (V : x_i')$, the region where $x_i$ receives the value 0 – by the function $V^{0}(x_i) = (V : x_i) \wedge (V : x_i')'$, and the region of contradiction − by the function $V^{*}(x_i) = (V : x_i) \wedge (V : x_i')$.

Using the same example we obtain:

$V = acf' \vee be'f \vee a'd'e \vee b'df \vee b'c'd'$,

$V : f = be' \vee a'd'e \vee b'd \vee b'c'd'$,

$V : f' = ac \vee a'd'e \vee b'c'd'$,

$V^{-}(f) = (be' \vee a'd'e \vee b'd \vee b'c'd')' \wedge (ac \vee a'd'e \vee b'c'd')'$,

$V^{1}(f) = (be' \vee a'd'e \vee b'd \vee b'c'd')' \wedge (ac \vee a'd'e \vee b'c'd')$,

$V^{0}(f) = (be' \vee a'd'e \vee b'd \vee b'c'd') \wedge (ac \vee a'd'e \vee b'c'd')'$,

$V^{*}(f) = (be' \vee a'd'e \vee b'd \vee b'c'd') \wedge (ac \vee a'd'e \vee b'c'd')$.

## DEDUCTIVE INFERENCE IN FINITE PREDICATES

In the case of multi-valued attributes the disjunctive knowledge matrix $D$ turns out to be a sectional Boolean matrix presenting some finite predicate. There is some specific in dealing with it [7, 10].

Let us state the central problem of deductive inference: a disjunctive matrix $D$ and a disjunct $d$ mated with $D$ (that means defined on the same pattern) are given, the problem is to find out whether $d$ is a logical consequence of $D$. In other words, the question is if the conjunctive term $d$ is derived from CNF $D$, that means it becomes equal to 1 on all elements of the space $M$ where CNF $D$ takes value 1?

Two ways for solving such problems are known: the direct inference and the back inference.

When the direct inference is executed, the initial set of disjuncts is expanded consecutively by including new disjuncts following from some pairs of disjuncts existing already in the set. This procedure continues until the disjunct $d$ is obtained or the set expansion is exhausted without finding $d$. At the last case it is proved that $d$ does not follow from $D$.

Any pair of disjuncts $u$ and $v$ can generate several disjuncts-consequents $w_i$, obtained formally by the operation $w_i = u<x_i>v$ which may be called the *resolution in regard with the variable* $x_i$ and which can be considered as the generalization of the resolution operation, well-known in the theory of Boolean functions, onto finite predicates. It is defined as follows: the domain (section) of $w_i$ corresponding to the variable $x_i$ equals the component-wise conjunction of the corresponding domains from $u$ and $v$ (this can be considered as the unification by the variable $x_i$), and the rest domains equal the component-wise disjunction of the corresponding domains from $u$ and $v$.

But far not all disjuncts obtained in such a way deserve subsequent consideration. There is no sense in including into the regarded set a disjunct which follows from some other disjunct belonging to the set, because it represents only some expansion of the latter one. For example, disjunct 110.0111.00 follows from disjunct 010.0110.00. It is reasonable to look only for non-trivial consequents. Such is a disjunct which follows from some pair of disjuncts $u$ and $v$ but does not follow from $u$ or $v$ taken separately. Let us call it a *resolvent* of disjuncts $u$ and $v$, and determine the rules for its obtaining.

Disjuncts $u$ and $v$ are called *adjacent in regard to the variable* $x_i$ if and only if the corresponding domains are incomparable (their component-wise disjunction differs from each of these domains) and there exists in each of the rest domains a component with the value 0 in both vectors. Note that at violating the first condition a disjunct is obtained which follows either from $u$ or from $v$, whereas at violating the second condition a trivial (identical to 1) disjunct is found, which follows from any other disjunct.

*Affirmation 1*. If disjuncts $u$ and $v$ are adjacent in regard to the variable $x_i$ and $w = u<x_i>v$, then the disjunct $w$ is a resolvent of the disjuncts $u$ and $v$.

For example,

$$\begin{array}{rc} & a\quad b\quad c \\ u = & 100.10.0011 \\ v = & 010.00.0110 \end{array}$$

It is easy to see that these disjuncts are adjacent in regard to $a$ and also to $c$, but not to $b$. Hence, they give rise to the following two resolvents

$$\begin{array}{l} u<a>v = 000.10.0111 \\ u<c>v = 110.10.0010 \end{array}$$

The direct inference is simple but time-consuming because the number of obtained consequents could be very large. The back inference is more efficient. It solves the problem by transforming the initial system of disjuncts into such a system which is consistent if and only if $d$ does not follow from $D$. So, the problem is reduced to the regarded above problem of checking some disjunctive matrix for consistency.

Denoting by $\neg d$ the vector obtained from $d$ by its component-wise negation, and by $D \wedge \neg d$ the matrix obtained from $D$ by the component-wise conjunction of each of its rows with vector $\neg d$, the following rule may be formulated.

*Affirmation 2*. A disjunct $d$ follows from a disjunctive matrix $D$ if and only if the disjunctive matrix $D \wedge \neg d$ is not consistent.

Checking this condition is rather easy: 1s are expelled from all columns of $D$ which correspond to components of the vector $d$ having value 1, then the obtained disjunctive matrix is checked for consistency.

For instance, if

$$D = \begin{array}{l} 001.0010.00 \\ 110.0011.01 \\ 010.1100.10 \\ 001.0100.00 \end{array},$$

$$d = 011.1000.00,$$

then the following disjunctive matrix should be checked for consistency

$$D \wedge \neg d = \begin{matrix} 000.0010.00 \\ 100.0011.01 \\ 000.0100.10 \\ 000.0100.01 \end{matrix}$$

This matrix is not consistent, hence the disjunct $d$ follows from $D$.

## CONCLUSION

Implicative regularities were proposed to fix the knowledge extracted from data on the stage of inductive inference and to play the role of axioms on the stage of deductive inference, when the values of goal attributes should be forecasted. A set of algorithms was developed to implement those operations. The suggested means were used when constructing several expert systems of various purposes where the pattern recognition problem was the central one. The computer experiments testified the high efficiency of the proposed approach.

## Bibliography


1. Bongard M. Pattern recognition. – Spartan Books, New York, 1970.
2. Chang C. L., Lee R. C. T. Symbolic logic and mechanical theorem proving. – Academic Press, New York – San Francisco – London, 1973.
3. Frawley W. J., Piatetsky-Shapiro G., and Matheus C. J. Knowledge discovery in data bases: an overview. – In: Knowledge discovery in data bases (ed. by Piatetsky-Shapiro and Frawley), Cambridge, Mass: AAAI/MIT Press, 1991, pp. 1-27.
4. Nilsson N. J. Problem-solving methods in artificial intelligence. – McGraw-Hill Book Company, New York, 1971.
5. Zakrevskij A. D. Revealing of implicative regularities in the Boolean space of attributes and pattern recognition. – Kibernetika, No 1, 1982, pp. 1-6 (in Russian).
6. Zakrevskij A. D. Logic of recognition. – Minsk: Nauka i tekhnika, 1988 (in Russian).
7. Zakrevskij A. D. Matrix formalism for logical inference in finite predicates. – Philosophical bases of non-classical logics. Institute of Philosophy AN SSSR, Moscow, 1990, pp. 70-80 (in Russian).
8. Zakrevskij A. D., Levchenko V. I., Pechersky Yu. N. Instrumental expert system EDIP, based on finite predicates. – Applied systems of artificial intelligence. Mathematical researches, issue 123, Inst. Math. AN Mold. SSR, Kishinev, 1991, pp. 24-40 (in Russian).
9. Zakrevskij A. D. EXSYLOR – expert system for logical recognition. – Upravlyayushchie sistemy i mashiny, 1992, No 5/6, pp. 118-124 (in Russian).
10. Zakrevsky A. D. Logical recognition by deductive inference based on finite predicates. – Proceedings of the Second Electro-technical and Computer Science Conference ERK'93, Slovenia Section IEEE, Ljubljana, 1993, v. B, pp. 197-200.
11. Zakrevsky A. Logical recognition in the space of multivalued attributes. – Computer Science Journal of Moldova, 1994, v. 2, No 2, pp. 169-184.
12. Zakrevskij A. D., Vasylkova I. V. Inductive inference in systems of logical recognition in case of partial data. – Proceedings of the Fourth International Conference on Pattern Recognition and Information Processing, Minsk-Szczecin, May 1997, v. 1, pp. 322-326.
13. Zakrevskij A. D. Pattern recognition as solving logical equations. – Special Issue 1999 – SSIT'99 (AMSE), pp. 125-136.
14. Zakrevskij A. D. A logical approach to the pattern recognition problem. – Proceedings of the International Conference KDS-2001 "Knowledge – Dialog – Solution", S.- Peterburg, June 2001, v. 1, pp. 238-245.


## Author information


Arkadij Zakrevskij - The United institute of informatics problems of the NAS of Belarus. Surganov Str. 6, 220012 Minsk, Belarus. e-mail: zakr@newman.bas-net.by

# MULTI-AGENT INFORMATION PROCESSING AND ADAPTIVE CONTROL IN GLOBAL TELECOMMUNICATION AND COMPUTER NETWORKS

*A.V.Timofeev*

*Abstract: The problems and methods for adaptive control and multi-agent processing of information in global telecommunication and computer networks (TCN) are discussed. Criteria for controllability and communication ability (routing ability) of dataflows are described. Multi-agent model for exchange of divided information resources in global TCN has been suggested. Peculiarities for adaptive and intelligent control of dataflows in uncertain conditions and network collisions are analyzed.*

*Keywords: adaptive and intelligent control of dataflow, models for multi-agent processing and information transfer, telecommunication and computer networks.*

## Introduction

The main aim of control for global telecommunication and computer networks (TCN) is a fast search and delivery (transportation) of necessary information to TCN users-agents with high quality of submitted services. So the general problem of control and multi-agent processing of information in TCN may be divided on four interconnected subproblems:

- control for dataflows between TCN agents with adaptation to different kinds of heterogeneous traffic;
- organization for multi-agent dialogue between TCN agents;
- telecommunication equipment control;
- administrative control for productivity and configuration of TCN.

In global TCN (for example, Internet) control for transfer and distribution (routing) of dataflows between agents should be done not on severe program but “in loose” on changing unpredictably users or nodal TCN components requests.

Traditional approach to TCN control does not provide interactivity in real time (for example, speech dialogue without relays) often. The other disadvantages are non-adaptivity (by relation to changing traffic) of control for dataflows, impossibility for automatic avoiding of network conflicts, fault recognition and TCN reconfiguration without human (administrator or network operators) participation.

## 1. Problems of Adaptive Control and Multi-Agent Processing in TCN

At design and exploitation of modern TCN important role is given to theory and tools of control for flows of transferred data. However today theory of TCN control (differently from, for example, theory of automated control for motions of aircraft, robots and the other mobile objects) has been developed weakly.

Therefore necessity in setting, formalisation and solution of problems for control, processing and information transfer in TCN in dialogue mode in uncertain conditions arises.

Specific character of TCN as control object is in distributed character of TCN components and controlled flows of different (heterogeneous) data, transferred through TCN nodes along different routes and communication channels. Consequently TCN control system should be distributed and have multi-level hierarchical structure.

Specific aims and problems of control arise on every level of this system. However many of those aims are not formalized and problems are not solved (it means that theoretically proved models of “collective dialogue and control algorithms” do not exist for these problems). So in the paper significant attention is devoted to setting and solution of problems for organization of multi-agent dialogue and dataflows control in uncertain and network conflicts conditions.

Uncertainty of TCN exploitation conditions is in unpredictable character of change and in heterogeneity of network traffic. The number of TCN users-agents is also uncertain and change significantly during the day. TCN may have misworks or faults of separate components as well as different kind of network conflicts. So it is necessary to have adaptation for traffic, monitoring and diagnosis of TCN states and also classification and resolution of network conflicts.

Noted TCN specifities require investigation and development for adaptive approach to solution for problems of “collective” dialogue and dataflows in TCN on the base of modern intelligent and multi-agent technologies. Suggested methods for adaptive and intelligent control provide adaptation to traffic, changing unpredictably, adaptive routing of dataflows, multi-agent information processing, TCN functional diagnosis, recognition and resolution of network conflicts [Timofeev A.V.,1996,2001,a,b,2002].

A wide class of complex distributed TCN may be presented as multi-agent system (MAS). Here users of TCN or nodal computers or local TCN as TCN segments work as TCN agents.
Typical features of these TCN agents are existence of local databases and knowledge bases and telecommunicational channels for information exchange between agents in the process of mutual use of distributed informational resources and processing of transferred information.
The main peculiarity of multi-agent processing of information and control is that firstly complex problem is decomposed (fragmented) on the series of local problems, which solution is distributed (disparallel) between agents, and then results of solution of these local problems are agregated (integrated) and realized with the help of telecommunicational resources.
Information processing MAS work is supported by TCN, realized network technologies of data transfer between agents. Communication between remote agents is executed on the level of their local databases and knowledge bases by the way of controlled message exchange in the process of solution for local (individual) or general (global) problems.
Significant theoretical and practical interest is given to two strategies for multi-agent processing and information transfer:
- with coordinator (when one of the agents is responsible for coordination of behaviour of all other agents);
- without coordinator (when all agents have “equal rights” and do not submit to leading agent-coordinator).
At multi-agent control for dialogue and dataflows in TCN it is necessary to develop the methods for automatic avoidance or resolution for conflicts, which may occur between TCN agents.
In connection with it a significant value belongs to multi-agent models and algorithms for information processing (code replication, data fragmentation, adaptive routing etc.)
At design for TCN dataflow control systems a reliability of used equipment plays important role. Reliability of global TCN is lower if TCN consists of more nodal computers. It is explained that with increased number of TCN nodes the probability for failure of one or several computers increases. That is why necessity for adaptive control and multi-agent processing for information in TCN, that guarantee problem solution at unpredictable traffic change, fault or failure of one or several nodal TCN computers, appears.

## 2. Global Controllability and Dialogue Possibilities of TCN

The main requirement for global TCN is the possibility to provide the main function - to provide for real and potential users a controlled access to distributed informational, computing and telecommunicational resources of all nodal computers or local TCN, united in a global network. To satisfy this requirement TCN, as complicated dynamics object for informational control, should be globally controlled.
Modern control theory means global control of dynamics control object as existence for control, providing conversion of this object from any allowable initial state to any allowable finite (aimable) state for a finite time. If control object has control property, it means that there are one or several algorithms for allowable control, providing aim achievement. Otherwise it is generally impossible because corresponding algorithm does not exist simply.
In connection with TCN a global controllability means a possibility for access from every nodal computer in arbitrary time moment, called as initial, to informational and telecommunicational resources of any other TCN computer for a finite time. At it a controlled access of user to existing distributed resources is provided by automatic planning of one or several computers, connecting user’s computer with aim computer, which contains requested resources and transfers them to user in the form of informational dataflow.
However it should be note that TCN literature (see, for example, [Olifer V.G., 2001]) means network controllability as “a possibility to control centralizely a state of main network element, find and solve problems, arising at a network work, execute an analysis for productivity and plan a network development”.
Such definition of TCN controllability is too general and fuzzy. It mixes in fact property for controllability of TCN as dynamics control object with requirements to TCN control systems, modes of their work, TCN fault-stability, their productivity etc.
Usefullness of concept of TCN global controllability and necessity for automatic routing, control for information dataflows and dialog organization appears most brightly in complex corporative and global (public) networks. However nowadays TCN control area has many unsolved problems. Consequently convenient, multi-protocol and multi-agent TCN control systems, able to adapt to heterogeneous traffic, unpredictably changing, and to solve automatically network conflicts have not been created yet.
Many modern TCN are controlled in fact by qualified men-administrators or network operators. Existing automatic tools do not control TCN but provide observation (monitoring) for their work on the base of change

of some important TCN operation indexes. Really they track only for rightness of TCN operation, form and store “history TCN exploitation” and partly control arising faults.

Existing systems for monitoring and information processing are not able in fact to control automatically for dataflows in global TCN, to adapt for changing traffic, to solve network conflicts, to diagnosis and eliminate faults and failures.

Modern tools for TCN automation provide control only for separate network elements, i.e. are tools for TCN local control. Functions for global control, observation and diagnosis of TCN belong to man (for example, to a network administrator or a network staff).

The most important characteristic for controlled global TCN is its productivity. Real productivity of TCN depends significantly on used control system and may approximate to a certain limit, which is naturally called as a potential productivity of TCN.

Productivity of global TCN is characterized by the following main indexes [Olifer V.G.,2001, Wallrand G.,2001]: TCN reaction time, TCN capacity and dataflow transfer delay. TCN reaction time is the duration T of time interval between the initial moment $t_0$ of users request to TCN and the finite moment $t_T$ of answer receiving on this request. The value of this index $T=T(\alpha_1, \alpha_2, \ldots)$ depends on a series of factors: $\alpha_1$ - the type of TCN service, requested by a user; $\alpha_2$ - the type of nodal server, requested by a user; $\alpha_3$ - the current state of TCN elements (load of segments, routers, commutators, through which a request passes, etc.); $\alpha_4$ - the daytime, when user is requesting TCN etc.

TCN reaction time is a global characteristic for TCN productivity with the point of an agent-user view. It is determined by the formula

$$T = \sum_{i=1}^{5} T_i \equiv t_T - t_0 , \qquad (2.1)$$

where $T_1$ - the time for preparation of request on a user computer, $T_2$ - the time for request between a user and nodal TCN server through communication equipment and TCN segments, $T_3$ - the time for request processing on TCN server; $T_4$ - the time for answer transfer from server to a user (client); $T_5$ - the time for processing of obtained reply on a user (client) computer

Additive decomposition of the reaction time $T$ on components $T_i$ is not interesting for a user. The finite result is interesting for him, i.e. minimal value of global TCN time reaction on his request in a dialogue mode.

Knowledge of the network local components $T_i$ of the reaction time T allows evaluating productivity of separate TCN elements, to find the most unproductive TCN elements and to minimize the global reaction time $T$ by control tools or by means of modernization of used TCN equipment.

Thus, a principal possibility for increasing of TCN productivity by optimization (on speed) of control appears. For example, it is possible to minimize $T_3$ by controlled choice of the shortest or fastest routes “client-server”, satisfying an informational user’s request. For it it is possible to use neural models for optimal routers.

TCN capacity $V$ is a volume of dataflow, transferred by TCN or its component for a time unit. Usually the value $V$ is measured in bits per a second or in package per a second. In contrast to TCN reaction time, depending particularly on a user, TCN capacity $V$ is an objective index of network productivity, characterizing a speed of dataflow transfer between TCN nodes through different communicational equipment. It is accepted to differ an average, instant and maximal TCN capacity [Olifer V.G., 2001].

Average TCN capacity $V_*$ is determined as a dataflow volume, transferred for a sufficient large time (for example, hour, day or week). Instant TCN $V_0$ - volume of dataflow, transferred for sufficient small time (for example, 1msec or 1sec). Maximum capacity $V_{max}$ is determined as the largest value of $V_0$ , fixed on all interval for observation of TCN work.

Among these indexes of TCN or its components capacity the most informative are the values $V_*$ and $V_{max}$ . Average TCN capacity $V_*$ evaluates TCN productivity on a sufficient large time interval, when unpredicted by them accidental “picks” and “decays” of traffic particularly compensate each other. Maximum capacity $V_{max}$ characterises TCN work ability at graph “picks” (for example, in the beginning of work time, when many users are registered simultaneously and requested for TCN).

Global TCN capacity $V$ depends on local capacity abilities $V_j$ of its components, measured, for example, between network nodes or input and output router ports. Because of consequent character of dataflow transfer by different TCN components a global capacity of any complex route of data motion in TCN will be equal a minimal of capacities of elements for this route. So to increase global TCN capacity it is necessary to increase a capacity of its the most slower components. It may be done, particularly, by means of TCN control

tools (for example, by way of routers optimization) on criterion for search of the shortest or fastest communicational routes.

## 3. Heterogeneity and Uncertainty of TCN Traffic

Traffic of messages, sent in telephone networks or in cable television networks, differs significantly from traffic of informational dataflows in global TCN of Internet type. These flows transfer not only files, DB etc., but also multi-media data, presenting in a digital form images and speech. Exactly so global TCN are used more widely for teleconference holding, remote learning etc.

For controlled transfer of multi-media information it is necessary to have not only a special equipment but and new protocols and control algorithms, providing adaptation to changing multi-media traffic. The matter is in that sound oscillations and light waves are continuous processes. Therefore for their high-quality reproduction it is necessary to measure, to code and transfer them sinchro that significant distortions and delays would not appear.

Traditional traffic for data, transferred by TCN, may change unpredictably in wide limits. Really this traffic has "pulsating" character and unknown beforehand intensity. For example, users request in DB of remote computer generates a datalow between its local and remote computer, depending on many factors (text edition etc.), and its insignificant delay does not depend practically on a quality of TCN users service.

However at traffic change in wide limits a service quality may significantly worsen. Therefore necessity for adaptation of TCN control system to unknown traffic, changing in wide limits unpredictably, appears.

The other reason of appearance of necessity for adaptive control is connected with that in global TCN both usual dataflows and multi-media information should transfer. It means that real traffic for global TCN is usually uncertain and heterogeneous. Therefore systems for control of global TCN dataflow should adapt to unknown beforehand specifities of heterogeneous traffic and provide a high quality of service for users, which structure can also change unpredictably.

## 4. Decentralized Control of "Client– Server" Dialogue

Among computer networks global TCN appeared first by chronology and united between each other computers, distributed on different cities and countries and providing possibility of controlled information change on a big distance.

At design of global TCN the main ideas, mechanismes and technology of modern computer network operation were suggested and completed. Principles of multi-level building of telecommunicational protocols, routing technology and information packages communication etc. are related to them [Wallrand G.,2001,Timofeev A.V.,2002]. These principles and technologies found realization in the most popular TCN – Internet and continue to perfect theirselves with consideration of new needs of informational society.

Global TCN belong to distributed computer networks with decentralized control for local centres (nodes) of processing and transfer of information on the base of computers with network adapters, connected between each other by communicational channels. For local control of computers in a mode of request for its resources it is necessary to have programme modules – servers, constantly waiting and serving constantly for requests of global TCN users on resource access.

From the other side, for local control of computers in mode of request forming to resources of TCN remote computers it is necessary to have programme modules - clients, moderating and transferring requests to address of necessary computer of TCN.

Pairs of modules "client-server" are important part of network operational system. They provide controlled joint access to certain type of remote informational resources (for example, to files).

A set of those pairs (on all types of resources) makes a network service, controlling information exchange between computer-server, presenting its resources to the other computers of TCN, and computer-client, using these resources. The same TCN computers may play simultaneously both server and client.

Control system for global TCN consists of network operational system, control system for local and distributed resources, named a network service, and control systems for informational flows, connected with so named network applications (network DB and KB, systems for data archiving, mail systems, systems for network conflicts resolution etc.).

TCN as a control object is characterized by topology of physical links between controlled TCN computers. As topological model for global TCN we will name a graph for TCN network configuration, nodes of which are

corresponded by computers (or local computer networks), and edges - by physical (electric) connections between them. Computers (or local networks) of global TCN are named as network nodes.
Optimization of topological TCN models is executed usually (because of economic reasons) on criterion of minimization of total length of physical communication channels. However for improvement of TCN reliability and balancing of separate channels traffic (load) redundant (reserve) communicational channels are introduced sometimes.
In local TCN with a small number of computers (several decades) topological model is homogeneous and corresponds to one of typical network configurations: "ring", "general bus", "star", "cellular" and "full-connectional [Wallrand G.,2001].
Global TCN are created on the base of connection of remote local TCN, which play role of big TCN. Therefore global TCN are decomposed on homogeneous local TCN, having one of typical configurations and their topological model is mixed (heterogeneous).

## 5.Multi-Agent Model for Controlled Exchange of Divided Information Resources of TCN

Let describe the set of agents-users of global TCN : $A=\{a_1,a_2,\dots a_N\}$, (5.1)
consisting of N≡|A| agents.
Let present a model for distributed information resource, allowable for an every agent from the set A, in the kind of the vector : $I=|I_1,I_2,\dots,I_M|$. (5.2)
The vector (5.2) is in geometry as a point in M-dimension Euclide information space $R^M$. Components of this vector are the values of the quantity $I_j$ of information of j-th type, connected with concrete types of data and knowledge in distributed system of local DB and KB of agents.
Let give to an every agent $a\in A$ in accordance the vector of its personal (local) information resources

$$I(a)=|I_1(a),I_2(a),\dots,I_M(a)|, \quad (5.3)$$

which components are the values of quantity $I_j(a)$ of local information of j-th type, accessible for this agent. Let note, that if $I_j(a)=0$, it is meant that an agent has no access for information of j-th type or it is equal zero.
It is suggested that agents from the set A may contact each other through telecommunication channels and unite in $A_K \subset A$ group (brigade) work to use jointly an accessible information. Here the agent $a_i \in A_K$ may increase or maximize its information state, consisting of local (single-agent) information $I(a_i)$ and group information state $I(a_j)$, $a_j \in A_K$, $j\neq i$, accessible for other agents of the group $A_K$.
Quantity of interconnected agents of the group $A_K$ is written as $N_K = |A_K|$.
Then the every agent $a_i \in A_K$ has in the limits of one tact of communication simultaneous (parallel) access to local information $I(a_j)$ of other agents from the group $A_K$ with the help of $N_{k-1}$ communication channels. Thus, every agent from work group $A_K$ has two-side information communications with $N_{k-1}$ partners from $A_K$.
Every agent $a_i$ has a right to request from other agents from the group $A_K$ only that type of the information $I_j(a)$, which it is interested in. At it an agent understands (knows), that it is preferable for it to connect with such agents that have bigger information on interested for it types of transferred information, i.e. on determined components of vectors of local (single-agent) information (5.3).
Using global TCN, agents try to maximize their information state (knowledge state) by information resources of other accessible for it agents. So such agents create information unions (groups, brigades) on interested for them information types.
Determine information interests (requests) of the agent $a\in A$ with the help of binary vector of interesting in information of the different types

$$R(a)=|R_1(a), R_2(a),\dots,R_M(a)|, \quad (5.4)$$

which components are predicates of the interests of the kind

$$R_j(a) = \begin{cases} 1, \text{if the agent a is interested in j-th type of information,} \\ 0-\text{otherwise,} \end{cases} \quad (5.5)$$

where $j=1,2,\dots,M$. For example, if $R_1(a) = R_2(a)=\dots=R_{M-1}(a)=0$, and $R_M(a)=1$, then information interests of the agent a are reduced only to the information of the M-th type, and other information is not interested for it.
Degree of interest of the agent $a \in A$ in distributed information resources (5.2) of other agents in the simplest case may be characterise by the value

$$\|R(a)\| = \sum_{k=1}^{M} R_k^2(a) = \sum_{k=1}^{M} R_k(a). \quad (5.6)$$

Digital index of interest of the agent (5.6) is Euclide norm of the vector (5.5) and is Hemming distance in $2^M$-dimension binary space of information interests (request) of agent. At $||R(a)|| = 0$ the agent *a* has no interest in any accessible information and at $||R(a)|| = M$ the agent *a* is interested in all types of accessible information, which other TCN types have.

Let name as $W_{ij}$ a degree of the agent $a_i$ wish to request the agent $a_j$ and let consider that $W_{ij} \in [0,1]$. Then necessary and sufficient condition of two-side wish of the agents $a_i$ and $a_j$ to communicate (request information from each other) has the kind

$$W_{ij} W_{ji} > 0, \; i \neq j. \qquad (5.7)$$

At it every agent $a_i$ may request and obtain information of any type from other agent $a_j$. If $W_{ij} = 0$, it means that the agent $a_i$ has no desire or denies to request information at the agent $a_j$.

The case, when $W_{ij} = W_{ji} = 1$ means, that between the agents $a_i$ and $a_j$ there is an open channel for two-side communication and information request of all types is possible.

It is important to consider also for every agent the right for "veto", i.e. possibility to refuse to divide definite types of information with other agents. Need in it may be occur, for example, in the cases, when information of determined type is a secret and is not allowed for transfer through open TCN communication channels.

Let name as $n=n(a)$ quantity of telecommunication channels of the agents $a \in A$, by which it is communicated with other agents to have a possibility for access to distributed informational resources of these agents.

In ideal case global TCN gives a possibility for every agent $a \in A$ to connect with all other agents from A, i.e.

$$n=n(a)=|A|-1= N-1, \; a \in A. \qquad (5.8)$$

In the case of the group $A_K$ from $N_K = |A_K|$ of interconnected agents

$$n_k=n(a)=| A_K|-1= N_k -1, \; a \in A_k. \qquad (5.9)$$

In this case each agent $a \in Ak$ has a possibility of access only to $N_k - 1$ agents from a set. For example, the set can consist of all agents interested in k-th type of an information, i.e.

$$A_k = \{a: R_k(a)=1 \}. \qquad (5.10)$$

Global TCN allows its agents to share for any part (the type of information) from divided information resource or all its local information (5.3) with other agents.

Let name as $p(a)$ this part of the information type (5.3), which agent $a \in A$ is ready to share (give) to other agents. Obviously $p(a) \in [0,1]$. At $p(a)= 0$ the agent *a* hides (makes as a secret) all its local information (5.3) from other agents. At $p(a)= 1$ all components of the vector (5.3), i.e. all types of stored information $I_j(a)$, $j=1,2,\ldots,M$, are open for requests of other agents.

## 6. Singularities of Adaptive and Intelligent Control for Informational Flows in TCN

Complexity of modern global TCN, providing access to enormous informational and computational resources is that in process of their control a central role is played by men-professionals in the area of informational and telecommunicational technologies. However on specialist evaluations, if such TCN, as Internet, will grow so fastly as it happens in the beginning of XXI century, in 10 years 200 million of people will be necessary for control and service for Internet–users. Therefore a sharp necessity for automation of control processes in global TCN on the users requests appears.

Solution of this problem is difficult because there are no formalized models of basic components and processes, happening in TCN, changing conditions of TCN exploitation are not learned etc. Without overcoming of these difficulties it is impossible to develop a theory for automatic control of TCN with adaptation to changing traffic of tools for automatic design for control systems and communicational equipment in TCN of a new generation.

Access of distributed informational and computational resources for TCN users is done in conditions of uncertainty and non-stationarity, i.e. at lack of information about current state of TCN and constantly changing informational environment.

The major factors of uncertainty and non-stationarity are:

1) uncertainty or unpredictable changes of heterogeneous traffic in TCN;
2) unpredictable changes of TCN users number;
3) apriori unpredictable character of agent-users request, connecting with their current local interest;
4) possibility of network conflicts in TCN etc.

In these real conditions of uncertainty and non-stationarity a necessity for robust, adaptive and intelligent control appears. It is a matter that precisely robust and adaptive control systems are possible to provide right

operation of TCN in beforehand uncertain and unpredictable changing conditions by signals of feedback about TCN current state and compensation of arising factors of uncertainty [Timofeev A.V.,1996].

Intelligent control systems, inheriting properties of robustness and adaptivity, have additional functions of artificial intelligence. In other words, they can solve autonomously (i.e. independently) some intelligent problems. For example, intelligent systems can diagnosis TCN states, recognize network conflicts and provide their avoidance or automatic resolution.

TCN with adaptive control will be named as adaptive TCN and TCN with intelligent control will be named as intelligent TCN. Such adaptive intelligent TCN behave as self-adjusting "clever" systems, able to overcome independently difficulties and to solve appearing problems in TCN before a user knows about them.

Such TCN, related to new generations of computer networks, have the following distinctive features:

constant accessibility and readiness to satisfy any (admissible) users requests, i.e. dialogue ability;

adaptability (adaptivity) to factors of uncertainty and ability to react "reasonably" on unpredictable events (traffic or users number changes, network conflicts etc.);

ability for autonomous reconfiguration and reconstruction in the case of faults and failures (for example, at disconnection or fault of some communication channels or TCN nodes) or unpredictable changes of external environment;

self-defence from possible threats and attacks, directed on work ability loss of TCN or its main components.

Since a global TCN consists of many local components, it is very important, that all controlling components are adaptive and intelligent. Then superposition and network integration of these components provides adaptivity and intelligence of TCN in a whole. Here it is appropriately to make an analogy of local and global TCN of a new generation with man or people team, which consist of many interconnected self-adjusting adaptive and intelligent subsystems, maintaining not only their "vitality" but also achievement by them local (particular) or global (collective) aims and needs [Timofeev A.V.,a,b,2001].

## Conclusion

In context of above described an interest is given to a new concept of IBM, stated in 2002 in the manifesto "Autonomic Computing: IBM Perspective on State of Information Technology". Searching for solution IBM Research turns its attention to control organization in alive nature. Any alive organism consists of many self-adjusting systems and subsystems. Elementary system with self-adjustment is named autonomic and superposition of such systems is named autonomic computing. Memorandum has been built as appeal for scientists of academic institutions and representatives of commercial companies to realize that "time of great changes" has become.

The work has been done under support of grant of Ministry for Industry, Science and Technologies of RF № 37.029.11.0027 and grant of RFBR № 03-01-0024 and grant of RHSF № 03-06-12019b.

## Bibliography

[Olifer V.G., 2001] Olifer V.G., Oliver M.A. Computer Networks. Principles, Technologies, Protocols. – Saint-Petersburg: Peter, 2001. – 672 p.

[Wallrand G., 2001] Wallrand G. Telecommunicational and Computer Networks. Introduction Handbook. - Moscow: Postmarks. - 2001. - 480 p.

[Timofeev A.V.,1996] Timofeev A.V. Intelligent Control Applied to Non-Linear Systems and Neural Networks with Adaptive Architecture.- Journal of Intelligent Control, Neurocomputing and Fuzzy Logic, 1996, pp. 1-18.

[Timofeev A.V.,2001,a] Timofeev A.V. Intelligent Control and Multi-Agent Navigation. - Proceedings of the IAR Conference and Associated IAR/ICD Workshop (Strasbourg, November 22-23, 2001), pp. 123-127.

[Timofeev A.V.,2002] Timofeev A.V., Syrtsev A.V. Neural Methods for Dataflow Optimal Routing. – Materials of the Conference "Regional Informatics-2002" (Saint-Petersburg, November 26-28, 2002),part 1,pp.87-88.

[Timofeev A.V.,2001, b] Timofeev A.V. Intelligent and Multi-Agent Control for Complex Systems. – In collection book "Extreme Robotics" (Materials of XI Scientific-Technical conference) – Saint-Petersburg: SPSU Publishing, 2001, p. 9-16.

## Author information

Adil Vasilievich Timofeev - Saint-Petersburg Institute for Informatics and Automation of RAS, 199178, Russia, Saint-Petersburg, 14-th line, 39; E-mail: tav@iias.spb.su

# INTERFACE ENGINEERING AND DESIGN: ADAPTIBILITY PROBLEMS[1]

*T. Gavrilova, E. Vasilyeva*

*Abstract: The paper describes some investigation's problems intended to construct user model via psychological assessment for further interface adaptation. The proposed concept of user model comprises formal representation of demographic, professional, physiological and psychological data about the user with stress put on psychological features. Essential users physiological, communicative and cognitive peculiarities are in the center of view in the research theoretical part. The presented research develops flexible user-centered approach as a unity of two main investigation directions – user modeling and adaptive interface's design. The adaptation engineering procedure is investigated via special interface modeling approach. The targeted implementation is distance learning process. Special software tool InterTrivium for interactive questionnaire presentation is described*

*Keywords: human-computer interaction, user model, adaptive interface, web-based learning system.*

## Introduction

The user interface development is one of the main stages in the computer software design process now. The main requirements for user interface are friendliness, comfort and easy studied. The branch of science called usability is working to build main principles of the user interface design now. Psychologists, ergonomists, computer specialists try to find out basic criterions of the user interface development.

In this paper a major research focus we concentrate on the problems of the human-computer interaction in the distance learning systems with the stress on individualised teaching. Web-based system design problems are actively discussing by the usability engineers [Nielsen, 1999]. The main idea of there work is to think out fundamentals of user interface design as "User Interface for all". This approach is not corresponding to the main psychologist's conception. Users are differing by their own characteristics and peculiarities such as age, education, psychological and cognitive peculiarities. The other approach to user interface design – developing of adaptive interfaces on the base of the user model - is supposed to be more fruitful for the user interface elaboration to such computer applications class as web-based systems.

The main reason of user model developing for the distance learning system is to adapt learning process. Use model may include the group of user characters that can influence the learning material generation, navigation via learning course parts, type of students testing. In this paper problems of the user model generation for the distance learning systems and the task of the interface adaptability in this class of the intelligent systems are discussed.

The paper discussed the first results of the current co-operative Russian-Byelorussian project intended to work out the methodology of distance learning systems interface adaptation and to develop special software tool to carry out interface adaptation. User and interface models' assembling for the distance learning system are the initial project milestones. The major distance learning interface properties able to adapt were combined into several special groups of the user interface model as well as key user features were arranged to the set of the user model groups. The preliminary propositions of the user and interface model's correlation were formulated to prove by the series of experiments. Special software tool InterTrivium was developed to provide the experimental part of research. It carries out the user questioning and user model design. Some single-purpose questionnaires were composed to determine user's interface preferences.

## Interface Model of the Web-based Learning System

The distance learning system interface model involves four groups of interface parameters (see Fig. 1). There are functional, interactive, service and lay out features of the distance learning user interface. More than 50 different interface characteristics were primarily selected to include to the interface model. The investigation of

[1] The research work reviewed in this paper has been carried out in the context of the Russian Foundation for Basic Research funded project "Adaptable Intelligent Interfaces Research and Development for Distance Learning Systems"(grant N 02-01-81019). The authors wish to acknowledge the co-operation with the Byelorussian partners of this project.

all interface parameters adaptation attainability is a very difficult task. Thereby several interface characteristics, which have the significant influence the human-computer interaction and education process, were selected for our research.

Functional interface parameters contribute greatly the system interaction behavior. It means difference in the representation of learning material and tests' performance. The examination of distance learning systems allowed to choose the following user functional parameters, which can be adapted:

- The set of the available working processes. This characteristic implies several distance learning system access modifications – for students, for administrator and for lecturer.
- The learning course material's structure. The user is allowed to study a certain limited suite of learning material according to his education, psychological peculiarities, interaction time and tests' results.
- Tests' content. Each user is provided by individual set of tests according to his user model.
- The navigation tool. The navigation tool adaptation supposes individual route for learning material study.

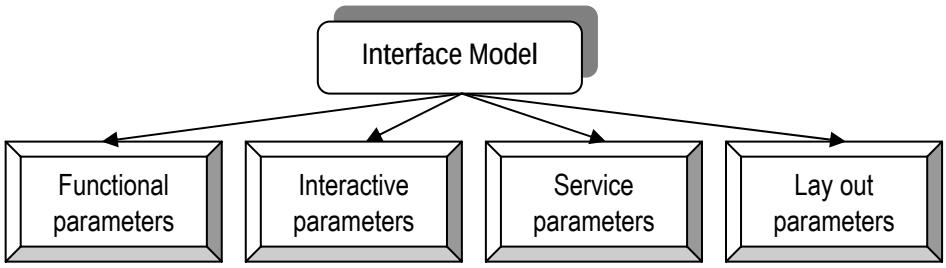



Fig. 1. Distance Learning System Interface Model Structure.

Interactive interface parameters determine the usability during the interaction with the system, the interaction scenario features. These characteristics group includes the information layout in the distance learning systems. it Is advisable to select the following parameters within this group:

- Dialogue type. The dialogue type management supposes providing different forms of tests.
- Controlling elements composition (menu configuration). The individual menu configuration should be formed for each user.
- Learning materials and tests' content being shown to the user at a time. The hypothesis of this parameter's adjustment according to the user's psychological peculiarities and his skills is put forward.
- Learning material performance format. Adaptation of this interface component assumes learning material presentation (font size, structure, graphic material availability) in the format most suitable to the user individual peculiarities.
- The highest possible hyperlink level (hypertext depth). According to the individual user model parameters' values the hyperlinks' number and hierarchy are generated to the distance learning system user.
- Navigation status. This characteristic's adaptation task is to follow the user navigation via learning material, to remind last visited page for starting with it during the next interaction with the system.

Service interface parameters include all objects participating in the reference and information dialog interface's functions. In our project we study adaptability of the reference information level, in other words we study individual user support system development feasibility.

Lay out interface parameters characterized information layout on the user display and level of user participation in it. In our project we tries to investigate adaptability of the following lay out interface parameters:

- Current window set-up. Distance learning system interface should be optimized to user screen sizes.
- Information lay out influences the learning material assimilation effect. Taking into consideration of psychological, physiological user's features will let to adapt this web-interface parameter.
- Menu appearance. The task of this parameter's adaptation is to design menu, which will be suit to user's psychological characteristics (as a text, icons, special images etc.).
- Background color.

- Text color.
- Hyperlinks color.

## User Model

User model is not a new concept. First it was introduced in 1974 by the Institute of Informatics of the USA Congress. Now user model is interpreted as system's notion about the user, which generates either on the base of predetermined information about the user or on information acquired in the process of human-computer interaction. In spite of the fact that the user modelling is studying for a long time, there are no common principles of user model generation and it's implementation as a complex adaptation criterion.

The basic tendencies of the current user modelling research are:

- the number's increase and variety expansion of parameters included to the user model;
- user modelling use for adaptive systems development,
- user modelling implementation for the wide range of software systems development,
- attempts of generalized user models generation.

The network technologies expansion ensured new application fields for user modelling. Thus user model generation uses in adaptive hypertext navigation systems.

In the adaptive hypertext navigation systems the user model includes:

- users' goals,
- user's knowledge,
- user's hyperspace experience,
- user's background,
- user's preferences.

Also the more simple stereotype model (Rich model - [Rich, 1983]) is used for user knowledge representation. The stereotype model differentiates several groups of typical or «stereotype» users. For each user model measuring the system should offer several possible stereotypes.

The distinctive feature of our research is an attempt of user model generation process systematization and also including of psychological, physiological and cognitive features into the user model. Up to nowadays the major part of users modelling approach have comprised only the group of the human-computer interaction parameters (number of errors, main executed commands, visited pages) and also user knowledge about the subject domain. We suppose that including of psychological, physiological and cognitive features into the user model should improve human-computer interaction process quality as well as it will considerably increase user interface adaptation flexibility.

In our research we propose the concept of the user model [Rich, 1983; Wagner, 1982] as a set of formal representation of different factors, which affect the user's productivity in distance learning system environment. The user characteristics are grouped into several classes. The proposed distance learning system user model structure is shown on the Fig.2.

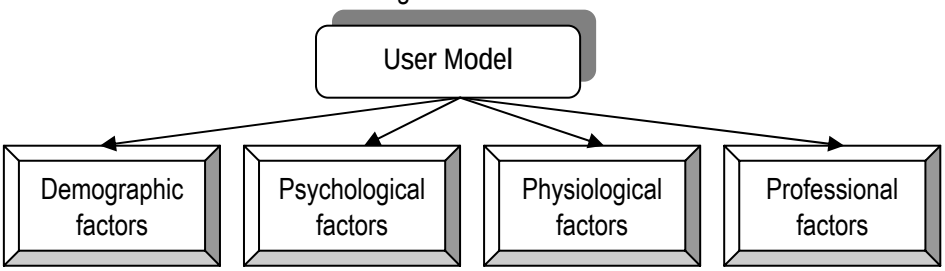



Fig. 2. Distance Learning System User Model Structure.

Demographic factors comprise such essential user's parameters as age, gender, first language, place of birth, social and cultural peculiarities. In our research we study two demographic factors – gender and age.

The group of the psychological factors in the distance learning system's user model consists of cognitive and communicative psychological peculiarities. The distance learning system's user model in our study

includes the following psychological factors: *ability to study, conformism, the level of locus control, cognitive style, logical mentality style*.

The study of the psychological characteristics group is in the centre of attention in our research. For the distance learning systems tasks the user cognitive style [Witkin, 1981] is an extremely important factor. It considerably influences the problem solving way. The study of the user logical mentality style or deductive/inductive strategies can help to present the learning material more comfortable for the student, because those who are using deduction always perform their cognitive activity with the top-down strategy from the higher level of abstraction to more and more detailed schema and in the variant of induction the users ascend from the unconnected elementary concepts to metaconcepts.

The user's physiological parameters have the greatest influence the productivity of the human-computer interaction. We have included two factors into the distance learning system user model - *attention* and *mistakes frequency* - for the purposes of our research.

In our study professional factors group consists of the following user features: *expertise level, user professional experience in the subject domain, user education and user computer skills.*

In this part we have described the most common structure of the distance learning user model. In our project we study all factors mentioned above and some other factors to be included at the distance learning systems user model final structure.

## Adaptation of the Distance Learning System

We consider the process of the distance learning system adaptation twofold – as interface adaptation and scenario adaptation. Scenario adaptation implies adjustment of the learning process scenario to the user peculiarities. Therefore we have included some characteristics, which in our opinion influence the learning materials navigation scenario, to the user model. Our adaptation comprehension corresponds the classical adaptation notion [Brusilovsky, 1996] (see Fig. 3).

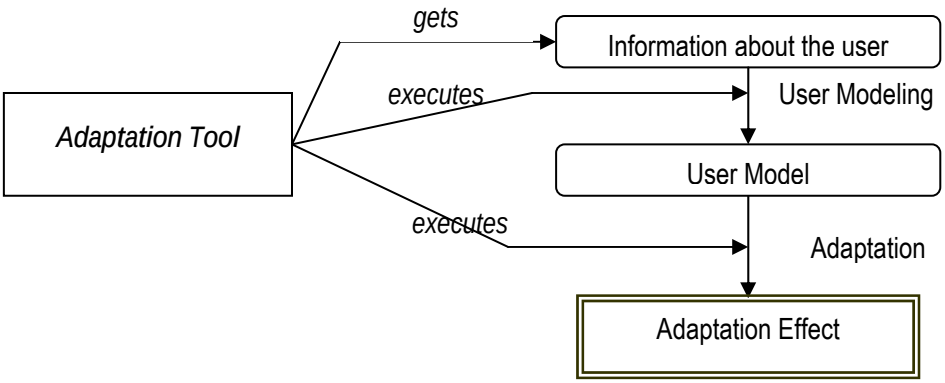



Fig. 3. Classical notion of the adaptation process in the adaptive systems

The major aim of the reported project is to develop interface adaptation tool for the distance learning systems. The process of the interface adaptation supposes inferences rules implementation to the interface generation. We are going to form knowledge base comprising user/interface models correlation rules. Now we carry out special experiments, which will allow to study correlation between user and interface models. The special software tool InterTrivium was developed to provide this experiments.

## InterTrivium – User Model Acquisition Tool

InterTrivium is a specially designed software system to form distance learning system's student model (first version called TOPOS was developed by Voinov, second version TRIVIUM developed by Geleverya T.).

It is an application for multi-factor quiz's data interpretation developing and the user model generation. The system can work with all types of question-answer tests (graphical tests, multi-factors test etc.). InterTrivium includes tools for interactive visual editing of tests' descriptions and tests' scales.

The main InterTrivium major targets are:

- test development,
- quiz/questionnaire executing,
- the result data interpretation,
- user model generation.

In the system there is an intelligent tool for automatic verbal interpretation of test results for each respondent using rules, defined by experts-psychologists.

The prototype of InterTrivium is implemented in the framework of PHP scripting language and can store data in MySQL database or in text files.

The outcome of the described application may then be used both by Internet-based and standalone computer-aided learning systems. InterTrivium can serve as a user model generation tool and as an application for the different testing and queering support.

Now InterTrivium supports several tests on the user interface preferences and some professional psychological tests. Some experiments aimed to find correlation between the user model and distance learning interface components are providing. The interface adaptation tool for the distance learning system will be developed on the base of experiments results.

## Acknowledgments

The research work reviewed in this paper has been carried out in the context of the Russian Foundation for Basic Research funded project “Adaptable Intelligent Interfaces Research and Development for Distance Learning Systems”(grant N 02-01-81019). The authors wish to acknowledge the co-operation with the Byelorussian partners of this project.

## Conclusion

The importance of the user interface adaptability is evident. The intelligent interface development can noticeably improve distance learning systems outcome. The approach described in this paper can be titled as design and building of adaptive interfaces embedded in the distance learning systems via user modelling. This approach is based on the user-centred technology that puts stress at the usability, handiness and efficiency of human-computer interaction.

The described project is under active development. Currently, different system components are studied – up to considerable extent – separately. This is referred to, e.g., user modelling, distance learning, Internet programming, description of subject domain.

## Bibliography

[Brusilovsky, 1996] Brusilovsky, P. Methods and techniques of adaptive hypermedia. User Modelling and User-Adapted Interaction, 6, 87-129.

[Nielsen, 1999] Nielsen J. Designing Web Usability: The Practice of Simplicity // New Riders Publishing; 1999. – 432 p.

[Rich, 1983] Rich E. Users are Individuals: Individualising User Models // Int. Journal of Man-Machine Studies, vol.18, 1983. – pp. 199-214.

[Wagner, 1992] Wagner E. A System Ergonomics Design Methodology for HCI Development. In: Proceedings of East-West Intern. Conference on Human-Computer Interaction EWHCI'92. - St. Petersburg.- pp. 388-410.

[Witkin, 1981] Witkin, H.A. & Goodenough, D.R. (1981). Cognitive Styles: Essence and Origins. NY: International Universities Press.

## Author information

Tatjana Gavrilova - Saint-Petersburg State Polytechnical University, Intelligent Computer Technologies Dpt., Computer Science School Politechnicheskaya 29/9, 195251, Saint-Petersburg, Russia; e-mail: gavr@limtu.spb.su

Ekaterina Vasilyeva – Saint-Petersburg State Electrotechnical University, IIST, Prof. Popova St., bl.5, Saint-Petersburg, 198156, Russia; e-mail: vki@mail.ru

# USING THE SIMULATION MODELING METHODS FOR THE DESIGNING REAL-TIME INTEGRATED EXPERT SYSTEMS

*G. Rybina, V. Rybin*

*Abstract. Certain theoretical and methodological problems of designing real-time dynamical expert systems, which belong to the class of the most complex integrated expert systems, are discussed. Primary attention is given to the problems of designing subsystems for modeling the external environment in the case where the environment is represented by complex engineering systems. A specific approach to designing simulation models for complex engineering systems is proposed and examples of the application of this approach based on the G2 (Gensym Corp.) tool system are described.*

*Keywords: integrated expert systems, real-time, simulation modeling, object-oriented model, rule, complex engineering systems, electrophysical complex.*

## Introduction

The more and more sophisticated character of modem software systems is due to the fact that their architecture comprises a great number of subsystems and components with different functional characteristics, which interact with different groups of users. Because many components are created and developed autonomously, without any provision made for the possibility of joint operation and support of integration processes in the course of evolution of systems, this results, as a rule, in a substantial deterioration in the reliability of such systems. On the other hand, tendencies towards the integration of investigations in different fields, which have prevailed over the last ten years, have necessitated the integration of semantically dissimilar objects, models, methods, concepts, and technologies. This circumstance has inevitably led to the emergence of new classes of systems, such as integrated intelligent systems, integrated expert systems, integrated information systems, integrated manufacturing systems, etc.

Thus, research and development in the field of advanced integrated systems are of particular importance at present. Among the studies in this field, we can mention certain results on the theory and technique of designing integrated expert systems (IES) for static application domains (AD) obtained by the author in the course of conducting the AT-TECHNOLOGY research project (see, e.g. [Rybina,1997]).

However, the integration problems are most conspicuous in the construction of dynamic IES operating in a real-time mode (real-time integrated expert systems (RTIES)), because in this case it is necessary to ensure the following (see [Rybina,1998]): simulation of the external world and its various states; representation, storage, and analysis of time-varying data incoming from external sources; simultaneous temporal reasoning about several distinct asynchronous processes (tasks), support in functioning the inference mechanism under conditions of resource (time and memory) limitations, and other capabilities.

In this connection, special software tools (ST) are needed. These tools must make it possible to design and develop RTIES that can operate in dynamic AD, including the case where the correction of search strategies and knowledge acquisition are possible directly in the process of searching for a solution. The most widely known ST of such a kind are G2 (Gensym Corp.) and RT works (Talarian Corp.).

Over recent years, the author has accumulated certain experience in designing RTIES on the basis of the G2 tools for diagnostics and control problems, such as the control of modem electrophysical complexes [Rybin,Rybina,1998a, Rybin,Rybina,1998b], the diagnostics of complex engineering systems [Rybina,1998], the launch readiness verification of carrier rockets (prelaunch monitoring of carrier rockets) [Rybin,Rybina,1999], radioecological monitoring of areas adjacent to nuclear power plants [Kosterev et al, 1998].

By and large, despite the external dissimilarity of AD, complex engineering systems (CES) were studied. These systems are objects of a technical nature characterized by the following [Rybin,Rybina,1998]: their parameters constantly vary (in real time); they comprise from several hundred to several thousand functionally and structurally interrelated components, subsystems, modules, units, etc.; the diagnostics of these objects can be considered as a specific control process with the goal of determining the technological state of objects at each current instant (the general task of diagnostics of the object status) and, in addition, the task of fault finding (as a special case of the general diagnostic task); the functioning of these objects is a complex technological process accompanied by a multitude of abnormal conditions, rapid changes in the environment,

and the lack of time for decision-making in response to abnormal conditions; a high price is paid for errors made by operators.

Therefore, any RTIES for diagnostics and control of CES (which are discrete and discrete-continuous for the main part) must ensure, in the general case, support for the execution of the following tasks: the dynamic modeling of all processes of functioning the of CES; monitoring the CES operation, detection of deviations from the prescribed regime, prefailure alerting and abnormal condition warning, emergency cut-out, etc.; studying the actions of the operators who control CES and training of personnel; a convenient graphic user interface for monitoring variations in the basic parameters characterizing CES operation, etc.

The architecture of an IES that is designed for real-time operation undergoes substantial changes due to these circumstances, because practically all basic components of a static IES are modified and two new subsystems are added—one of them is intended for the environment simulation, and the other supports the interface with the physical equipment.

The primary emphasis in the present paper is made on the problems of designing one of the most important components of the RTIES, the subsystem for simulation of the environment, which is represented by CES.

## 1. Statement of the Problem

The problems of real-time modeling of the external environment and its various states are very important in designing any RTIES. The methods and tools of simulation modeling (SM) are the most appropriate here, because the dimensionality of problems being solved and the unformalizability of complex objects and systems for which the RTIES are designed do not allow one to use rigorous mathematical methods.

On the whole, the SM is quite efficient, but, at the same time, it is a rather labour-consuming method and entails a number of problems such as the necessity to provide an adequate description of systems and processes in these systems, a correct interpretation of the results obtained, questions of stochastic convergence of simulation processes, the necessity to overcome difficulties caused by the problem of dimensionality, etc.

As was noted above, the application of methods and tools of SM in RTIES is caused primarily by the necessity for real-time modeling of the external environment and, in particular, to include the corresponding subsystems that adequately reflect all the processes and laws of functioning of the CES in the architecture of RTIES.

In [Rybina,1997] the author consider and solve these problems within the framework of the general-purpose problem-oriented methodology (POM). This methodology is intended for use in static and dynamic applications; it is a set of models, methods, algorithms, and procedures for designing of applied IES.

Since IES are implemented by integrating the methods and technologies of ES with those of conventional programming, the basic problem of integration within the framework of POM should be considered in the following way (see [Rybina,1997]): integration in the framework of IES of different components that execute formalized and unformalized tasks and determine the specific character of functioning of the entire IES (the top integration level); integration (functional, structural, and conceptual) related to basic system designs and concepts of development and design of particular classes of IES and their components (the medium integration level); integration (informational, software, and hardware) related to the technologies, ST, and platforms used (the bottom integration level).

Analyzing the problem of the top-level integration problem, we proposed the classification of IES and introduced the concepts of IES with the superficial and deep integration of components. It was also shown that the methodology for developing simple ES can be used only for designing IES with superficial integration and is completely inapplicable for IES with deep integration. In this case, we propose to apply the approach that improves ES by incorporating functions of a certain component N (where N is a database management system, an application package, SM system, etc.) that are unconventional for such ES, which is an important conceptual basis of the POM.

Here, the problems of integrating ES with SM in the framework of RTIES with deep integration of components are of primary interest, because, in this case, it is necessary to ensure the following (see [Rybin,Rybina,1998a]): the conceptual uniformity of approaches, models, and methods being used; the combination of rigorous mathematical methods of search for solutions with unformalized heuristic methods based on expert knowledge; due regard for the time factor both in the construction of models of AD and in the search for solutions, and other capabilities. The problems of designing RTIES that integrate simulation, conventional ES, as well as other constituent components of IES, are not clearly understood yet; therefore, the present paper deals with these problems. The modular principle of designing RTIES on the basis of POM,

as well as the similarity of certain concepts used in ES and SM, makes it possible to integrate these technologies.

Below, we describe the results of experimental approbation of the elaborated models and methods of POM that ensure the integration of ES with SM in designing prototypes of RTIES for problems of control and diagnostics of CES; moreover, we focus mainly on one of the most complicated problems in the construction of RTIES, namely, the control problem for CES.

## 2. Construction of Simulation Models of CES

As noted above, the complexity of CES under study does not allow one to apply rigorous mathematical (analytical and numerical) models when describing these systems. There is no way in which this problem can be solved except by constructing a simulation model (MSM); moreover, preference should be given to the application of methods of the intelligent SM. Now, we consider an example of designing the simulation model MSM for the life-support and survival system (LSS) of the electrophysical complex (EPC) [Rybin,Rybina,1998a].

It should be noted that, due to their complexity, modern EPC are designed to eliminate the effects of subsystems on each other to the greatest possible extent; hence, each particular subsystem can be considered as a CES, and EPC can be treated as the set of CES, i.e., CES = ($CES_1$,..., $CES_N$}. Therefore, the control system (CS) for a given object is a hierarchical CS whose control object at the top hierarchy level is the control system of a particular subsystem of the LSS.The simulation model $M^{SM}$ can be subdivided into the model of internal stochastic perturbations ($M_{SD}$) and the model of the control system (Mcs). To simplify the description of the structure of the simulation model $M^{SM}$, we assume that the LSS comprises only one subsystem, for instance, the subsystem of vacuumizing. In this case, the simulation model $M^{SM}$ of the life-support and survival system of the EPC will be represented by the simulation model $M^{SM}$ of the subsystem for vacuumizing.

Thus, from the standpoint of SM, the life-support and survival system of the EPC is a discrete-continuous system. In the general case, the set-theoretic model of this system has the following form:

$$M^{SM}_{LSS} = \{M_{CO}, M_{CONTR}, M_{SD}, V^X, V^U, V^E, V^Y, V^Z, S, F^{Y\to U}, F^{XEU\to YZ}\},$$

where $M_{CO}$ is the model of a control object, $M_{CONTR}$ is the model of a controller; $M_{SD}$ is the model of internal stochastic perturbations; $V^X = \{ v_i^x \}$, $i = (1, m)$, is the set of monitored uncontrolled inputs; $V^U = \{ v_j^u \}$, $j = (1, s)$, is the set of monitored controlled inputs of the $M_{CO}$; $V^E = \{v_h^e\}$, $h = (1, k)$, is the set of stochastic perturbations, $V^Y=\{v_l^y\}$, $l = (1, r)$, is the set of output parameters of the Mco (this set is used in the controller); $V^Z = \{ v_g^z \}$, $g = (1, q)$, is the set of output parameters of the Mco; $S = \{s_c\}$, c = (1, n) is the set of possible (admissible and abnormal, i.e., inadmissible) states; $F^{Y\to U}$ is the function generating the control vector $u(t_{i+1})$ on the basis of the incoming output vector $y(t_i)$; $F^{XEU\to YZ}$ is a function mapping the input of the CO into its output.

Let $I$ denote the input of $M_{CO}$ ($V^x$, $V^u$, $V^e$), and let $O$ denote the output ($V^y$, $V^z$); then, we have $o(t) = F^{xeu\to yz}(i(\tau), \forall\tau \in [v, t], o^{(k)}(v), k = \varnothing; (n - 1), t)$, where $[o^{(k)}(t)(t=v)= o^{(k)}(v), k = \varnothing; (n - 1)$, are the initial conditions (IC). Thus, at any instant $t$, the output is a certain function of the input and the IC.

The output of $M_{CO}$ has the dimension $(r + q) \le n$; therefore, $O \subseteq S$, and each particular state of the CO is described by the set of selected (on the basis of different criteria) properties (characteristics), i.e.,

$$C^{\sim} = \{c_1^{\sim},...,c_\lambda^{\sim}\}$$

where $C^{\sim}$ are the valuated properties of the CO. Thus, the set $C^{\sim}$ can be used to describe the set of states $S \subseteq S^{ad} \subseteq S^{ab}$, where $S^{ad}$ is the set of admissible states and $S^{ab}$ is the set of abnormal states.

This formalized representation of the simulation model $M^{SM}$ of LSS of the EPC describes the operation of the entire system, but it is still too abstract for further implementation. In the present paper, in order to make the obtained model more specific and universal, we have used the Rational Rose Real-Time 6.0 CASE tool and the UML with enhancements for the support of real-time system engineering (UML-RT) is used as a language of model designing; these enhancements include the structural elements of the UML-RT-like Capsule and behavioral elements of type of Protocol. The idea is to represent all units using the diagram of classes and to write the diagram of states and transitions for each class. For instance, all units of the equipment making up $M_{CO}$ and $M_{CONTR}$ are represented in the form of capsules (abstract representations of real-world objects, i.e., the equipment of the LSS of the EPC) with a necessary set of ports (abstract representations of data-transfer channels of real-world objects) through which the messages from other elements (from capsules in abstract declarations, from EPC's equipment in the real world) are incoming.

Since $O \subseteq S$, the output ($V^y$, $V^z$) of the $M_{CO}$ is completely described by the attributes of capsules that represent $M_{CONTR}$; the component ($V^u$) is described by the capsules of the component representing the generator of

stochastic perturbations; and each arrow ($V^e$) is represented by a message (information) transfer channel, i.e., by a protocol.

For each capsule, a diagram of states and transitions is constructed; the sets of these diagrams for all capsules are defined (represented) by the functions $F^{y\to u}$ and $F^{xeu\to yz}$. Thus, the representation of the simulation model $M^{SM}$ in terms of UML-RT has the following form:

$$M^{SM}_{LSS} = <C, P, S, T, E, R_C, R_p, A>,$$

where C is the set of capsules; *P* is the set of ports, *S* is the set of states of capsules; *T* is the set of transitions; *E* is the set of events initiating a transition to another state; $R_C$ is the set of relations between capsules, $R_p$ is the set of relations between protocols; *A* is the mechanism for event tracing and the initiation of transitions (actions) corresponding to an event.

The representation of the model $M^{SM}$ obtained above allows one to pass to its implementation at the level of tools; moreover, no constraints are imposed on the choice of ST for this purpose. In the present paper. the G2 system is used as a toolkit for implementing the simulation model $M^{SM}$ of the life-support and survival system of the EPC and as a software development environment of the entire RTIES. In the G2 object-oriented environment, the above model is unessentially modified, namely: the diagram of classes, which was developed in the Rational Rose RealTime, is turned into an analogous diagram in the G2 environment; the hierarchy of protocols transforms into the hierarchy of connections and relations; the logic of the diagram of transitions and states is described using the G2 rules for the *whenever* construction; the application of these rules allows one to form the mechanism of *event tracing.*

An *event* means that the system is in one of the following a priori known states: a variable, a parameter, or an attribute of an object received a new value; an error occurred when a value was assigned to a variable; a variable lost its significance (the value is no longer significant); an object of some class was created; an object was moved (changed its coordinates) on the desktop; an object passed into an active or dormant state; two objects became related by a certain relation; two objects was connected to each other.

Therefore, $M^{SM}$ in the G2 environment can be represented as the following set:

$$M^{SM}_{LSS} = <CL, O, C, E, RL_E, R_{C1}, R_0>,$$

where *CL* is the set of system classes; *0* is the set of objects; C is the set of connections between objects; *E* is the set of model events; $RL_E$ is the set of rules of the event-tracing machine; $R_{C1}$ is the set of relations between classes; $R_O$ is the set of relations between objects.

## 3. Methodology for Designing Simulation Models of CES

Thus, the following particular methodology for constructing the simulation model MSM of a CES, which is oriented towards use in the G2 environment, was developed within the framework of POM.

1. An AD is analyzed; the basic concepts of the AD, as well as the characteristics and operations of functioning of these concepts, are specified.
2. The abstractions of these concepts are described as classes in G2. The characteristics of concepts of a CES are represented by attributes, and the operations are represented by methods.
3. A powerful visual editor is employed to construct a scheme of the equipment of the real-world CES using the program objects and instances of described classes. Moreover, the interrelations of objects of the real-world CES are assigned by connections and relations.
4. As a result, one obtains the scheme $S = <0, R>$, where *0* is the set of objects of the scheme and *R* is the set of relations-between these objects.
5. Then, a set of model and temporal events, $E = \{e_i\}$ is constructed. A model event is an a priori specified state, i.e., the set of valuated attributes of one or several particular objects. A temporal event is a priori preset model or real time.
6. The set $D = \{d_j\}$ of actions, which are associated with the set of events, *E,* is constructed (an action is the totality of methods of objects of the scheme *S*), as well as the Scheduler (sequence monitor) of their joint functioning (sequential, concurrent, or with time delay). An action has a duration, which is realized in G2 through the use of the *wait for t* construction, where *t* is the delay time.
7. Based on Items 4 and 5, an event-tracing machine is constructed. This machine is applied for scanning the states of the system and for the initiation of an action associated with an event, when the system passes into a state for which this event is described in the set *E.* This mechanism is implemented in the G2 environment through the use of the rules of the *whenever* construction.
8. The set *B* of initial states is determined. The initialization of the scheme *S* with one of the initial states is carried out by executing constructions of the type *initially*.

## 4. Example of Application of Methodology for the Construction of Simulation Models of CES for RTIES

We use the example of modeling the LSS of the EPC for an operational prototype of the RTIES [Rybin, Rybina, 1999] to illustrate the elaborated methodology for designing simulation models of CES in the G2 environment. The LSS of the EPC, the charged-particle accelerator in the case under consideration, comprises interrelated subsystems for electric power supply, water supply, vacuumizing, magnet cooling, tunnel ventilation, high-frequency electric power supply, radiation protection, and fire safety.

We restrict ourselves to the consideration of the vacuumizing subsystem designed for the development and maintenance of vacuum in vacuum chambers. When the accelerator is started up, the air is evacuated in two stages; moreover, different types of pumps are used. At the first stage, a low vacuum is developed; at the second stage, the air is evacuated so as to develop a high vacuum. If an abnormal condition arises in the process of operation of the accelerator, i.e., if the pressure is at variance with that existing under high vacuum, then a specific control signal is sent to the emergency system and the vacuum chamber is blocked; the corresponding message is generated and sent to the control desk with the aid of the RTIES (see Fig. 1).

The construction of the simulation model $M^{SM}$ of the vacuumizing subsystem is started simultaneously with the object-oriented analysis of the AD when the basic concepts of the AD are specified and the relations (their type, multiplicity, etc.) between objects are refined. In the case under consideration, one can distinguish the following: the vacuum chamber, the vacuum sector, and the exhaust units (VN1-MG, NEM-300, TMN-200). As has already been noted, the model is designed using, the Rational Rose for RealTime (although this is not obligatory and the designing can be carried out directly in the G2 environment). A diagram of classes is constructed; in our case, we apply the UML-RT, so this diagram is a diagram of capsules; moreover, the relations between capsules can be of the following four types: association, utilization, aggregation, inheritance.

Then, the interacting capsules and the message flows that describe their interaction are specified. For each such interaction, the concept of the message transmission channel and, as a consequence, that of the protocol of transmission of these messages are introduced abstractly. Thus, the diagram of capsules is refined by the diagram of protocols. Further, the sets of attributes that characterize the abstract state of a given capsule as a consequence of the state of the real-world equipment are specified, the conditions for transition from one state to another and the corresponding actions (i.e., the variation of a particular variable, the start-up of the capsule method, the transmission of a message to another capsule) are specified. On the basis of the information thus obtained, a diagram of transitions and states is constructed with the use of the UML-RT tools.

After this, the complete source information for the implementation of the obtained model in the G2 environment is available. In order to generate a list of events, it is sufficient to simply write out the conditions for transitions between states. The actions that are performed when a certain event occurs are also represented in the diagram of states and transitions using Rational Rose for RealTime. It only remains to connect the events and the actions associated with them by *whenever* constructions; these constructions represent one of the types of rules of the G2 system:

<*whenever rule*> :: = whenever <*event declaration*>

[or <*event declaration*>] [and when <*logical expression*>] then <*list of actions*>.

Then, a set of system states is constructed. To simplify the further description of the model, we assume that the vacuumizing subsystem can be in one of the nominal states only, for instance, the system is in the "off" state, which is described by the condition *none of the pumps is powered;* the system is at the stage of maintaining a low vacuum, which is described by the condition *the vacuum is in the range from $10^{-3}$ to $10^{-2}$ mm Hg;* the system is at the stage of maintaining a moderate vacuum, which is described by the condition *the vacuum is in the range from $10^{-3}$ to I0$^{-6}$ mm Hg;* the system is at the stage of maintaining a high vacuum, which is described by the condition *the vacuum is in the range from $10^{-6}$ to $10^{-9}$ mm Hg.*

Then, a set of actions is constructed in the G2 environment. The actions in this case are the G2 procedures that initiate and suspend the methods of objects and the methods themselves; for instance, the procedure that initiates all the methods of air evacuation used by pumps of the second type is declared as follows: *power_all_p2(VC),* etc.

And, finally, the event-tracing machine is constructed. In our case, the machine for event tracing and initiation works is as follows:

*whenever the pressure of any vac_chamber VC receives a value and when the pressure of VC < 10 and the status of VC /=* 2 *and the status of VC /= 3 and the status of VC /= 4 then conclude that the status of VC = 1;*

*whenever the status of any vac_chamber VC receives a value and when the status of VC = 1 then conclude that the status of VC* = 2 *and start power_all_p2(VC);*

*whenever the pressure of any vac_chamber VC receives a value and when the pressure of VC < 10e-5 and the status of VC /=4 then conclude that the status of VC = 3;*

*whenever the status of any vac_chamber VC receives a value and when the status of VC = 3 then conclude that the status of VC = 4 and start power _all_p3( VC).*

The initial conditions are determined using the rules of the *initially* type, which invoke the procedures of the system initialization when the applications are started up, i.e., *initially start initial_top_value (top_value) and start initial_bottom_yalue (bottom_value) and start initial_stub posts (stub_posts).*

An example of operation of the current version of the prototype of the RTIES for control of the EPC is presented in Fig. 1. This figure presents the case of detection of an abnormal condition in the operation of the vacuumizing subsystem. This condition is simulated with the aid of the simulation model $M_{LSS}^{SM}$; it is detected by the RTIES; and, using the rule-based inference, a message indicating possible causes of this condition and suggesting remedies for the trouble is issued to the operator.

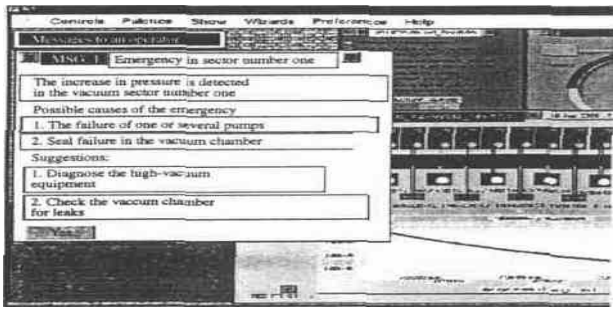



Fig. 1. An example of operation of the prototype of the RTIES for the control of the EPC

## Conclusion

Thus, a simulation model $M^{SM}$ of any CES in the G2 environment is an object-oriented model; i.e., it is the set of program objects that simulate the dynamics of the behavior of the real-world CES described with the use of these objects; these latter are interrelated both by data transfer channels and by logical circuits. The real-time simulation process as such is supported by the G2 Scheduler (sequence monitor), which coordinates the processing of model and temporal events; this circumstance substantially facilitates the elaboration of the simulation model $M^{SM}$.

It should be noted that, if necessary, the MATLAB system or its analogs can be additionally used for modeling complex continuous processes described, in particular, by differential equations of the second or higher order; in this case, integration with G2 is carried out on the basis of tools of the GSI interface (G2 Standard Interface).

## Bibliography

[Rybina,1997] G.V.Rybina. Task-Oriented Methodology for Computer-Aided Construction of Integrated Expert Systems for Static Problem Domains. Izv.Ross.Akad.Nauk, Teor.Sist.Upr. 5. 125-137. 1997.

[Rybina,1998] G.V.Rybina. Specific Features and Principles of Design of Integrated Expert Systems for Diagnosis of Complex Industrial Systems. Pribory i Sistemy Upravleniya. 9. 12-16. 1998.

[Rybin,Rybina,1998a] V.M.Rybin., G.V.Rybina. G2 Real-Time Expert System for Control of Electrophysical Complex. Proc. 2nd IMACS Int. Multiconf. CESA '98. Computational Engineering in Systems Applications, vol.1. Tunisia. 659-663. 1998.

[Rybin,Rybina,1998b] V.M.Rybin, G.V.Rybina. Using the Tools Complex G2 For Control of an Electrophysical Complex. Accelerator and Large Experimental Physics Control Systems. Proc. Int. Conf. on Accelerator and Large Experimental Physics Control Systems, Beijing (China), Zhaanel, J and Daneels, A., Eds.. Cern: Science Press. 107-109. 1998.

[Rybin,Rybina,1999] V.M.Rybin, G.V.Rybina. Dynamic Real-Time Integrated Expert Systems: Analysis of the Experience of 'Study and Development. Pribory Sistemy Upravleniya. 8. 4-8. 1999.

[Kosterev et al, 1998] V.V.Kosterev, E.A.Kramer-Ageev, G.V.Rybina, et al. A Prototype of Real-Time Expert Systems for Radio Ecological Monitoring of Areas Adjacent to Nuclear Power Stations. Proc. VI Nat. Conf. with Int. Participation II'98. Pushchino: RAII. V.2. 433-439. 1998.

## Author information

Galina Rybina - Moscow Engineering Physics Institute (State University),Kashirskoe shosse, 31, 115409, Moscow, Russia,Email: galina@ailab.mephi.ru

Victor Rybin - Moscow Engineering Physics Institute (State University),Kashirskoe shosse, 31, 115409, Moscow, Russia,Email: rybin@aie.mephi.ru

# EXPERIMENTAL ESTIMATION OF ERGONOMIC PARAMETERS OF THE COMPUTER TRAINING PROGRAMS ON ELECTROTECHNICAL DISCIPLINES

*D.A. Yakovetc, L.H. Zainutdinova*

*The abstract: The present research offers the technique of the experimental estimation of the ergonomic parameters of the computer training programs, created on the basis of the method of theoretical images are submitted.*

*Key words: the computer training programs, ergonomic parameters, method of theoretical images*

## Introduction

Computerization as well as the information process built on it are being widely implemented in all spheres of human life, including such an important sphere of social activity as education. New information technologies and computer-based facilities being the nucleus of the information process of education. Nowadays a computer classroom in the educational institution has become the same necessary and habitual attribute, as library. New means of training: all kinds of computer training programs (CTP), including electronic textbooks (ET) are developed.

The learning efficiency is influenced by various factors, among which comfortable conditions for work with the computer program are great importance. The analysis of this aspect of interaction of the person and computer will be carried out within the framework of ergonomics – the science that studies in complex the law of interaction of the person and engineering during this or that activity with the purpose of development of the requirements for increasing the efficiency of this activity. In the present research the task was to estimate the ergonomic parameters of a number of computer educational programs on Electrical Engineering.

## The basic part

The electronic textbook, as a special case of the man - machine system (MMS), is characterized by system-technical and ergonomic properties. According to [Зараковский и др., 1993, с.27], system-technical properties are those that cause the adaptation MMS to performing its appropriate functions (efficiency, reliability, cost, etc.). Ergonomic property - characteristic of MMS and its elements, which are defined by biomechanical, physiological and psychological possibilities of human activities.

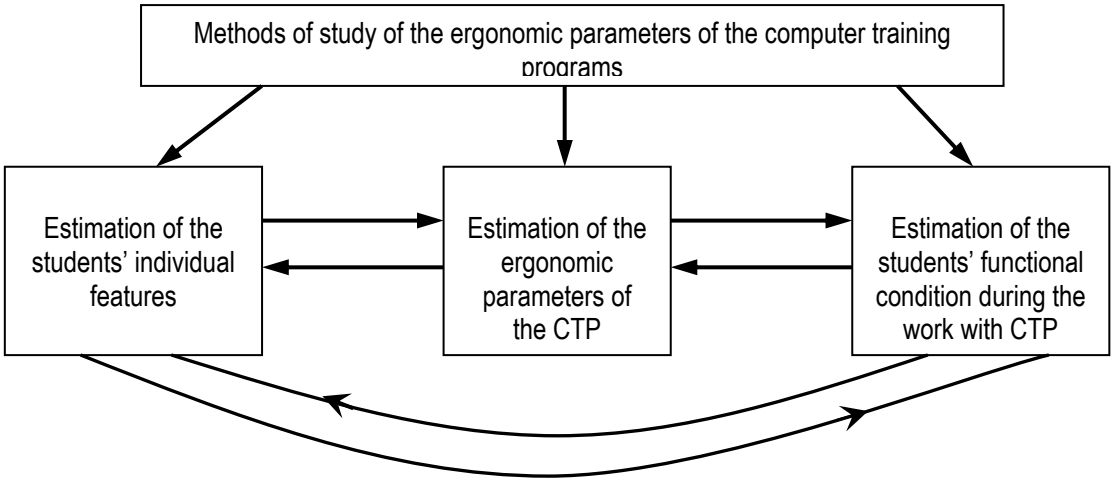
Already at the first stages of the development of computer facilities the problem of interaction of the person and computer has become a subject of research in engineering psychology, ergonomics, psychology of labor. Various aspects of an operator's work with the use of a computer were originally examined in Automatic Management System. The most attention was paid to optimization of hardware component of the human-machine interface (HMI). Now, because of the constant perfection of computer engineering the ergonomics attention was switched to software components of HMI. Recently some researches have been ergonomic carried out which dealt with ergonomic designing of interaction of the man and computer in the system of educational: preschool, secondary, university; the questions of maintenance of optimum work conditions being mostly examined: such correct mode, the microclimatic conditions, the light exposure, the correct pose during with work, the organization of the workplace. But the questions of the ergonomic of optimization of properties of CTP are paid less attention to. The recommendations on organizing the procedure of dialogue, the adaptations of the program to students' individual features, on color and spatial design of the information on the display are often of generalized character. There are not any works investigating the ergonomic properties of CTP of a narrow purpose and having their own specific character, for example, for teaching general-engineering disciplines (GED). Thus, the question of the estimation of quality of educational software is extremely urgent.

One of central problems of ergonomics is the study of functional condition of the users arising during this or that activity, and the development of adequate methods of their estimation and correction. Functional condition (FC) of the person - complex of the characteristics of those functions and qualities of the person, which directly or indirectly determine performance of working operations [Введение в эргономику, 1974, с.94].

Traditionally researched types of FC are exhaustion, tension, stress. As the phenomena of physical and mental exhaustion, as well as the way of their description can depend in great degree on the examinees, personal peculiarities and aims, it is also advisable to get his personal profile, especially concerning his

stability and introversion / extraversion. The factors, causing the increased motivation of the user also require deep study.

Taking the above-stated facts into account, as well as the specificity of teaching GED with the use of information technologies and using the data of the analysis of the scientific literature devoted to the ergonomic criteria of efficiency of activity and the questions of psychological testing, Yakovetc D.A. has developed a technique of an experimental research of ergonomic parameters of computer training programs (picture 1.).




Picture 1

Nowadays we have a situation when the greatest number of CTP under development belongs to the branches of humanitarian disciplines and natural sciences disciplines for the school system of education. The quota of CTP on general engineering disciplines (electrical engineering, heat engineering, hydraulics, theoretical mechanics, etc.) and special technical disciplines (according to the profile of the graduating faculties) is insignificant. One of the reasons that slows down the development of CTP on technical disciplines for system of secondary and higher education is the lack of theoretical bases of ergonomic designing of the similar programs.

The present research offer the method, of an experimental estimation of the ergonomic parameters of CTP enables the comparative analysis of the existing CTP. This technique has been applied for the estimation of the several computer training programs on electrical engineering. The research were carried out on «The Three-phase circuits» (TZ) program, created on Electrical Engineers Faculty of Astrakhan State Technical University (ASTU), as well as «The Theory of electrical circuits » (TEC) program, developed in Siberian State Academy of Telecommunication and Computer Science [Бакалов и др., 1998].

We consider that the TEC computer program represents the most widespread type of modelling program in the branch of engineering disciplines. The software for mathematical and imitating modeling allows expanding the borders of experimental and theoretical researches, to supplement a physical experiment. During the work with the TEC program the student is given the possibility to choose the necessary parameters of the electric circuit, the TEC program demonstrates the change which occurs in the work mode of the circuit. The drawback of the similar programs is the lack of feedback and of active educational activity of the students. The student acts in a role of the observer. On the finishing work with the TEC program the student isn't given any mark for the lesson.

The TZ program is an electronic textbook. According to [Зайнутдинова, 1999[1], с.35]: " The Electronic textbook (ET) is an educational program system of a complex purpose, ensuring the continuity and completeness of the didactic cycle of the teaching process, which gives the theoretical material, ensuring the drilling activity and the control of the level of knowledge, as well as the information retrieval activity, mathematical and imitating modeling with the computer visualization and service functions on the condition of realization of an interactive feedback".

The program TZ is designed on the basis of the method of theoretical images [Зайнутдинова, 1999[2]]. The educational material of the GED is know for its high level of abstraction and therefore is perceived and acquired by the students with difficulty. The offered method of theoretical images ensures the increase of availability of the form of an abstract educational material.

The scientific knowledge is in the overwhelming majority of passed on in verbal form as theories, laws, concepts and is accompanied by some symbolical explanations, for example, mathematical formulas, it enables to speak about the verbalized form of scientific knowledge. It is know form psychology that for understanding and mastering the educational material well, the semantics of the text should be represented in the form of visual images. On the basis of the above-stated facts, the author of the present work introduces the following definition: "The Theoretical image is a visual-figurative representation of the semantics of the verbal forms of scientific knowledge (concepts, laws, theories)" [Зайнутдинова, 1997, с.164].
One of interest, but poorly developed problems is the problem of passing on the image from one person to other during dialogue [Ломов, 1991, с.69]. It is that problem that comes to the foreground in computer training systems. The theoretical image that has been formed in experienced teacher's consciousness for many years, can not be directly handed down to, the student. During the application of the traditional technology an image passes on to other person on the thinking-in-words level. The application of new information technologies principally new possibilities appear, which enable to pass on the visual - figurative concepts from the teacher to the student. The necessity of a verbal description is reduced. The theoretical image being property of the experienced teacher can be passed down to the student with the least losses and distortions through a didactic program system. But there is a necessity for the development of new technologies or strategy of such transmission. On the basis of the psychological and pedagogical theories of training the present research gives recommendations for designing theoretical images for CTP [Зайнутдинова, 1999[1], С.144-145].
The method of theoretical images is capable to some extent to affect all the components of the student's learning process, but the most significant changes take place during the perception, understanding, memorizing and revision of the educational material. Besides the method is supposed to improve the emotional attitude t the process of learning and to reduce the necessity for volitional efforts on the part of the student. While using the method of theoretical images the educational information is given not only as texts and formulas, but also in a visual - figurative form. Both the verbalized left-hemispheric and the visual right - hemispheric information are percepted simultaneously. The processing of the right -hemispheric information is carried out at high speed, in complex. There is a sort of a complex (synergism) effect of interaction of the left - and right -hemispheric mechanisms of our thinking. At the achieved level of the quality of both hardware and software the visuality and vividness of the information can attract the user's attention to such extent that the necessity for a volitional regulation of the process of perception and understanding is reduced considerable. The method of theoretical images has high potential possibilities of presentation of the educational material with the support on the interrelation and interaction of conceptual, figurative and effective components of thinking, what is of immense importance for the creation of the CTP for general-engineering disciplines.
The high pedagogical efficiency CTP, developed with application of the method of theoretical images, was confirmed by research [Зайнутдинова, 1999[2]]. The aim of the present work is comparative experimental estimation of the ergonomic parameters of a number of CTP oh Electrical Engineering: the TEC modeling program and the TZ program, the latter being created with the application of the method of theoretical images.
The experimental research was carried out among 85 ASTU full-time students of the second year. It was participated by the students of the following specialties:

- « The Automated systems of processing of the information and management (AS) - 220200 », 31 persons;
- « Automation of technological processes and enterprises (AP) 210200 », 22 persons;
- « Networks of communication and commutation system (AC) 200900 », 32 persons.

The research was carried out in two stages. At the first stage the individual peculiarities of the personalities of the students who take part in the experiment were being estimated. At the second stage the ergonomic parameters of the computer educational programs and student functional state during the work with these programs were being estimated.

## The first stage of research

The estimation of the students individual personal peculiarities was carried out once at the specially arranged time (at the lecture).
For the estimation of the student's individual features (IF) it is necessary to fulfil some requirements:

1. The techniques of the estimation of the student's IF should be reliable and valid.

2. The techniques of the estimation of the student's IF should be convenient for a group testing. It is necessary to ensure the equal conditions of testing. The procedure of measuring, processing and interpretation of the investigated parameters should be limited in time.
3. The time required for carrying out the tests is to be taken into consideration (the IF-estimation is to be carried out during one lecture).

The first stage of research included:

1. The definition of the student's professional propensities.
2. The definition of the level of logical ability of thinking.
3. The definition of the basic properties of the nervous system:
   - The steadiness of the nervous system (the balance of nervous processes, expresses the correlation between the processes of excitement and inhibition in the cells of the cortex);
   - Mobility of nervous system (ability to react quickly to changes in the environment);
   - The force of nervous system (reflects a limit of efficiency of the cell of the cortex - their capacity to endure either very strong or very long-termed though not strong excitement, without coming to the state of inhibition.

The definition of the student's professional propensities. As the authors see for the estimation of the level of the student's interest in the study of GED it is very important to reveal their professional propensities. Presence or absence of a positive motivation during the work with a computer program is an e important factor that influences to a learning efficiency, the perception of the properties of CTP, the creation of the state of functional comfort or functional discomfort. According to [Чайнова, 1985] The functional comfort is the optimum functional condition of the working person, during which the correlation between the means and conditions of work is achieved. In this case the person develops a positive attitude to his activity, which ensures the adequate activization of his psycho-physiological processes, postpones the development of exhaustion, promotes long and high capacity for work without determent to health.

On the basis of the analysis of the literature on psychological testing and concordance with the above-state requirements, for the definition of the student's professional propensities differential - diagnostic questionnaire by E.A. Klimov [Горбатов, 1998].

The definition of the level of logicalability of thinking. As it is noted in the work [Зайнутдинова, 1999[2], c.95], while teaching GED, where also electrical engineer belongs, the student mast form in his memory rather great number ( a bank) of theoretical concepts, taking into account their interrelations. Besides the system of the scientific notions of GED are characterized by a high level of hierarchism and abstraction and a high degree logical interrelation of its components. Therefore, the authors maintain, that the level of the development of the logical ability of thinking is an important factor that affects the efficiency of the students work with the date of CTP. The level of the development of the logical ability of thinking was investigated with the help of the increasing difficulty test (Roven's technique) [Столяренко, 2000, c.111]. The given test corresponds to the above mentioned requirements.

The definition of the basic properties of the nervous system of the students participating in experiment, is of great importance for the given research. According to [Словарь …, 1998, c.600], the property of nervous system are the steady features of nervous system, that influence, when all the other conditions are equal, the person's individual psychological peculiarities. The whole set of the nervous system, forming a certain type of the nervous system, makes a physiological basis of the individual originality of the person's activity, directly influences a functional condition (exhaustion, stress, functional comfort, productive or unproductive tension, etc.) during various activities. Laboratory techniques for the diagnostics of the basic properties of nervous system require special conditions of realization and equipment. They are rather labor-consuming. Therefore in this research for the definition of steadiness and mobility of the nervous system the authors used the questionnaire Y. Strelyau [Столяренко, 2000, c.180], and a Tepping-test [Столяренко, 2000, c.187] for the definition of the force of nervous system.

The results of the first stage of an experimental research (the estimation of individual features of the students' personality) have shown:

1. Among the students of the three specialties (AS - 83 %, AP - 86 %, AC - 63 %) the great majority displayed propensity for trades, whose labor objects are both / either technical and / or sign information, i.e. for the activity that presuppose the usage of various machines, materials, other products of a civilization and / or based upon the preferences to processing figures, letters, codes, other symbols.

Thus, it was found out, that for the quota of the students participating in the experiment, the choice of trade is basically adequate to the person's propensities. Hence, it is possible to assume a positive motivation of the students and their interest in mastering material in the Electrical Engineering.
2. The majority of students, participating in the experiment, have a high or middle level logical ability of thinking: in the stream AP - 83 % from the general number of the students of this specialty, AC - 97 %, AS - 95%. Therefore, the students participating in the experiment have a sufficient level of the logical ability of thinking for perception and mastering of the educational material in GED.
3. The students of all the three specialties have weak or middle-weak nervous system. The advantage of the weak nervous system over a stronger one is in its capability to react to stimuli of lower intensity. The weak nervous system is more delicately organized, is more sensitive. Hence, the influence of various properties of CTP on the students' functional condition during the work with the given programs will be displayed to a great extent..

## The second stage of the research

The estimation of the ergonomic parameters of the computer training programs and the students' functional state during the work with CTP on Electrical Engineering was carried out in practical lesson in a display classes during few lessons.
While operating a computer the significant part of the information is perceived by the person through the visual analyzer. The increase of quality of the displayed visual information can be performed in two directions: the optimization of the light mode (the brightness, the light exposure in the room, the contrast of elements on the screen) and optimization of human-machine dialogue. It is necessary to note, that the continuous perfection of computer engineering leads to the reduction of the urgency of an estimation and control of hardware parameters of display. Therefore ergonomists' attention is concentrated today, mainly, on the area of the software of the processes of the information display. The present research will be carried out within the framework of the given direction.
On the basis of the analysis of the scientific literature, devoted questions of the ergonomic designing of software of various purposes and a long-termed practice of using electronic textbooks in educational process in ASTU it was suggested in the present work that the following ergonomic parameters should be introduced for the estimation of the quality of the training programs:

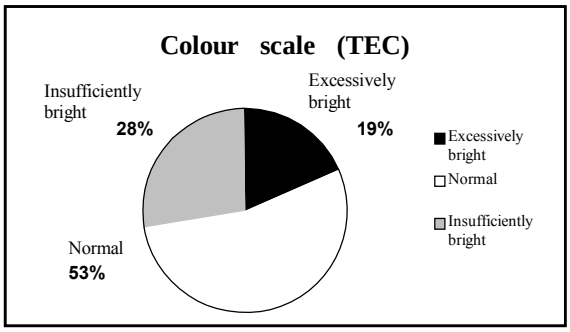
1. The used color scale: excessively bright; normal; insufficiently bright.
2. The convenience of reading the information: small print, normal; excessively large.
3. The spatial arrangement of elements of the information on the screen: inconvenient; rather convenient; convenient.
4. Dynamic presentation of the information (moving objects on the screen): promotes a better understanding of the educational information; does not improve the perception of the educational information; irritates.
5. The degree of clearness of the sequence of actions during the work with the program: the sequence of actions clear; sometimes there are difficulties in understanding the sequence of actions; the sequence of actions not clear.

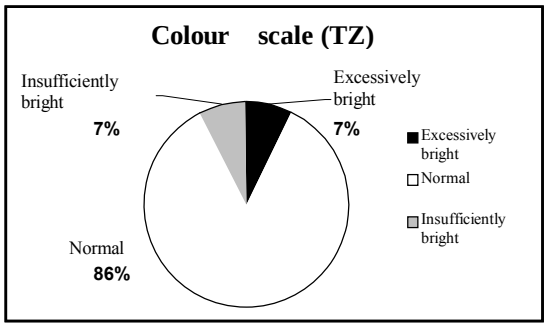
The authors used the method of filling a questionnaire. The students estimated a particular computer training program according to the given parameters at the end of the lessons in this program.
The results of an experimental research of the ergonomic properties of TZ and TEC are show in pict.2. The diagrams illustrate a great advantage of TZ over TEC in all parameters: a spatial arrangement of the elements of the information, the degree of clearness of the sequence of actions, a color scale, the convenience of reading of the information (a print).
While estimating the students' functional state (FS) during the work with CTP it was necessary to meet some requirements:

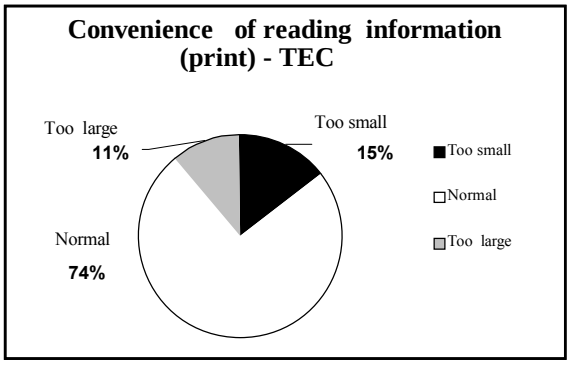
1. The estimation techniques should be reliable and valid.
2. The estimation techniques FS should be convenient for a group testing.
3. The FS estimation should not interrupt the plan of the educational process. The time spent on the FS estimation should not exceed 10 minutes (5 minutes at the beginning of the lessons and 5 minutes at the end).

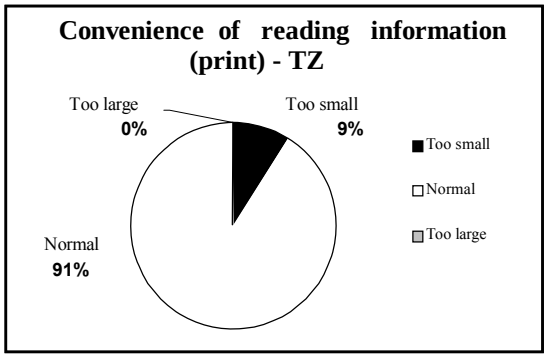
For an estimation of the functional state (FS) ergonomics uses two type of methods: physiological and psychological [Эргономика, 1988]. From physiological methods of the estimation FS the given research uses the measurement of pressure and pulse.

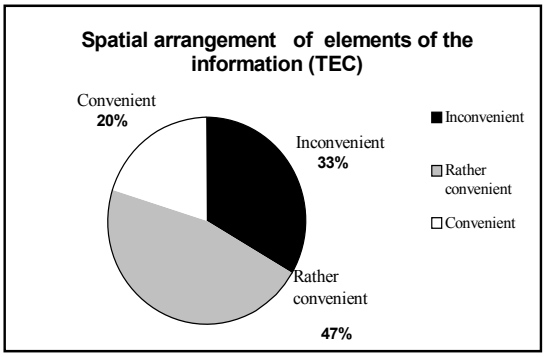
Picture 2. THE ESTIMATION OF THE ERGONOMIC PARAMETERS
OF THE TEC AND TZ COMPUTER PROGRAMS

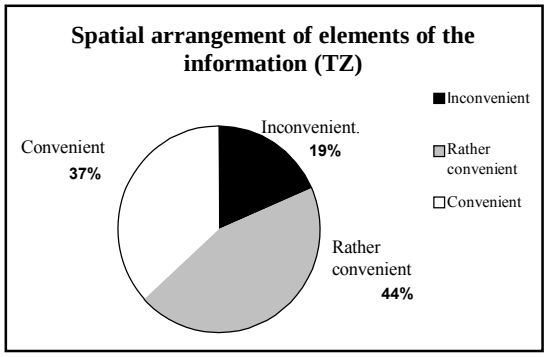


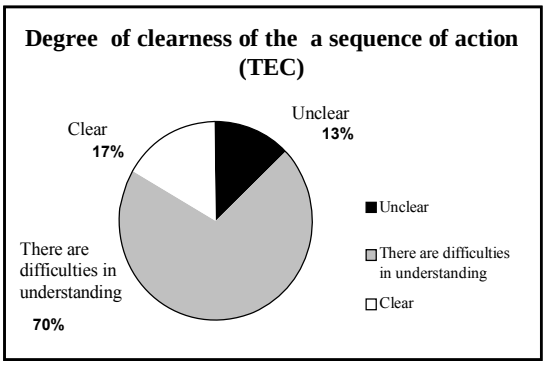



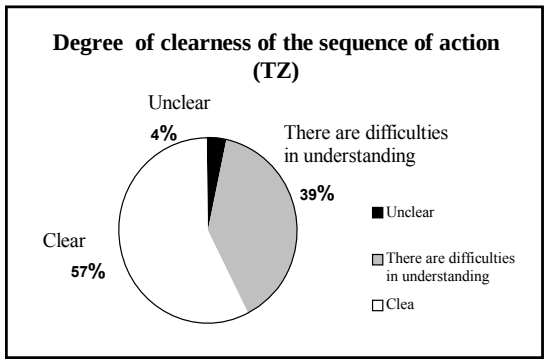



These physiological parameters were measured selectively of some students at the beginning and at the end of the lesson. The students for the control were chosen on the basis of the results of the tests on the properties of nervous system. Students with the weakest and the least steady nervous system were chosen as well as with the strongest and the steadiest nervous systems.

For diagnostics of the FS changes psychological tests for the volume of short-term visual memory (a test for storing of numbers) and also volume of distribution and switching of the attention («A Numerical square») test were used [Рабочая книга …, 1996, с.172-174].

The results of the FS estimation with the usage psychometric techniques have shown:

The reduction of the volume of the short-term memory during work with TZ was registered of almost half of students, whereas during the work with TEC the reduction of the volume of short-term memory was observed of a smaller number of students (AS - 19 %, AP + AC - 28 %).

The reduction of the volume of distribution and switching of the attention during the work with TZ is lager that with TEC (TZ -43 %, TEC- 30 % accordingly).

It is possible to explain the received results by a greater intensity of the learning activity of the students during the work with the TZ program (the students do the calculation of electrical circuits, draw vector diagrams and get a mark for each task). The work with the modeling program TEC is reduced only to observance of diagrams and vector diagrams, a mark for the lesson is not given.

For a subjective estimation of their functional state at the end of the lesson the students were offered a test, prepared on the basis of the test of the differentiated self-estimation (SAM) [Столяренко, с.367]. The parameters were chosen according to the scales «State of health», «Activity», «Mood» that has the clearest and the most precise formulation. The examinees were asked to link their sensations with a number of properties; the definition of each of them should be as curt as possible. The following parameters were estimated:

1. The self-estimation of state of one's health (feels well; badly; it's difficult to answer).
2. Tension (tense, relaxed, it's difficult to answer).
3. Vivacity (vigorous, languid, it's difficult to answer).
4. Mood (good, bad, it's difficult to answer).
5. Satisfaction from the work (satisfied, not satisfied, it's difficult to answer).
6. The degree of concentration of attention (attentive, absent-minded, it's difficult to answer).
7. The degree of excitation (exited, sleepy, it's difficult to answer).

The analysis of the results of the students 'state after the work with TZ and TEC program has shown:

· A significant advantage of TZ over TEC according to the parameters: «vivacity» (by 19 %), « satisfaction with the work » (by 24 %).

· An advantage of TZ over TEC according to the parameters: «state of health» (by 13 %), «mood» (by 11 %).

· The both programs have caused the same degree of tension.

· TZ has required a higher degree of concentration of attention and caused a greater excitation of the students. Apparently, it is explained by the fact, that on the results of his work with TZ the student is given a mark, but after the work with TEC marc is not given.

## Conclusion

The results of the present research have shown the validity of the developed technique for an experimental estimation of the ergonomic parameters of CTP. The present investigation has carried out the control and a comparative analysis of the ergonomic parameters of a number of computer training programs on Electrical Engineering, one of the general engineering disciplines. The received results have shown the advantage of the TZ computer training program, which was worked out on the basis of the method of theoretical images over the TEC modeling program.

The account of the ergonomic aspects of perception of the information is very important for designing programs of educational purpose. Creating a comfortable interface, we shall promote the students ' effective mastering of the knowledge, form their positive attitude to the learning educational activity, what causes the adequate mobilization psycho-physiological processes, postpones the development of exhaustion, promotes a long and highly effective serviceability without damage to health.

## Bibliography

[Бакалов и др., 1998] Бакалов В.П., Крук Б.И., Журавлева О.Б. Компьютерный учебник по теории электрических цепей // Материалы четвертой междунар. науч.-метод. конф. «Новые информационные технологии в преподавании электротехнических дисциплин». - Астрахань: АГТУ, 1998. – С.58-62.

[Введение в эргономику, 1974] Введение в эргономику. / Под ред. Зинченко В.П. –М.: Сов. Радио, 1974. -352с.

[Горбатов, 2000] Горбатов Д.С. Практикум по психологическому тестированию: Учебное пособие – Самара: Издательский дом «Бахрам-Х», 2000. –С.84.

[Зайнутдинова, 1999[1]] Зайнутдинова Л.Х. Создание и применение электронных учебников (на примере общетехнических дисциплин) // - Астрахань: Изд-во «ЦНТЭП», - 364с.

[Зайнутдинова, 1999[2]] Зайнутдинова Л.Х. Теоретические основы создания и применения дидактических интерактивных программных средств по общетехническим дисциплинам. Дис. докт. пед. наук. Астрахань-Москва, 1999. – 410с.

[Зайнутдинова, 1997] Зайнутдинова Л.Х. Создание теоретических образов как метод повышения эффективности электронных учебников // Материалы науч.-технич. конф. «Новые информационные технологии в региональной инфраструктуре (НИТ РИ-97)». - Астрахань: АГТУ, 1997. – С.163-167.

[Зараковский и др., 1993] Зараковский Г.М., Мунипов В.М., Шлаен П.Я. Эргономика в вопросах и ответах. Материалы понятийной базы эргономики. – Тверь, 1993. –68с.

[Ломов, 1991] Ломов Б.Ф. Проблема образа в психологии. В кн.: Вопросы общей, педагогической и инженерной психологии. (Труды д.чл. и чл.-кор. АПН СССР). - М.: Педагогика, 1991. - С. 65-72.

[Рабочая книга…, 1996] Рабочая книга практического психолога: Технология эффективной профессиональной деятельности – М.: ИД «Красная площадь», 1996. – 400с.

[Словарь…, 1998] Словарь практического психолога. / Сост. С.Ю. Головин. – Минск: Харвест, 1998. –800с.

[Столяренко, 2000] Столяренко Л.Д. Основы психологии. Практикум. –Ростов н/Д.: Феникс, 2000. –566 с.

[Чайнова, 1985] Чайнова Л.Д. Функциональный комфорт как обобщенный критерий оптимизации трудовой деятельности. // Техническая эстетика. –1985, -№2. –С.19-20.

[Эргономика, 1988] Эргономика: Учебник для вузов по спец. «Психология». / Под ред. А.А. Крылова, Г.В. Суходольского – Л.: ЛГУ, 1988. –182с.

[Bacalov etc., 1998] Bacalov V.P., Kruk B.I., Guravleva O.B. The computer textbook under the theory of electrical circuits // Materials by fourth international conference «New information technologies in teaching electrotechnical disciplines». - Astrakhan: ASTU, 1998. - pp.58-62.

[Introduction in ergonomics, 1974] Introduction in ergonomics. / Editor: Zinnchenko V.P. -M.: Soviet radio, 1974. –352p.

[Gorbatov, 2000] Gorbatov D.S. A practical work on psyhological testing: the manual. – Samara, The publishing house «Bahram-H», 2000. -p.84.

[Zainutdinova, 1999[1]] Zainutdinova L.H. Creation and application of the electronic textbooks (after the example of traditional technical disciplines) // - Astrakhan, «ЦНТЭП», - 364p.

[Zainutdinova, 1999[2]] Zainutdinova L.H. Theoretical bases of creation and application of didactic interactive software on traditional technical disciplines. Dissertation. Astrakhan-Moscow, 1999. – 410p.

[Zainutdinova, 1997] Zainutdinova L.H. Creation of theoretical images as a method of increase of efficiency of the electronic textbooks // Materials of the conference «New information technologies in a regional infrastructure (НИТ РИ-97)» . - Astrakhan: ASTU, 1997. - pp.163-167.

[Zarakovsky etc., 1993] Zaracovsky G.M., Munipov V.M., Shlaen P.Y Ergonomics in questions and answers. Materials of conceptual base of ergonomics. - Tver, 1993. –68p.

[Lomov, 1991] Lomov B.F. A problem of an image in psychology. In books: Questions of general(common), pedagogical and engineering psychology. (APH USSR). - M.: Pedagogics, 1991. - pp. 65-72.

[The Working book…, 1996] The Working book of the practical psychologist: technology of effective professional activity (manual for the experts working with the personnel). - M.: Publishing house «Red area », 1996. – 400p.

[Dictionary…, 1998] Dictionary of the practical psychologist. /Compiling S .U. Golovin. - Minsk: Harvesr, 1998. –800p.

[Stolyarennko, 2000] Stolyarennko L.D. Bases of psychology. A practical work. –Rostov-Don.: Fenix, 2000. –566p.

[Chainova, 1985] Chainova L.D. Functional comfort as the generalized criterion of optimization of labour activity. // An industrial Art. -1985, - № 2. -pp.19-20.

[Ergonomics, 1988] Ergonomics: the textbook for high schools on .speciality «Psychology». /Under edit. A.A. Krulov, G.V. Suhodolsky - Л.: LSU, 1988. –182p.

## Author information

Yakovetc Diliara Ahtamovna - Astrakhan state technical university, assistant of department «The Automated systems of processing of the information and management »; Russia, 414025, Astrakhan, Tatichev street, 16; e-mail: dl_sun@mail.ru ;

Zainutdinova Larisa Hasanovna - Astrakhan state technical university, Head of electrotechnics department, doctor of pedagogical sciences, professor; Russia, 414025, Astrakhan, Tatichev street, 16; e-mail: zain@astranet.ru

# THE STRUCTURE OF INFORMATION DIALOGUES: A CASE STUDY

*M. Koit*

*Abstract: In the paper we consider the structure of information dialogues. Our study is based on Estonian dialogue corpus which contains two kinds of dialogues – transcriptions of spoken conversations, and dialogues collected with the Wizard of Oz method. We are using two ways for describing the structure of dialogues – a typology of dialogue acts, and a system of communicative strategies. We depart from the notion of communicative strategy introduced by Kristiina Jokinen in her Constructive Dialogue Model. The analysis of our empirical material shows that people are using similar communicative strategies in telephone conversations and computer interactions. In the same time, the structure of human-human conversation is much more complicated.*

*Keywords: Computer intellectualization*

## Introduction

Estonian dialogue corpus consists of two kinds of dialogues. Firstly, 255 spoken dialogues are recorded and transliterated by the transcriptional system of conversation analysis (Jefferson 1979). 150 of the dialogues are telephone conversations where a person calls an office (railway station, bus terminal, travel agency, etc.) aiming to get some information. The remained 100 are face-to-face conversations. Secondly, we have collected 20 dialogues by the Wizard of Oz (WOZ) method. All the WOZ dialogues are information requests. The participants of our WOZ experiments were allowed to ask questions about bus schedule in Estonia and ship or plain traffic between Estonia and Finland. Therefore, we have a reasonable number of information dialogues in our corpus.

Building our corpus, we have two goals. The first goal is studying of spoken human-human conversation, and the second is modelling of human-computer interaction. Our further aim is to build an experimental dialogue system which could act as a rational agent and provide the needed information to the user. The dialogue system will integrate several language technology modules built up for Estonian so far (morphological and syntactic analysis, text-to-speech synthesis etc.). To work out a dialogue manager, we are studying the structure of our information dialogues.

There are several ways to describe the dialogue structure. From one side, we can use a system of dialogue acts and represent dialogue as a sequence of such acts. From the other side, communicative strategies for achieving certain communicative goals can be found in dialogues, and dialogue can be represented as implementation of the strategies. Both of these developments are methods for expressing and achieving the coherence of dialogues.

## Typology of Dialogue Acts

There exist several typologies of dialogue acts. The first well known typology was worked out by J. Sinclaire and M. Coulthard on the ground of the study of real dialogues (Sinclair, Coulthard 1975). The system of dialogue acts was further developed by A.-B. Stenström (Stenström 1994). Several researchers are considering practical problems of dialogue acts determination during the last decade – corpus linguists, discourse and conversation analysts, language technologists (Hakulinen 1989; Allwood et al. 2000; Stolcke et al. 2000; Jokinen et al. 2001).

Choosing a dialogue act mark-up system we have had two goals: to study spoken human-human conversation and to model human-computer interaction. We started with analysis of existing dialogue act systems and typologies (Klein, Soria 1998, Francis, Hunston 1992, Stenström 1994, Dybkjær 2000). It proved difficult to take over a ready-made typology because most of them are domain-oriented (eg. furnishing an apartment, guessing a journey on the map, determining a meeting, etc.). Therefore, we decided to work out our own typology. We departed from the Stenström system which is based on conversation analysis.

There are 140 dialogue acts in our system divided into 8 groups:

1) rituals – greeting, introducing, etc.;

2) acts for re-structuring of conversation, with help of which the speaker starts a new topic or changes the type of conversation;

3) acts for exchanging of turn-takings, with help of which the speaker is asked to continue, or the existence of contact is checked;
4) repairing acts, with help of which partners are solving communication problems;
5) directive acts for giving and receiving of commands, requests, etc.;
6) questions and answers – pairs of acts, with help of which one partner asks a question and another answers it;
7) acts for taking up of attitudes, with help of which one partner represents an attitude (belief, evaluation, charge) and another responds it;
8) the last group contains the remaining acts (additional information, argument, conclusion, promise, acknowledgement, signal of new information, etc.).

The acts from all the groups, except of the last, can form adjacency pairs. For that reason, they are divided into 2 sub-groups: the first and second parts. The first parts are used to give commands, ask questions, etc. The second parts express reactions to commands, answers to questions. Acts from the 8th group can supplement both the first and second parts.

A simplified formal grammar determining our dialogue acts system is as follows (cf. Koit 2001). The terminals (dialogue act names) are written in capitals.

```
  interaction::= (transaction)+
  transaction::= (exchange)+
  exchange::= organisational-exchange | conversational-exchange
  organisational-exchange::= ritual | repair | CONTINUER
  ritual ::= CALL RESPONDING-THE-CALL | GREETING RESPONDING-THE-
GREETING | THANKING RESPONDING-THE-THANKING | LEAVE-TAKING RESPONDING-
THE-LEAVE-TAKING
  repair::= hearer-initiated-repair | self-repair
  hearer-initiated-repair::= INITIATING-REPAIR CARRING-OUT-REPAIR |
INITIATING-REPAIR CARRING-OUT-REPAIR EVALUATION
  initiating-of-repair ::= NON-UNDERSTANDING | RE-QUESTION | SPECIFYING-
CONDITONS-OF-THE-ANSWER
  self-repair::= REFORMULATION
  conversational-exchange::= directive-exchange | question-exchange
  directive-exchange::= directive's-pre-member directive's-re-member
  directive's-pre-member::= ORDER | REQUEST | PROPOSAL | WISH | CALL-UP |
OFFER | REQUEST-TO-WAIT
  directive's-re-member::= FULFILMENT | REFUSAL | AGREEMENT | POSTPONING-
THE-ANSWER | FULFILMENT-WITH-RESERVATIONS | YOU-ARE-WELCOME
  question-exchange::= question's-pre-member question's-re-member
  question's-pre-member::= CLOSED-YES/NO-QUESTION | OPEN-YES/NO-QUESTION
| WH-QUESTION | SPECIFYING-THE-CONDITIONS-OF-ANSWER
  question's-re-member::= AGREEMENT-(YES)| AGREEMENT-(NO) | NON-AGREEMENT
| open-answer | POSTPONING-THE-ANSWER | ANSWER-AS-AN-ALTERNATIVE
  open-answer::= GIVING-INFORMATION | INDICATING-THE-ABSENCE-OF-
INFORMATION
```

Our typology does not allow to annotate dialogues on several levels as it is possible, for example, in DAMSL (Allen et al., 1997). However, some levels can be differentiated indirectly. Communicative status is indicated by the dialogue act REFUSAL which marks a non-interpretable or unfinished utterance. Information level is expressed by the conversational exchanges (as opposite to organisational ones). The role of forward-seeking functions is played by the first parts (pre-members) and the role of backward-seeking ones by the second parts (re-members) of adjacency pairs. Our scheme is more detailed as DAMSL. For example, the group of rituals consists of 34 acts (there are only 2 acts in DAMSL – opening and closing). Such particularity is very useful for study of human-human conversation even though it makes the annotation process more difficult. If we had only one goal – training a question-answering system – then we could to be satisfied with a more superficial typology of acts. But our primary goal is to study human-human conversation.

Our studies are currently centred on information seeking dialogues. We are using our system for annotating our corpus. Supposedly, the typology can be reduced in process of the work.

## Dialogue Acts in Information Dialogues

For this paper, we annotated 10 spoken (telephone) and 10 WOZ dialogues from our corpus. It is possible to outline the structure of information-seeking dialogue as consisting of four parts with different functions (Fig. 1).

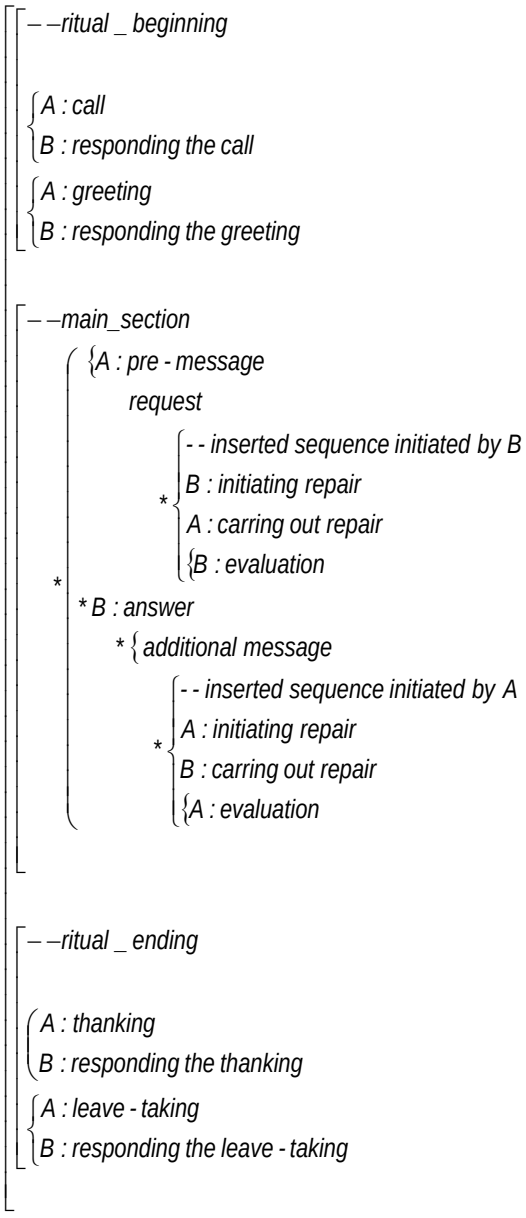



Figure 1. The structure of information dialogue. Notations: ( – adjacency pair, [ – connects the whole dialogue or its section, { – dialogue act, adjacency pair or triad which is optional, * – dialogue act, adjacency pair or triad which can be repeated, -- – start of comment.

The four parts are

- a ritual beginning (greeting, introducing etc.);
- a ritual ending (thanking, farewell);
- requesting and giving information (answering questions, giving telephone numbers, etc.);
- solving communication problems (misunderstanding, inaudibility, unreliability of information) in cooperation of partners. This part often follows after the first question and forms an inserted sequence within the first adjacency pair, also it can be repeated within the following adjacency pairs.

Ritual parts can be missed in conversations. It is usual in WOZ dialogues that the user (A) does not greet the computer (B), (s)he starts interaction with request. A's information request is expressed by directive's or question's pre-member (usually, open yes/no question, wh-question or wish). Pre-messages can be added to request (for example, 'I have a question'). B's answer is expressed as directive's or question's re-member,

usually as open answer: giving information. B often asks adjustable questions to specify the conditions of answer.

Let us consider two examples from our corpus (cf. Examples 1 and 2). The first dialogue is a telephone conversation and the second one is a WOZ dialogue. In the last case, the user put in his/her questions from the keyboard, and got answers from the wizard on the screen. The ritual beginning and ending parts are put out in the examples.

Example 1. A – client, B – a travel clerk. (Translated from Estonian.)

| *No* | *Utterance* | *Dialogue act* |
|---|---|---|
| 1 | A: I'm interested in trips to Scandinavian states. | WISH |
| 2 | B: Yes? | YOU ARE WELCOME |
| 3 | More precisely? | SPECIFYING THE CONDITIONS OF ANSWER, POSTPONING THE ANSWER |
| 4 | A: Which variants do you have | WH QUESTION, POSTPONING THE ANSWER |
| 5 | to Sweden, Norway? | ADDITIONAL INFORMATION: SPECIFICATION |
| 6 | B: mmm… You can buy tickets by us. | ANSWER AS AN ALTERNATIVE |
| 7 | A: So. | ACKNOWLEDGEMENT, SIGNAL OF NEW INFORMATION |
| 8 | B: Plane and ship tickets. | ANSWER AS AN ALTERNATIVE, ADDITIONAL INFORMATION: SPECIFICATION |
| 9 | Unfortunately, we don't offer a whole travel packet. | OPEN ANSWER: ANOTHER |
| 10 | A: So. | ACKNOWLEDGEMENT, SIGNAL OF NEW INFORMATION |
| 11 | B: I mean a group trip. | ADDITIONAL INFORMATION: SPECIFICATION |
| 12 | A: mmm | CONTINUER |

Various adjacency pairs appear in our dialogues: in the first example, A's wish (1st utterance) is closed by B's open answer (9), A's open question (4) is closed by B's answer as alternative (6), etc. In the second example, A's open question (1st utterance) is closed by B's open answer (8-11). Any adjacency pair can contain other pairs within. In the second example, the pair open question – open answer contains a pair adjustable question – open answer: giving information (utterances 4 and 5). An utterance can hold more than one function, as utterance 3 in the first example (postponing the answer that closes the utterance 1, and adjustable question that opens a new adjacency pair).

Some of dialogue acts do not occur in WOZ dialogues, for example, checking contact, continuer, taking up of attitudes. In our experiments, clients seldom asked questions which could be answered shortly, i.e. giving information about one or two trips. For that reason the computer asked very many adjustable questions followed by an open answer: giving information. In some degree, the questions were necessary for the wizard to extend the time for information retrieval.

Example 2. A – user, B – WOZ. (Translated from Estonian.)

| *No* | *Utterance* | *Dialogue act* |
|---|---|---|
| 1 | A: When will the bus depart from N. to Tallinn, | OPEN QUESTION |
| 2 | in order that to arrive in Helsinki by ship at 12 o'clock? | ADDITIONAL INFORMATION |
| 3 | B: Wait please! | REQUEST TO WAIT |
| 4 | Are you interested in a certain week-day? | SPECIFYING THE CONDITONS OF ANSWER |
| 5 | A: I am interested in workdays | OPEN ANSWER: GIVING INFORMATION |
| 6 | B: Do you agree to spend a night in Tallinn? | SPECIFYING THE CONDITONS OF ANSWER |
| 7 | A: Yes. | AGREEMENT (YES) |
| 8 | B: The bus departs at 9.10 p.m. from N. | OPEN ANSWER: GIVING INFORMATION |
| 9 | Arrives at 10 p.m. in Tallinn. | OPEN ANSWER: GIVING INFORMATION |
| 10 | The ship departs at 8 a.m. from Tallinn, arrives at 10.30 a.m. in Helsinki, | OPEN ANSWER: GIVING INFORMATION |
| 11 | or departs at 9.30 a.m., arrives at 12. | OPEN ANSWER: GIVING INFORMATION |
| 12 | Do you need an earlier bus from N.? | CLOSED YES/NO QUESTION |
| 13 | A: No. | NON-AGREEMENT |
| 14 | Thank you. | THANKING |

The wizard asked such adjustable questions as 'Do you prefer a certain time-interval?', 'Do you prefer a certain weekday?', 'Are you interested in certain arrival time?' and so on.

The wizard often needed relatively long time to compile answers. In order that the user did not suspect of an error of the program or slow Internet-connection, the wizard used the dialogue act request to wait in the form 'Wait please!'.

## Communicative Strategies in Information Dialogues

Let us depart from the notion of communicative strategy, considered in (Jokinen 1996a,b) as a part of the Constructive Dialogue Model (CDM). The departure point of the CDM is in general communicative principles which constrain cooperative and coherent communication. Dialogue participants are engaged in a cooperative task whereby a model of the joint purpose is constructed. Contributions are planned as reactions to the changing context. Communicative strategy is used by a participant to build up the next turn as a reaction to partner's previous one. Thus, communicative strategies express the coherence of the dialogue similarly as adjacency pairs of dialogue acts. Four context factors are used in CDM to determine communicative strategies:

1) expectations – is the turn expected or not;
2) the central conception – does the partner's turn keep the topic or not (related or unrelated);
3) goals – are the speaker's goals fulfilled or not;
4) initiatives – has the speaker initiative or not.

The first two parameters are hearer-related and the last two speaker-related.

All the context factors have binary values (1 or 0) in CDM which gives $2^4$=16 communicative strategies. Every strategy can be represented by a vector of factors with coordinate values 1 or 0, for example, finish/start (vector 1111, i.e. expected-related-fulfilled-speaker), new request (0010, i.e. non-expected-unrelated-fulfilled-partner), subquestion (0101), follow-up old (1100), object (0001), etc.

By means of communicative strategies changing initiatives, achieving goals, changing topics, digressing from normal talk can be traced in dialogue structure.

We annotated dialogue strategies in 10 spoken and 10 WOZ dialogues. The same information-seeking dialogues were analysed as for dialogue acts. The more frequent strategies were follow-up-old (represented by the vector 1100, i.e. expected-related-unfulfilled-partner), finish/start (1111), subquestion (0101) in spoken as well as in WOZ dialogues. Wizard often implemented the strategy unrelated (0000) ('Wait please!' in our example) which is unusual in spoken dialogues. From the other side, there are more changes of topic in telephone conversations as the WOZ dialogues. The user more strictly keeps the topic when interacting with the computer. Likewise, the initiative more often goes from one participant to the other in telephone conversations. Wizard tried to keep initiative and control interaction. The general structure of information

$$
\left[\; * \left[\; \begin{array}{l} \text{A: finish/start} \\ * \left\{ \begin{array}{l} \text{- - inserted sequence initiated by B} \\ \text{B: subquestion} \quad | \text{ object} \\ \text{A: follow - up old} \;|\; \text{continue} \;|\; \text{backto} \end{array} \right. \\ \text{B:} \left\{ \begin{array}{l} \text{notrelated} \\ \text{follow - up old} \;|\; \text{continue} \\ \{\text{somethingselse} \end{array} \right. \\ * \left\{ \begin{array}{l} \text{- - inserted sequence initiated by A} \\ \text{A: subquestion} \\ \text{B: follow - up old} \end{array} \right. \\ \{\text{specify new} \;|\; \text{new dialogue} \end{array} \right. \\ \{\text{A: continue} \end{array} \right.
$$

dialogue is represented on Figure 2. The ritual beginning and ending parts are omitted.

Figure 2. The structure of information dialogue: communicative strategies. Notations: [ – connects the whole dialogue or its section; { – an optional strategy or a sequence of strategies, * – strategy or a sequence of

strategies which can be repeated, | – variants of strategies; -- – start of comment.

Let us go back to the examples and use now communicative strategies for expressing the structure of dialogue (Examples 3-4).

Example 3 (cf. Example 1). A – client, B – a travel clerk.

| No | Utterance | Vector of factors | Strategy |
|---|---|---|---|
| 1 | A: I'm interested in trips to Scandinavian states. | 1111 | finish/start |
| 2 | B: Yes? | 1100 | follow-up-old |
| 3 | More precisely? | 0101 | subquestion |
| 4 | A: Which variants do you have | | |
| 5 | to Sweden, Norway? | 1101 | backto |
| 6 | B: mmm… You can buy tickets by us. | 0101 | continue |
| 7 | A: So. | 1100 | follow-up-old |
| 8 | B: Plane and ship tickets. | 1100 | follow-up-old |
| 9 | Unfortunately, we don't sell a whole travel packet. | 1100 | follow-up-old |
| 10 | A: So. | 1100 | follow-up-old |
| 11 | B: I mean a group trip. | 1100 | follow-up-old |
| 12 | A: mmm | 0001 | object |

Example 4 (cf. Example 2). A – user, B –WOZ.

| No | Utterance | Vector of factors | Strategy |
|---|---|---|---|
| 1 | A: When will the bus depart from N. to Tallinn, | | |
| 2 | in order that to arrive in Helsinki by ship at 12 o'clock? | 1111 | finish/start |
| 3 | B: Wait please! | 0000 | unrelated |
| 4 | Are you interested in a certain week-day? | 0101 | subquestion |
| 5 | A: I am interested in workdays | 1100 | follow-up-old |
| 6 | B: Do you agree to spend a night in Tallinn? | 0101 | subquestion |
| 7 | A: Yes. | 1100 | follow-up-old |
| 8 | B: The bus departs at 9.10 p.m. from N. | | |
| 9 | Arrives at 10 p.m. in Tallinn. | | |
| 10 | The ship departs at 8 a.m. from Tallinn, arrives at 10.30 a.m. in Helsinki, | | |
| 11 | or departs at 9.30 a.m., arrives at 12. | 1100 | follow-up-old |
| 12 | Do you need an earlier bus from N.? | 0111 | new dialogue |
| 13 | A: No. | | |
| 14 | Thank you. | 1110 | follow-up-new |

## Discussion and Conclusion

When constructing the next utterance, a participant must act cooperatively and follow certain conversational norms. The reason is that dialogue can be considered as a negotiation process where each participant is responsible for continuation of communication. When we are speaking in terms of communicative acts, it means that there are certain acts that typically can follow an act, and if a speaker does not choose one act from this set then it can be treated as a violation of the norm. From the other side, when we are speaking in terms of communicative strategies, then context factors determine the next strategy, and similarly, they guarantee the coherence of interaction.

A many-to-one mapping can be determined from the set D of dialogue acts to the set S of communicative strategies. The strategies where the speaker has initiative correspond to the first parts of adjacency pairs, and any act sets up a new goal. For example, wish and open question represent the finish/start strategy, specifying the conditions of answer – the subquestion strategy, opposing – the continue strategy, etc. Therefore, when interacting with a user, the dialogue system which uses information both of dialogue acts and communicative strategies, is able to respond to the user more adequately.

In our previous work, we have considered argumentation dialogues and determined communicative strategy as an algorithm for achieving a certain communicative goal (Koit, Oim 2000a,b). We also determined communicative tactics as algorithms for building the next utterances. Tactics of enticement, persuasion and threatening were considered. Thus our communicative tactics correspond to communicative strategies in (Jokinen 1996a,b). So far, we were interested in such conversations where participants could have

antagonistic goals. The information-seeking communication, in opposite, is cooperative. Starting conversation, one of participants, A, has a communicative goal 'A get information P'. The communicative goal of the (cooperative) partner B is this same. This type of dialogues clearly will be the area where in the next few years already systems will be required that would be practically reliable, but at the same time could follow the rules of natural human communication.

Our further work will be concentrated on a formal model which integrates both a dialogue grammar and communicative strategies with our previous (a kind of BDI) model, and implementation of the model in information-seeking interactions.

This work was supported by Estonian Science Foundation (grant No 4555).

## Bibliography

[Allen et al., 1997] Allen, James, Mark Core. Draft of DAMSL: Dialog Act Markup in Several Layers; http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html, 1997, used 10.02.2003.

[Allwood et al., 2001] James Allwood, Ahlsen E., Björnberg M., Nivre J. Social activity and communication act-related coding In "Gothenburg Papers in Theoretical Linguistics 85. Dialog Coding – Function and Grammar. Göteborg Coding Schemas". Ed. Jens Allwood. University of Göteborg, 2001 , pp. 1–28.

[Dybkjær, 2000] Dybkjær, Laila. MATE Deliverable D6.2. Final Report; http://mate.nis.sdu.dk/about/deliverables.html, 2000, used 10.02. 2003.

[Francis, Hunston, 1992] Francis, Gill, Susan Hunston. Analysing everyday conversation. Advances in Spoken Discourse Analysis. Ed. By Malcolm Coulthard. Routledge, London and New York, 1992.

[Hakulinen, 1989] Auli Hakulinen Keskustelun luonnehtimisesta konteksti- ja funktionaalisten tekijöiden nojalla In Kieli, No 4. Suomalaisen keskustelun keinoja 1. Ed. by Auli Hakulinen. Helsinkin yliopiston suomen kielen laitos. Helsinki, 1989, pp. 41–72.

[Jokinen et al., 2001] Kristiina Jokinen, Hurtig T., Hynnä K., Kanto K., Kaipanen M., Kermanen A. Self-organising dialogue management In Proceedings of the NLPRS Workshop Neural Networks and Natural Language Processing, Tokyo, Japan, 2001.

[Jokinen, 1996a] Kristiina Jokinen Cooperative Response Planning in CDM: Reasoning about Communicative Strategies. In "TWLT11. Dialogue Management in Natural Language Systems", S. LuperFoy, A. Nijholt & G. Veldhuijzen van Zanten, ed. Enschede: Universiteit Twente, 1996, pp. 159-168.

[Jokinen, 1996b] Kristiina Jokinen Rational Agency. In "Rational Agency: Concepts, Theories, Models, and Applications", M. Fehling, ed.. Proceedings of the AAAI Fall Symposium. MIT, Boston, 1996, pp. 89-93.

[Klein, Soria 1998] Klein, M., Soria, C. 1998. MATE Deliverable D1.1. Supported Coding Schemes. Dialogue Acts. – http://www.dfki.de/mate/d11/chap4.html, used 10.02.2003.

[Koit, 2001] Mare Koit Annotating dialogue acts in Estonian Dialogue Corpus In Proceedings of the Euroconference Recent Advances in Natural Language Processing (RANLP'01). Ed. G. Angelova, K. Bontcheva, R. Mitkov, Nicolas Nicolov, Nikolai Nikolov. Tzigov Chark, Bulgaria, 2001, pp. 273-275.

[Koit, Oim, 2000a] Mare Koit and Haldur Õim Dialogue management in the agreement negotiation process: a model that involves natural reasoning In The 1st SIGdial Workshop on Discourse and Dialogue. Ed. L. Dybkjaer, K. Hasida, D. Traum. HongKong, Association for Computational Linguistics (ACL), 2000, pp. 102-111.

[Koit, Oim, 2000b] Mare Koit and Haldur Õim Developing a model of natural dialogue. In "From spoken dialogue to full natural interactive dialogue-theory, Empirical analysis and evaluation. LREC2000 Workshop proceedings", L. Dubkjær, ed. Athen, 2000, pp. 18-21.

[Sinclair, Coulthard, 1975] J.M. Sinclair, Coulthard, R.M. Towards of Analysis of Discourse: The English used by Teachers and Pupils. London: Oxford UP, 1975.

[Stolcke et al., 2000] Andreas Stolcke, Coccaro N., Bates R., Taylor P., Van Ess-Dykema C., Ries K., Shriberg E., Jurafsky D., Martin R., Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech In Computational Linguistics, 26:3, 2000, pp. 339–373.

[Stenström, 1994] Anna-Brita Stenström An Introduction to Spoken Interaction. London and New York: Longman, 1994.

## Author information

Mare Koit – University of Tartu, J. Liivi 2, 50409 Tartu, Estonia; e-mail: koit@ut.ee

# FROM AN ONTOLOGY-ORIENTED APPROACH CONCEPTION TO USER INTERFACE DEVELOPMENT *

*Kleshchev Alexander, Gribova Valeriya*

*Abstract: The paper describes a new approach to user interface development which is an evolution of the model-based approach. The aim of the new, ontology-based approach is to eliminate the demerits of demerits of the model-based approach but to conserve its merits and, as a consequence, to lower more the cost of user interface development and maintenance. The main idea of our approach is to exchange models of different interface components for corresponding ontologies. The ontology models accessible by the Internet are used to form the models of there components.*

*Keywords: Ontology, interface model, user interface development*

## Introduction

A user interface is a central component of any modern software system. The efforts that are necessary to design, implement, modify and maintain a user interface add up to 70% of all the labour consuming for a software system development. Recent trends in development of software systems generally and of user interface in particular are applying the tools that free developers from low-level programming. These tools based on computer languages of the 4-th generation lower the cost of the development and maintenance for the applied software systems. There are a number of different tools for user interface development. But many user interfaces are implemented with interface builders as before. At the same time, there is a lack of tools that help designers to put all the pieces of an interface design together [1], no support for specifying the dynamic parts, and even for the static parts this support is not adequate [2,3]. These defects have given impetus to the development of a model-based approach for constructing user interface and now several model-based interface development tools have been built, for example [4,5,6,7]. The main goal of this approach is an automatic translation of the declarative, high-level models of interface components into an executable program[8,9]. As a result, the number of procedural components developed in the course of designing an interface becomes considerably less, there is a possibility to reuse the knowledge making up a model, there are powerful tools supporting development [10]. Except these merits, this new technology has also a few demerits that are discussed in [11]. In addition to them it is possible to point out the following. First, up till now there has been no universally accepted standard of interface components. As a result, every model-based tool defines its specific model interface components. In the second place, the methods for implementation of different interface model components by the same model-based tool are different. For every model specific principles and mechanisms are used. This is a reason for difficulties in linking these models together. In the third place, many these tools require the description of application program in detail. This property makes the interface development and maintenance difficult. In the fourth place, different model-oriented tools are based on different declarative languages and data models. This fact makes also a transfer of the same models from one tool to another difficult. In the fifth place, a universally accepted terminology has not been formed within the model-oriented approach yet. As a result, the components, which are identical by meaning and effect, often have different names.

The aim of a new, ontology-based approach to the user interface development advanced in this report is to eliminate the existing demerits of the model-based approach but to conserve its merits and, as a consequence, to lower more the cost of user interface development and maintenance.

## THE BASIC IDEAS OF THE ONTOLOGY-BASED USER INTERFACE DEVELOPMENT.

In this report four principal more precise definitions to the model-oriented approach are suggested.

1. A model of a user interface should be considered as a representation of such information about it that should be modified if some conditions of using the software system are changed. The information uniform by meaning should be combined into interface model components.

* The presented work has discussed on the KDS-2003. It has corrected in compliance with remarks and requests of participants.

Every component of the interface model should be represented in the form of an ontology model [12,13,14,15]. The representation of knowledge in the form of an ontology model is a universally accepted practice for development of knowledge based systems.

2. The ontology models accessible by the Internet should be used to form the interface model components for which it is possible.
The great current interest in ontologies is caused by the fact that the ontologies of different domains, represented in specific computer languages, can provide access of people as well as computer programs to a huge volume of information and knowledge stored in the Internet and give a possibility to software systems to use these ontologies and knowledge for solving different tasks. It is the ontologies that make possible the development the Internet of the second generation or semantic Internet [16]. The main problem of the semantic Internet is to give direct access of all comers to whole knowledge accumulated by the human civilization.
To implement the approach suggested in this report, it would be necessary to develop, store and maintain the user interface ontology in the Internet. This ontology could be used to form ontologies of particular user interfaces. The user interface ontology as well as the others has to be perpetually maintained.

3. The user interface and application should be designed and implemented as independent components that interact asynchronously through a set of common variables.
This improvement gives a possibility to do away with a task model description and a tool for linking the interface and the application. This idea can permit to decrease the cost of development and to make better interface maintainability.

4. The tool for interface development should be provided with a set of system functions. The interface designer should have a possibility to include necessary system functions in the developed interface according to the customer's requirements.
Any user interface has a set of possible functions according to the functions and tasks of user interface defined, for example, in [17]. These functions are to input data, to solve the task, to exit the application program, to view its results and so on

## THE COMPONENTS OF AN INTERFACE MODEL

The user interface model has to contain all the information about this interface that can be modified during the life circle of this interface. This model has also to be appropriate to automatic implementation of the interface (by translation or interpretation).
Before describing components of an interface model we answer the next questions. What is the user interface? What components it consists of? The Xerox Palo Alto Research Center and its official David Liddle have investigated this problem and presented all the user interface components in the form of an iceberg (fig.1). According to the research the interface consists of three main components: a presentation of information to users, an interaction, and a relation among objects. The visible iceberg part is greatly smaller than its invisible part. The iceberg top is information for users (color, animation, objects form, sound, graphic, positioning information). It aggregates only 10% of the whole information and isn't the main user interface component. The middle user interface part is practice of interaction and feedback with users. It aggregates 30% of the whole information. And, finally, the lower and main iceberg part aggregates 60% of the whole information. It includes objects properties and relations between them. On the basis of the research we determine the components of an interface model. Since a dialog with an software system carried on within the framework of a domain concept system, and the concept system can be modified during the life circle of the software system according to user's wishes and also according to modifications of the domain and of the program functions, the user interface model has to contain some information about this concept system. The domain concept system has to be appropriate to express the input and output data of the system, the information about the application program control, about the interface control and also about an intellectual supporting user's actions.

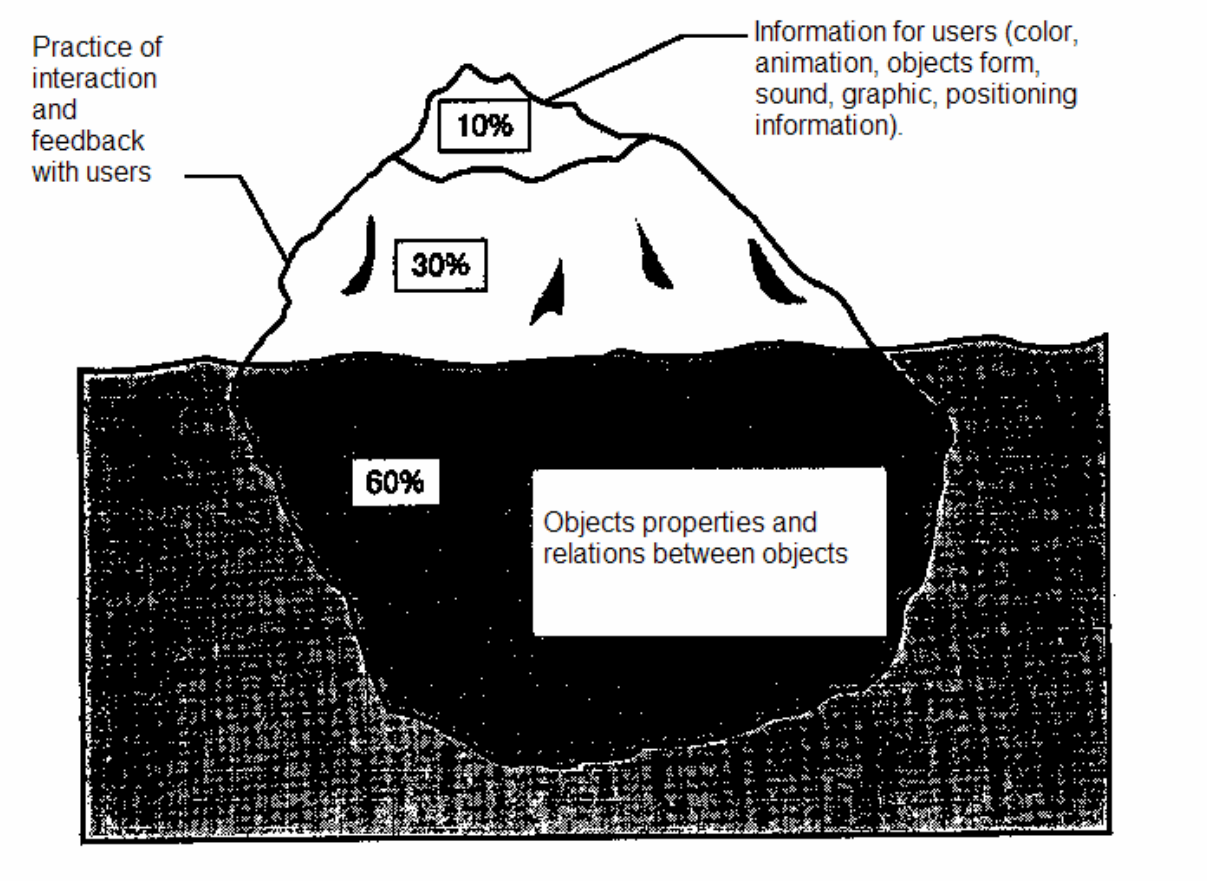



Fig. 1. User interface components in the form of an iceberg

Any user's dialog with the system is carried on using some display aspects of the interface such as methods and means for information transmission, for the control of user interaction with the application program, for dialog structures and so on. These display aspects can also be modified during the life circle of the system according to customer's wishes, to modifications of the program functions and to the development of our ideas about the display aspects of the interfaces. So this information about the display aspects has also to be a part of the interface model.

A software system consists of its user interface and application specific program (application program), which the user interface is closely connected with. The application program of the software system can be modified during its life circle according to modifications of the requirements to it. So the information about the application specific program has also to be a part of the user interface model. The less there is this information in the user interface model, the less it is probable that this information will have to be modified when the application specific program is modified, and so the better this part of the user interface model is.

The user interface presents the input and output information of the system to a user in terms of the domain concept system, but to the application specific program in the form of the values of its variables. In this manner there is a correspondence between the domain concept system and the set of the applied program variables. This correspondence can be modified when the domain concept system and/or variables of the application specific program are modified. So the information about this correspondence has also to be a part of the user interface model.

The display aspects of the user interface are used in a dialog to present in a certain form the information that is transmitted from a user to the application program and from it to the user. The user understands this information within the framework of the domain concept system. In this manner there is a correspondence between the domain concept system and the display aspects used in the user interface. This correspondence

can be modified according to a modification of the display aspects and of the domain concept system, and also according to user's wishes. So the information about this correspondence has also to be a part of the user interface model.

Any dialog is carried on according to a scenario. This scenario can be modified according to user's wishes, to a modification of the domain concept system and of the application program. So the information about the scenario of the dialog has also to be a part of the user interface model.

The customer's requirements to the set of system functions of the interface can also be modified in the course of using the application program. So the information about the set of system functions of the interface has also to be a part of the user interface model.

In this manner the model of any user interface of an software system can be considered as the set of the following models. They are the models of the domain concept system, of the display aspects for a presentation of these concepts in the interface, of the application program, of a correspondence between the domain concept system and the display aspects, of the scenario of a dialog and of the set of system functions.

## The Domain Concept System Model

The direction of attention towards the user means that the interaction between the user and the application program is realized in terms of the domain that the program is intended for. The domain is characterized by its concept system, which consists of concept definitions and of the descriptions for correspondences among them. An explicit representation of these definitions and correspondences is called domain ontology. Thus, a domain concept system model of every user interface is a formal model of a domain ontology.

The concept system used for an interaction between a user and an application program is a part of the concept system of an appropriate domain. It is possible to expect that in the near future formal ontology models for some domains will be presented in the Internet, and the number of models will be increased in time. The terminology of these ontologies will be standardized. If the domain ontology to which an software system relates has been formed, then the ontology of the concept system used by its user interface either is a part of this ontology or can be defined in terms of it.

Domain ontology often has the following property: many its components have the same structure. In this case it is convenient to form its description consisting of two levels. The first one is a reusable metaontology or an ontology of a domain class. It is the same for all the domains of the class. The second one is a domain ontology formed on the basis of the metaontology.

## A model for the display aspects of the interface

Now user interface development is an independent branch of software engineering, in which a rather stable concept system has been formed. The display aspects used in every specific user interface can be described as concrete definitions in terms of the concept system. Since designing the display aspects of a user interface is a professional activity, it should naturally be carried on within the framework of this professional concept system.

If this concept system is standardized, reduced to an ontology, formalized and made open to general use, designing the display aspects of the user interface will be considerably simplified. So practical use of the ontology-based approach to user interface development considerably depends on the fact, how quick the open to general use, standardized and formal and ontology model for the display aspects of the user interface will be formed. This ontology model has to contain descriptions for classes of objects having general structure and purpose. It should be possible to expand this ontology by adding new objects and their properties. The content of this branch of knowledge does not depend on a specific interface, but is determined by its achievements. Today a version of the graphic user interface ontology model is accessible by the Internet [18].

In that way, there are two possibilities for designing a model for the display aspects of the user interface in general case. The first one consists in extracting an appropriate concept system from the ontology model for the display aspects of the user interface had presented in the Internet. Extracting this concept system consists in defining the values of all the attributes in the definitions of general concepts. The second possibility is direct forming a specific model using the terminology accepted in this branch of knowledge.

## An Application Program Model.

Any software system can be considered as consisting of two main components which are the application program and the user interface. The application program solves some tasks, and the user interface supports an interaction between a user and the application program. According to [17], the user interface is intended for supporting an interaction between a user and an application program, that is a process fulfilling a task. The functions of this interaction are transmission of information for a control of running the application program, transmission of the input data from a user to the application program, of the output data from the application program to the user. In this manner, the interface should transmit the input data and maybe some control information to an application program, and the output data to a user.

An application program model is the better the less it contains information. It is obvious that the minimum information about an application program is a description of all the application program variables by which the exchange of information between the user and application program takes place. It may be considered that every variable is either input or output one, and has an identifier and a possible value range, i.e. an application program model can also be represented in the form of ontology.

When an application program ontology is described, it is necessary to choose an appropriate concept system for the description of the variables. This description is possible:

- - in terms of the domain concept system;
- - in terms of the implementation language;
- - in mathematical terms.

In the first case a transition from the application program ontology model to its implementation cannot be monosemantic. This fact will require an additional description of this transition.

In the second case using different implementation languages will require different models of the application program. This fact can considerably worsen their reusability.

In the third case the description of the variables in mathematical terms is monosemantically understood and reusable.

Thus, designing an application program model reduces to a description of its variables common with its interface in the form of ontology.

## The Correspondence between a Domain Concept System and an Application Program Model

There is a correspondence between domain concepts and application program variables. The input data should be transmitted from a user to the application program without information loss, as well as the output data from the application program to the user. In this way, this correspondence between the models of the domain concept system and of the application program has to be defined when an interface model is designed.

## The correspondence between a domain concept system model and a display aspect model of an interface

A specific ontology for the display aspects of an interface is a subset of the domain-independent ontology. It determines what interface elements are used in the concrete user interface, and what properties these elements have. The meaning or information components of these interface elements are domain terms described in the domain ontology. In other words, the domain term system forming the input and output data of the tasks is presented to a user in the user interface in the form of different interface elements such as lists, menus, tree control, edit field, graphical images and so on. Thus, there is a correspondence between the domain ontology and the specific ontology for the display aspects of the interface.

At the same time, different communications can exist which communicate the same information. They form a class of equivalent communications. Different users often need in the presentation of the same information in different forms. And what is more, flexibility is an occurring everywhere requirement to the modern interface. It means a possibility of adjusting the interface to user's requirements and its adaptability, i.e. self-adjusting to the user.

Usually a concrete user interface consists of a subset of repeated interface elements, having different meaningful components. For example, they may be menus of the same type which have domain terms as

their names and elements. In this case, there is no necessity to give a presentation in the interface for every domain term. It is enough to enumerate all the display aspects of the same type used in the interface and the domain term classes corresponding to them. The domain term classes are defined in the metaontology. But if there is no repetition in the interface, it is possible to assign explicitly every domain term its display aspect.

## A dialog scenario model

In previous sections the knowledge branches are presented which are necessary for user interface model development. Besides the declarative model descriptions and the correspondences between them, it is necessary to define the functions for control of the interface by the application program and by the user. To do this, the tool for interface development should have a set of system functions determining user interface behavior. It is necessary to define in a dialog scenario model a correspondence between system functions and their presentation in the interface, and also to define attributes of appropriate system functions.

## A DESIGN PROCESS AND IMPLEMENTATION TOOLS FOR THE INTERFACE MODEL

In general case, the design process and ontology-based tool architecture does not differ from the model-based ones. These tools also include design critics and advisors, automated design tools and so on for helping to interface developers.

The user interface design process begins with developing the ontologies or with editing them, i.e. with forming specific ontologies using general ontologies in the case, if the latter have been presented in the Internet. To do this, the tools should contain an ontology editor. The ontology editor gives a possibility to read an existing (for example, in the Internet) ontology and, using it, to extract a specific one. Since these ontologies may appear in the Internet in the near future, their inner representation format has to be standardized. This standardization gives the editor a possibility to read any ontology. If there has not been yet an ontology ready for use in the Internet, then it is possible to form this necessary ontology by the editor, too.

After forming the ontologies of the domain, of the display aspects of the interface and of the application program, models of all the correspondences between the different ontologies are defined by a linking editor. At last, a model of the whole user interface is formed by it. As well as in the case of the model-based approach, an implementation tool generates the source code either in a programming language or in the form of a certain format file, which can be read by an existing UIMS (User Interface Manager Systems). The interface model can also be interpreted at runtime.

## Conclusion

The ontology-based approach to user interface development presented in this report permits to decrease the cost of the user interface development and maintenance. The first reason of this fact is reusing ontologies and their fragments both newly developed and presented in the Internet. The second one is using a library of system functions. The third one is using minimum linking between a application specific program and an interface. This property also simplifies the maintenance of software systems.

## Bibliography

[1] Puerta, A.R., and Maulsby, D. Management of Interface Design Knowledge with MOBI–D. IUI97:International Conference on Intelligent User Interfaces, Orlando, January 1997, pp.249–252

[2] P. Castells, P. Szekely and E. Salcher. Declarative Models of Presentation. International Conference on Intelligent User Interfaces (IUI'97). Orlando (Florida), 1997, pp. 137-144.

[3] Puerta, A. R. Supporting User–Centred Design of Adaptive User Interfaces Via Interface Models. First Annual Workshop On Real–Time Intelligent User Interfaces For Decision Support And Information Visualization, San–Francisco, January, 1998.

[4] P. Szekely, P. Sukaviriya, P. Castells, J. Muthukumarasamy, E. Salcher Declarative interface models for user interface construction tools: the Mastermind approach.. In Engineering for Humand-Computer

Interaction, L. Bass and C. Unger Eds. Chapman & Hall, 1996 http://www.isi.edu/isd/Mastermind/mastermind-ia.htm

[5] Puerta A. the Mecano project: comprehensive and integrated support for model–based interface development. Computer–aided design of user interfaces, ed. by Jean Vandrdonckt. Pressed Universitaires de Namur, Belgium,1996, pp.19–25.

[6] Lonczewski F., Schreiber The FUSE–system: an Integrated User Interface Design Environment, Proc. CADUI 96, J Vanderdonckt, ed., http://www.info.fundp.ac.be/~jvd/dsvis/cadui96.html

[7] Foley J. History, results, and bibliography of the User Interface Design Environment (UIDE): An Early Model–Based System for user interface design and implementation. Proc. Eurographics Workshop design, specification, verification of interactive systems, F. Patern, ed., 1995, http://www.info.fundp.ac.be/~jvd/dsvis/dsvis94.html

[8] P. Szekely.User Interface Prototyping: Tools and Techniques. 1994 http://www.isi.edu/isd/humanoid-papers.html

[9] da Silva, P.P., Griffiths, T. and Paton, N.W., Generating User Interface Code in a Model-Based User Interface Development Environment, Proc. Advanced Visual Interfaces, V. di Gesu, et al. (eds), ACM Press, 155-160, 2000.

[10] P.Szekely Retrospective and Challenges for Model-Based Interface. 1996 http://citeseer.nj.nec.com/szekely96retrospective.html

[11] Puerta, A.R. Issues in Automatic Generation of User Interfaces in Model-Based Systems. Computer-Aided Design of User Interfaces, ed. by Jean Vanderdonckt. Presses Universitaires de Namur, Namur, Belgium, 1996, pp. 323-325.

[12] N. Guarino, Formal Ontology in Information Systems.Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press.

[13] Kleshchev A.S. & Artemjeva I.L. Domain ontologies and knowledge processing. Techn. Report, Vladivostok: IACP, FEBRAS, 1999. 25 p.

[14] A. Kleshchev, I. Artemjeva A structure of domain ontologies and their mathematical models. In proceeding of the Pacific Asian Conference on Intelligent systems 2001. Korea Intelligent Information Systems Society. 2001. p. 410-420

[15] M. Uschold Knowledge level modeling: concepts and terminology. The knowledge Engineering Review, Vol.13:1.5-29

[16] www.ontoweb.org

[17] Coats R.B. and Vlaeminke I. Man-computer interfaces. Blackwell Scientific Publications 1987

[18] http;//interface.es.dvo.ru/ontology.htm

## Author information

Kleshchev Alexander, Professor, Head of the Expert System Department, Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of the Sciences: Vladivostok, +7 4323 310424 kleschev@iacp.dvo.ru, http://www.iacp.dvo.ru/es

Gribova Valeriya, Ph.D. Senior Researcher of the Expert System Department, Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of the Sciences: Vladivostok, +7 4323 314001 gribova@iacp.dvo.ru, http://www.iacp.dvo.ru/es

# MUTUAL ADAPTATION OF THE COMPUTER ENVIRONMENT AND INDIVIDUAL

*V. Kolomeyko*

*Abstract: In article the problems of mutual adapting of the humans and computer environment are reviewed. Features of image-intuitive and physical-mathematical modes of perception and thinking are investigated. The problems of choice of means and methods of the differential education the computerized society are considered.*

*Keywords: image-intuitive modes of perception, physical-mathematical modes of perception, differential education.*

## Introduction

At present the human society and computerized environment are going through the complicated process of mutual adaptation. The new approaches to the solution of the tasks and problems, making scientific researches, studying, many more. Some of the processes are completely new for the humanity that is why they demand undivided attention of the researchers and society on the whole. It is important not only to study tendencies and to analyze possible consequences but to work out the means and methods of purposeful control of them. The purpose of this work is to single out some key-problems, to set possible ways of their solution.

## Image-Intuitive and Physical-Mathematical Modes of Perception and Thought

The existed modes of studying and description of objects, processes and phenomena may be conditionally united into two groups, named further as physical-mathematical and image-intuitive modes of perception and thinking. The basis of physical-mathematical modes is formed by traditional for physics and mathematics approaches and methods, including formalization, Laplacian determinism [1]. The basis of image-intuitive mode is orientation to the image thought, intuition, subconscious, collegial and other informal methods of making decisions [2].

It's natural that such division is rather schematic without clearly marked borders. It is easy to explain because very often in practice a kind of symbiosis of these two approaches and the attempts of strict division may seem to be controversial, artificial, incorrect.

For humanity image-intuitive mode is more usual, because it has appeared incomparably earlier historically. So there is no wonder that small children in their attempt to know the word use image-intuitive mode only. Even their abstractions and generalizations are based on the combination of imagines and inherent intuitive methods of perception. But the traditional views towards science and education prefer physical-mathematical mode. Such prevalence very often reveal itself in the primary school where to children with imaginary, intuitive thought physical-mathematical mode of cognition of the world is instilled. But only for minority it is useful.

Rather typical test, connected with description of the leaf's flight from the tree. The child of pre-school age describes such flight rather adequately. But as for the senior students very often they try to do this with the help of physic's textbook not taking into the consideration the striking difference between experimental and theoretical results.

It is well known [3] that every individual has his particular features of thought, perception, cognition of the world. For one people the complicated mathematical formulas bring the feeling of harmony and the understanding of the heart of the matter. For another they are not more than dry notes which are hard to understand, and the real harmony and understanding of heart of things is contained in quite another (the classical example — Mozart and Saliery). Wrong selected way of teaching contradicts with the innate world of the man, with his subconscious, mode of thought, becomes the obstacle for the development of the individual and society in whole.

The methods of differential education, taking into consideration the abilities, inclinations, the character of pupil were developed and used by different pedagogues [3, 4]. Not going into details of these methods, we should admit that they are oriented into pre-computer world reality. That is why they demand reconsideration and further development.

## The Formation Individual's Personality in His Intensive Coordination with Computer Environment

One of the most important factors making the influence on the forming of the modes of perception, thought and cognition of the world is global computerizing. A human being as one of the biological spices has got virtual intellectual assistants, advisors, prompters, persons to talk to, friends, the performers of the everyday work, which take a lot of duties, functions unknown before. The role of influence of these assistants is not always obvious and with only one meaning. But they exist, develop quickly enter different areas of human activity. So it is time to take into consideration that not simply an individual with his abilities, priorities, modes of perception and making decisions lives, studies, works, but the individual who has the virtual assistant.

More over, the development of the computer environment has reached such stage when the virtual assistant performs a lot of tasks and functions more effectively than the man. That is why to our mind the teaching of one or another specialist is advisable to consider as a process of creating and teaching the one human-computer team, in which a man adapts to his virtual assistant (computer), and computer adjusts to the one particular man. As the tasks, possibilities, needs of the man change during his whole life so the process of mutual adaptation of each of the participators from every man-computer team may last during whole life.

This problem is complicated, closely connected with the help of return contact with a lot of other different tasks and problems. That is why we should solve this problem from different sides. Because psychology, pedagogy, computer sciences, and fundamental preparation and plus a lot of professional human cultural knowledge abilities and skills are important here.

But we should start with the analysis of the influence of those principally new factors which computer brings into the creative, innate world of the man. To our mind the most important are as follows:

1. The high level of lability, disposition, adaptation of the computer according to the tastes, needs, and demands of different users.
2. A friendly programming environment which allows even unprepared users to establish effective cooperation with computer.
3. The elements of game which exist in the associating with the computer. Clearly expressed play origin arıses the level of motıvatıon of the users even in process of makıng everyday and burdensome work.
4. The possibilities of practical use of the most difficult methods of mathematics other sciences and subjects by non-specialists.
5. Completely new possibilities of communication and multimedia.
6. The presence of the means and methods of fixation and notation of events, actions, decisions making possible to simplify the following analysis and adaptation of the system.
7. The possibility to cooperate not with the real objects and phenomena but with their virtual copies and worlds [5].

This and other factors connected with the rapid development of the computer environment make great influence as well as on separate individuals and on the whole society. Some of these factors are considered lower.

## The Development of Means of the Differentiatial Education in the Computerized Society

Taking into the consideration upper said, we keep to the mind that for every student should be selected individual combinations and proportions of different methods of teaching. The most important factors in such selection must be:

- the particular features of the mind, perception and the character of the individual;
- the particular features of the future specialization and the work of the individual.

This supposes the necessity:

- the differentiation of the profession, the areas of activity, the teaching methods, in dependence from correlation different modes of thought and making decisions;
- effective distribution of functions, tasks, roles among the society, individual and computer as it applies to one or another situation, profession, area of activities.

To put it differently there are a lot of obvious and non-obvious dependencies and connections between the choice of teaching methods and the distribution of functions in the computerized society in general and human-computer systems in particular.

We do not claim to determine what subjects and in what amount should be studied by one or other student. Our purpose is to investigate the mutual influence of the tendencies in the computer environment

development and teaching methods. It is important to understand some things, to investigate general laws based on the simply and impressive examples.

As one of such examples we have chosen the tasks connected with the architect education. The foundation for such selection was follows. For the first, the profession of architect supposes the possession of both imagine-intuitive and physical-mathematical thought. For the second, the profession of architect is in the great demand according to creators of the systems of computer-aided design / computer-aided design manufacturing (CAD/CAM) that let us say about the experience and tendencies. For the third, strongly pronounced architect-builder inclinations reveal itself in a rather early childhood, that gives the reasons to consider this profession and part of the conclusions and recommendations rather broadly.

It is obvious that the education of the architect, who designs the objects with the help of computer is one thing. And the training of the architect who has to create computer programs of the calculation constructions and expenses is quite another thing. The first one based on the image-intuitive way which is inherent to the creative process. That is why he must “feel” mathematics, thermal physics, economics more than to know them (it is supposed that this general technical subjects are based upon corresponding CAD/CAM or the calculation group). The second one must know deeply all basic subjects because without it he can not “apply” them in his programs.

Both of these two architects must professionally use proper CAD/CAM and effectively make every possible constructions and calculations. The analysis of the present tendencies with the high level of confidence make possible prognosis of two things. For the first that CAD/CAM will become the main working instruments of the architects. For the second that for the majority of the architect-students the most effective methods of studying of the courses and the units of the general technical subjects will be admitted the methods, based upon the practical calculations, constructions, projects with the help of the intellectual instruments.

For example the methods of studying the courses oriented toward the first student to our mind must include the possibilities of making the virtual experiments. When a student chooses one or another construction, the results of this experiment are immediately shown: the gradients of thermal fields, voltage waveforms, the grades of heat irradiation. When student changes the construction the appropriate modifications of fields and loads appear on the monitor.

It is quite possible that at the beginning the student will advance to the acceptable decision with the help of the cut-and-try method. But the experience of use of game situation shows that very quickly appear a kind of understanding of the laws based possibly on the feeling of innate harmony. The chaotic selection of the variants is changed into the sensible use of the method of the successive approximations. A man begins to “feel” the problem, studies to solve it effectively.

At the set example with two students we have considered some extreme cases: image-intuitive thought or physical-mathematical. In ordinary life such cases take place not very often. The intermediate variants are more typical.

For example, in many professions connected with the design the best results achieve the specialists who effectively combine image-intuitive and physical-mathematical modes of thought. A lot of experienced specialists have learned on practice how to set themselves into one of another mode of thought, to select important proportions of these modes according to the character of the tasks which are being solved and works which are being made.

As people are very different and solve very different tasks so one of them are at a lack of the image-intuitive thought and other are in need of physical-mathematical thought. This lack could be compensated at the expense of choice of appropriate profession and work. But it is so not always. That is why a special attention is paid to the creation of intellectual assistants, which are able to compensate the lack of knowledge, skills and abilities of the individual.

We should stress that computer is not very reliable from the point of view of increasing the potential of the image-intuitive mode of thought. The main achievements in this field are qualitative visualization, wide multimedia possibilities. The rest is much worse, not talking about the intuition. This is significant argument in favor of abrupt increasing the pedagogues’ attention to the development of the image-intuitive mode of thought. Because the lack of development of physical-mathematical mode of thought computers compensate more successfully.

One of the most effective methods of purposeful development of one or another ways of perception and thought is the practical cooperation with the appropriate developing and teaching computer programs. These programs usually contain a big number of specially selected game situations, which substantially increase the level of user’s motivation and simplify the teaching. There are a lot of such programs and games. The

spectrum of them is rather wide too. Some of them are considered to be strictly specialized other is used for wide application. There is positioning in fields of activities, age, interests. For the last years developing computer games appeared which are ranged as the games for children of 3 to 5 years (we should notice that they are played even by 2 years old children with the help of adults).

The analysis of results of practical use of such games persuade us in advisability of fundamental expansion in this direction. The basing on the game approaches gives powerful stimulus for the development and teaching the child. At the same time the investigations directed on the revealing and use of another positive treats and particular features of child's perception are necessary. Without it the game may bring not only the benefit but harm. Besides this, it is very hard to develop the scenarios of games, create appropriate interesting understandable and at the same time useful play situations, without studying the particular features of perception. In this case it is advisable to our mind to base on such instincts of creation that the majority of people has. The child willingly builds something from the sand, bricks and details of constructor. That is why at the beginning the substitution of bricks, details of constructors by their virtual analogs may be possible. Then goes the rapid increase of the number and types of virtual details. A new effect will take place in that case if the computer begin to form a kind of evaluation figures of constructions created by the child. For example if it is said about the building of virtual house so such figures may be the durability of the house, the value of the flats, their temperature conditions, the isolation of sounds... In another words the things, which the child knows and understands.

Practice shows that when a child plays so-called strategic and emulation computer games he begins to understand very quickly that a lot of things in the world are interconnected. The improvement of one thing leads to the deterioration of another. It leads to multi-criterion optimization by the certain examples. In games, which are oriented towards the children of pre-school age, should not be a lot of criterions (1 - 2). In primary school classes the number of optimization criterions may run up to 3. As regards senior school, a lot of things here depend upon the definite subjects and teaching purposes.

It is obvious that the extension of range of the studied problems leads to increase of optimization criterions. The same example with the construction of buildings supposes the studying and analysis of many factors (including visual, social, ecological). That is why the number of optimization criterions may reach 10 and 20. As for the means and methods of education in the senior classes the author, taking into consideration the limits of article, can repeat the same arguments as in the example with architects. Adding that many things are still not clear, require studying, development and approbation.

Studying the questions of differentiation of education with the help of computers we should point to the possible risks and negative moments. The most obvious those demerits and difficulties which are on the top. For the first these are the factors of negative influence of computer environment upon the children, additional stratification of society, expenditure and difficulties connected with the considerable reorganization of educational process, the individual choice of educational programs, teaching forms, means and methods. Besides the differentiation of education at the primary school classes supposes the necessity of making appropriate decisions about career-guidance because such decisions are always connected with multitudinous risks.

As the humanity has started the path of rapid computerization and there are no other real alternatives so the demerits and possible risks only stress the main thesis of article: the problems of global computerization are complex, require special attention of investigators and society on the whole.

## Bibliography

1.Князева Е.Н. Одиссея научного разума. Синергетическое видение научного прогресса. – М.: ИФРАН, 1995.-228с.
2.Коломейко В.В. Методологические аспекты построения человеко-машинных систем поддержки принятия решений // Искусственный интеллект. – 2002. - № 3. – С.101 - 106.
3.Гипенрейтер Ю.Б., Романова В.Я. Психология индивидуальных различий. – М.: ЧеРо, 2002. – 776с.
4.Дифференциация в начальном звене. / Под ред. Ю.З.Гильбуха. Психолого-педагогические основы дифференциального обучения в общеобразовательной школе. – К.: Перспектива, 1996. – 53 с.
5. Коломейко В.В. Проблемы использования неформальной информации в задачах моделирования // УСиМ. – 2001. - № 6. – С.17-24.

## Author information

Vladimir Kolomeyko - V.M.Glushkov Institute of Cybernetics, Prospect Acad. Glushkova, 40, 03680, Kiev-187, Ukraine. Tel: (380+44) 234 58 51 e-mail: vvklm@ukrpost.net

# AN APPROACH TO NEW ONTOLOGIES DEVELOPMENT: MAIN IDEAS AND SIMULATION RESULTS

*B. Dobrov, N. Loukachevitch, O. Nevzorova*

*Abstract: In the paper we consider the technology of new domain's ontologies development. We discuss main principles of ontology development, automatic methods of terms extraction from the domain texts and types of ontology relations.*

*Keywords: Ontology, thesaurus, automatic term extraction, ontology relations.*

## Introduction

The present article is dedicated to the technology of creation of the so-called linguistic ontologies, i.e. ontologies, concepts in which are generally based on the semantics of the domain terms. Such ontologies are usually used for the automatic processing of texts in a natural language.

The technology has been developed in the process of creation of large and extra large ontologies and thesauri for various domains and their actual usage in multiple applications of automatic text processing.

Among such works are the following:

thesaurus of the social and political life (28 thousand concepts, 67 thousand terms, 100 thousand conceptual relations), - Sociopolitical thesaurus [Loukachevitch, 2002], which is a search means in the University information system RUSSIA (www.cir.ru) and is used in such applications of automatic text processing as conceptual indexing, automatic text categorization, automatic text summarization;

the Russian language thesaurus RuThes (43 thousand concepts, 100 thousand words and expressions, 166 thousand conceptual relations) [Loukachevitch N., 2002];

thesaurus for the domain «Elections» - is included into the Sociopolitical thesaurus;

thesaurus for the domain «Economic statistics»;

ontology for the domain «Software functionality» for decision making support during software testing;

Avia-Ontology for the domain, describing behavior of an operator (air crew) and board equipment in various flight operations (1200 concepts, 3400 terms). Aim of the Avia-Ontology development is the analysis of the completeness of documents, describing the logic of work in typical flight regimes [Nevzorova, 2001].

Avia-Ontology is currently being developed and will be used as a basic source of examples for the article [Dobrov, 2002].

It should be highlighted that the peculiarity of the proposed technology and the existing experience is namely the activity of a knowledge engineer, who at the beginning of work has a very superficial idea of the conceptual structure of a domain and its terminology.

## 1. The formation of text collection

One of the essential conditions of the successful development of a linguistic ontology is preliminary creation of an electronic text collection with reference to a domain. The collection may be of various genres and may include textbooks, scientific articles, technical data, mass media works and so on.

During the development of the Sociopolitical thesaurus we used the text collection of the University Information system RUSSIA, comprised of over 700 thousand documents: official documents, laws, scientific works on social studies, newspaper releases.

During the development of the Avia-Ontology a great effort had to be made in order to form a sufficient collection of electronic documents on the given domain. To a great degree scanning of the printed matter and search of relevant materials in Internet had to be done. As a result an electronic collection with the size over 100Mb was formed.

## 2. Automatic terms extraction from the domain texts

### 2.1. Terms extraction based on syntactic information

Having formed an electronic collection of domain texts it is necessary to obtain its "terminological portrait", for which the procedures of automatic extraction of terms, essential for the given domain, are employed. For the Russian language terminology the syntactic structure of over 90 percent of domain terms [Loukachevitch, 2002] refers to one of the following constructions:

single nouns, adjectives, and words not described in the morphological dictionary – usually abbreviations;

noun phrases (NP): noun + noun in the genitive case;

NP: adjective + noun

NP: adjective + adjective + noun

NP: noun + adjective + noun in the genitive case

Such types of constructions are collected on the basis of preliminary morphological processing of texts. In this process noun and adjective agreement was checked, after which the bulk of such noun phrases are syntactically correct groups. Decreasing frequency-ordered lists of such words and word expressions present important data for the formation of an idea about the domain.

It should be mentioned that the list includes many non-terminological linguistic expressions.

First of all, domain texts contain a great number of words of general meaning, e.g. *possibility, means, condition, type, point* and others.

The larger a domain and a text collection are, the bigger problem is to exclude multiword constructions of general meaning, such as *task solution, further development, uniform system, beginning of the year, present time*. During the development of Sociopolitical thesaurus we made use of special lexical filters, organized as a specialized vocabulary to exclude such general expressions from consideration [Loukachevitch, 2002]. Such filters are undoubtedly useful for the creation of extra large ontologies, however, they also depend on a domain. So, for instance, the word *argument* may be considered as a non-terminological one in the sociopolitical domain, however, *argument* is an important term in the domain of software production.

Besides, in a domain there may be longer terms or terms of a different syntactic construction, like those containing prepositions.

It should however be highlighted that the simplicity of the given algorithm of term extraction is an important factor of its use for the analysis of the domain structure.

### 2.2. Multiword terms extraction on the basis of the text structure

In order to extract longer terms and/or terms containing prepositions another method is used.

Many algorithms of multiword term extraction use the supposition that words, comprising a term, frequently occur together. [4, 5]. To find terms of a more complex syntactic structure that described earlier, we use our own variant of algorithms of such type.

If the text author use a certain term as a separate unit of narration, then exactly in this text words of the term will occur alongside more frequently than spaced out.

To reveal this during the text processing for every word (noun, adjective) an immediate neighbor word and neighbor-words in the text window of a given size are stored. A table of immediate neighbor words and a neighbor table in the text window are created and the frequency of word pairs occurrence is calculated.

Further it is expected that if a pair of words occurs as immediate neighbors in over half of cases of their occurrence in the same text window, this proves that this pair in aggregate serves as a reference point in the text, i.e. represents a term or a fragment of a term.

In this case the word pair is glued together to form a common terminological unit and the tables are recalculated as if this unit has been known from the very beginning before the text processing. This step gives the possibility to the further development of the term, thus forming units of length 3 and longer. Examples of terms obtained in this way in the aviation domain are as follows: *break off the attack*, *position of tactical advantage, flying in pairs* and so on.

In the domain “Elections” such terms as *member of election committee with right to deliberative vote, executive body of local government, local governor’s elections* were obtained.

When being reviewed, the terms obtained are ordered not by the frequency of their occurrence, but by the number of texts in which they occurred. It is supposed that a special attention should be paid to those word-combinations which remained stable in over two texts. As a result of the experiment on the text collection of 50Mb newspaper releases 1346 such word-combinations were obtained. 80% of them were qualified as terms. The comparison was drawn with the Sociopolitical thesaurus and there were singled out approximately 30 important terms, not present in it at that moment, such as *volume of output, economy in transition, capital flow.*

The algorithm drawback is that it practically does not extract terms from short texts - a term has to be used in a text at least twice, better three times. However this drawback is insignificant when dealing with extra large collections, such as the collection of UIS RUSSIA. Thus we suggest to process the whole collection using the abovementioned algorithm to single out terms not present in the Sociopolitical thesaurus and to further monitor the appearance of new terms in the sociopolitical domain.

We consider useful to apply both described algorithms to work with small domains.

## 3. Using general linguistic resource as a base for the development of applied ontology

The terminology of any domain contains both specific terms, used only in the given domain or in the range of similar domains, and rather commonly known terms. In the given domain the examples of such commonly known terms are *pilot, airplane, pursuit plane, weapon, attack* and many others. This allows us not to begin the domain model development from point zero, but to use the knowledge described in more general linguistic and terminological resources.

As such a source we used the Thesaurus of the Russian language RuThes [Dobrov, 2002]. The Thesaurus represents a hierarchical network of concepts, each of which has a number of text variants (linguistic expression means) and conceptual relations with other concepts of the thesaurus. The resource volume nowadays makes up 100 thousand words and word expressions, confined to 43 thousand concepts. Over 166 thousand relations were manually determined between the concepts. Over 1200000 relations between concepts of the Thesaurus were established on the basis of the transitivity and inheritance properties.

The RuThes Thesaurus contains two big parts. RuThes is comprised of the Sociopolitical thesaurus, including terminology of economic, political, military, social, scientific and other spheres (64 thousand words and terms). The zone of words and word expressions of the thesaurus, designating actions, situations, objects, which can occur in any subject area texts, and thus not included in the Sociopolitical thesaurus, is called general lexicon (33 thousand words and word expressions).

The presence of a big generally valid linguistic resource makes possible to compare the given resource with the domain texts, to single out the knowledge types described in the thesaurus (concepts, synonyms, relations between concepts), to transfer them to a special working domain as a basis for the creation of a domain ontology. Lists of gathered words and word-combinations in the domain were compared to the terms of the Sociopolitical thesaurus. If the comparison of the next word-combination was successful, a corresponding concept from the thesaurus together with all the terms that express it in the text, was copied into the domain model. During next step all the relations of the Sociopolitical thesaurus between the copied concepts were copied.

Besides, the closing of relations was performed: if concept B is superior compared to concept A, concept C is superior to concept B, while concepts A and C were copied into a domain, then concept B is also copied into the domain model together with its terms variants and relevant relations.

The comparison with the general lexicon zone was controlled manually, as part of such words preserved its general meaning, for instance *necessity, conditions*, etc. and , thus, do not need to be reflected in the domain model, other words like *chandelle, cover, escort* are important concepts of the domain, which therefore have to be reflected in the model. In the second case the corresponding concepts were also copied together with their terms and relations, just as during the comparison with the Sociopolitical thesaurus terms.

Undoubtedly, the transferred concepts and relations require a thorough additional test for adjusting the setup to a certain domain. For instance, as a result of transfer among the terms we can come across a synonymic variant, which is improbable in a given domain, for example, the word «rotorcraft» as a synonym of the word «helicopter» is unlikely to be found in the professional technical area.

In the present version of the Avia-Ontology about one-third (1100) terms were transferred from the thesaurus RuThes together with synonymic relations and relations between concepts, which provided fast development of the Avia-Ontology and considerable reduction of the new resource development time.

## 4. The main procedure of the ontology construction

Having exhausted the available linguistic resources, we begin analyzing the available texts and terms lists for the further ontology development.

The decision making in this process consists of the following steps:

on the basis of the available text material we look for a word or a word-combination, designating an important concept (how to determine the concept «importance», we will consider in the following sections).

Entering a new concept we must provide it with an understandable and unambiguous name. If possible it is better to enter the concept names which are unambiguous even outside the current domain. For instance, when entering the concept "airplane wing" it is possible to name it "*wing"* – in the given domain this word is unambiguous, but it is better to give it a clearer name, which will not lose its unambiguity with the domain expansion or after including this ontology into a larger ontology;

when entering a concept at least one relation of this concept with the other ontology concepts is entered. On the one hand, it is supposed that if it is rather complicated to set a relation for the entered concept, then it is too early to enter such a concept and additional analysis is required. On the other hand there is no need (and as a rule it is impossible) to describe all the necessary relations of a new concept at once. Practice proves that by entering further new concepts the initial position may become clearer – in order to enter something new, it is often necessary to correct the inaccuracies and distortions of the old. Properly speaking, inclusion of new concepts helps to reveal the problems of the existing description;

- and, at last, a concept must be supplied with a list of words and multiword expressions, which can be used to refer to the entered concept in texts. As such text entries single words (nouns, adjectives, verbs), noun and verb phrases can be included. We suppose that a multiword linguistic expression must be used in a text as an inseparable construction. A text entry may be ambiguous (may have another meaning), then it must be specially marked. Besides, a sequence of normalized forms of all constituents of a multiword expression must be entered (masculine gender, nominative case, singular), which will be used for term recognition in texts.

## 5. Ontology concepts selection

### 5.1 Single word-based concepts

The single words, occurring in domain texts, have two subgroups, on which it is easy to decide whether to include or not include the corresponding concepts into an ontology.

One of such groups - evident terms of a given domain, for example, *aircraft, flaps,* and corresponding concepts must be included in an ontology. Another group – evident words of general meaning, such as *necessity, possibility, creation, etc.,* for which inclusion into an ontology is not necessary.

It is more problematic to make a decision on the other groups.

One of such groups - terms of a given domain that appear on the basis of general lexicon words like *turn, reversal, dive*. Narrowing or other changes of basic meanings are characteristics of such words. So, in the aviation domain *turns* refer to the airplanes. Besides, to help in distinguishing such terms serves the fact that in the terminological word-combinations lists there is a considerable amount of different word-combinations with the inclusion of this word, which should also enter the ontology: *turn, combat turn, turn radius.*

Another «complex group» of words are words that clearly refer to the vocabulary of a general meaning, but ontology includes a certain number of concepts, based on the terms with this word: break off – *break off combat, break off attack*. The questions arise if corresponding generalizing concepts should enter an ontology. Two aspects should be considered. On the one hand, the appearance of such a generalizing

concept provides an additional structure to the ontology, which is a positive factor. However, on the other hand, if a word is very ambiguous in the framework of a given domain, and abstract, then it may cause serious problems in the lexical ambiguity resolution, and this corresponding concept should not enter the ontology.
The last group of words, which requires a special effort to decide if a corresponding concept should be included into an ontology or not, are words, that are on the borderline between domains or those that were relatively accidentally used in a text collection, for instance, *aircraft construction* (whether refers to the subject area or not), *adapter, heat insulation* and so on.

## 5.2. Multiword word-combinations-based concepts

Any domain text contains a great number of various multiword linguistic expressions. Selection of such expressions for the inclusion into an ontology is a serious problem. Existing terminological lists of a domain usually embrace only a small part of those term-like expressions that are met in texts. Experts may not also have a certain opinion on the bulk of such expressions. Therefore it is necessary to have a total of principles to decide which specific factors are to be taken into account for including concepts based on multiword expressions into an ontology.
We should highlight the main principle at this point. If such a concept is included into an ontology, it should happen not so much because the corresponding word-combinations refer or not to the vague category of the given domain terms, as which new information the appearance of this concept in the ontology gives. Thus, a new concept in an ontology is the application point of additional information which can be used in automatic text processing.
Such information may be divided into several types.

### 5.2.1. Existing and important

Any domain has a small number of main points which are extremely important in the given domain. Terms and other linguistic expressions that correspond them are highly frequent in the subject area texts. Such main points (concepts, single objects) must be reflected in the ontology. So, working in the domain «Elections» it is essential to have in the ontology a conceptual unit *CENTRAL ELECTION COMMITTEE OF THE RUSSIAN FEDERATION.*
If concepts entered into the ontology have a fixed and small number of narrower concepts, then they have to be reflected in the ontology. So, for the domain «Elections» types of elections are reflected, in the Sociopolitical thesaurus – types of budget, in the subject area of military aviation - bombing flight regimes *(dive bombing*).
Another important type of information is that two concepts have a common subtype. For instance, concepts *DEFENSIVE MANEUVER* and COMBAT TURN have a subtype *DEFENSIVE TURN*, concepts *FINE* (*PENALTY)* and *ADMINISTRRATIVE PUNISHMENT*– subtype *ADMINISTRATIVE FINE.*

### 5.2.2. Multiword expression has «interesting» synonyms

A concept can unite the total of various text expressions with the same meaning (derivatives (*to take off, take off*) are also considered as text entries of the same concept). So, the revealing of synonymic expressions or derivatives often leads to the introduction of a new concept for the fixation of the synonymy found. The variety of textual expressions in this case often points at the importance of a corresponding concept.
After the concept is set up, special effort is applied to find other ways of referring to the same concept in texts, i.e. the synonymic raw of the concept is maximally filled. These variants may seem obvious for a person, and their entry may seem tiresome, but as practice proves, during automatic processing of various texts direct matching is better that any logical inference. It is often supposed that this or that variant exists, and then its actual occurrence is checked by Internet. For instance, if a new concept is entered on the basis of the term *horizontal flight bombing,* then instantly existence of the synonym *horizontal bombing,* which actually exists, is checked. If afterwards a concept for the term *pitch-up bombing* is entered, then, naturally, the existence of the term *pitching bombing* is checked; such term was not found in Internet. Let us give example of synonymic row:

ENGINE POWER INCREASE
Increase engine power
*Increase engine thrust*
*Increase thrust*
*Engine acceleration*

### 5.2.3. Relations that do not follow from the structure of a multiword expression

The principle used to evaluate the necessity of entering a concept into many thesauri and ontologies is that a multiword term has relations that do not follow from its structure.
Examples of such relations are:
*accelerated turn – loss of speed,*
*attack evasion – defensive maneuver,*
*superiority in energy – tactically advantageous position.*
To fix this relation it is necessary to introduce the corresponding concepts.

### 5.2.4. Completion of ontology levels

An important principle of ontology completion is the "closing" principle, which has two subtypes.
In the first place, if a new concept, introduced by any reason has created a new inferior ontology level, then it has to be completed by other essential concepts of the same level. For instance, if *MISSILE LAUNCH* concept is entered as an inferior one for the concept *USE OF WEAPON,* then it is necessary to enter, for example, the concept *CANNON FIRING*, as the second most important type of using weapons in the given domain.
This principle is at the same time limiting: if we decide to enter a new-level concept, we must evaluate the consequences of such a step: how many concepts of the same level we are going to enter; if the number of potential concepts of this level is too big, then the entry limiting principles should be determined at once. So, for instance, in the *Sociopolitical thesaurus* there can be a lot of concepts inferior to the concept *GOODS FOR CHILDREN*. The appearance in the inferior row of concepts *CLOTHES FOR CHILDREN, SHOES FOR CHILDREN, TOYS FOR CHILDREN* is additionally justified by the existence and inclusion into the thesaurus of certain types of these goods, having single lexemes as text entries.
On the other hand, an opposite situation may emerge: several concepts sharing common features are found, it is necessary to find a common concept. For instance, on the basis of subject area texts analysis concepts *FLAPS* and *SLATS* are introduced, common features of which is that both are located on the wings and that they serve to control a flight. Extra attention is paid to searching for a generalization, which is found - *FLIGHT CONTROL SURFACES.*
Another example of generalizing two concepts that were entered: *AFTERBURNER LIGHTING* and *DECREASE THRUST* lead to the entry of the concept *THRUST CONTROL*.

### 5.2.5 Single words are ambiguous and word-combination is unambiguous

An important factor which helps to determine the entry of a new concept is the presence of ambiguous words inside an unambiguous multiword expression.
So, an ambiguous term *press* is important for the sociopolitical area, and to support the ambiguity resolution process we introduced such concepts as *LOCAL PRESS, CENTRAL PRESS, ILLEGAL PRESS* into the Sociopolitical thesaurus.
Working in the same broad sociopolitical area we may hesitate if the introduction of concepts *SHORT FILM–FEATURE FILM* is necessary, but as soon as it becomes known that *SHORT FILM* has the synonym *short subject*, the corresponding concepts are included at once.
We should emphasize that all the abovementioned does not mean that for every word-combination, consisting of ambiguous words, a corresponding concept is created; the described principle only helps to make the decision in such cases when we are almost ready to introduce a concept.

## 6. Ontology relations

The most important part of the prevailing number of ontologies is the total of relations between the concepts. This set of relations largely depends on the domain and on the task for solving which ontology is meant. We suggest to begin the construction of an ontology on a minimal set of relations and to determine the domain structure according to this set. Such a minimal set of relations does not depend on the type of a domain, on the type of a problem solved, as it is based on the fundamental properties of concepts. In the first place for a given concept we determine such concepts on which depend its existence or existence of the given concept examples, i.e. determining the so-called relations of ontological dependence, which are studied in the framework of the philosophical discipline «formal ontology»

The main instruments of entities analysis within Formal Ontology [Smith, 1998] are the following:

the theory of identity, integrity. The main problems of this type of analysis: what does the fact that two entitles are one and the same thing means, how can an entity change and preserve its identity, what properties are essential for preserving one's identity, etc.

the theory of part and the whole (mereology, mereotopology). The main problems here are the following: what is considered as a whole, what makes an entity a whole, what is the connection of parts in the whole, what properties such a connection relation has, how is the whole separated from the «background», what are boundaries and so on.

the dependence theory [Guarino, 1998].

The main question of the dependence theory is if an entity can exist by itself or it supposes the existence of something else:

whether the existence of an entity supposes the existence of something else (rigid dependence), for instance, *boiling* is impossible without the existence of a certain volume of liquid which boils;

whether existence of examples of a certain class (generic dependence) is supposed, like, the appearance of the concept *garage* is impossible without the existing concept *motor vehicle,* though a certain garage may appear without any reference to a certain motor vehicle;

- when the existence of an entity in moment T presumes the existence of another entity in moment T1 before T (historical dependence), so, for instance, straw historically depends on threshing, as straw can not appear without a preliminary threshing process, altogether this work comes to an end, while straw continues its existence for a long time.

Examples of conceptual dependence relations in the Avia-Ontology are as follows:

*ALTIMETER* depends on *FLIGHT ALTITUDE* (generic dependence),

*TANKER AIRCRAFT* depends on AIRCRAFT FUEL (generic dependence),

*AIR PATROL* depends on *FIGHTER AIRCRAFT (*rigid dependence*)*

Thus, for each ontology we suggest to develop a sort of an "initial" ontology, in which non-taxonomic relations are relations of the conceptual dependence. Such ontology can serve as a basis for explication of the domain structure and determination of a new set of relations, necessary for solving the main problem. In this case the conceptual dependence relations are so important for any domain, that there is no need to delete them, it is only necessary to re-name them in the newly introduced relations system.

A specific set of relations, which is used by us now besides taxonomic relations (BROADER-NARROWER relations), is the following:

PART– WHOLE – is used to describe the traditional parts, participants of situations, properties. Here the conceptual dependence of concept-part on the concept-whole is required;

unsymmetrical associations ASC1-ASC2 – are used for the rest of conceptual dependence relations;

symmetric association is used for concepts, similar by meaning.

Thus, two types of relations in the relations set employed by us are significantly bound with the concept of ontological dependence. Relations of these types occupy approximately half of all relations in our thesauri and ontologies.

## Conclusion

The described technology of constructing ontologies for various domains was employed to create the so-called linguistic ontologies, which are used to solve problems of the automatic text processing. However, application of such technologies, connected with the processing of large text collections is also useful for the creation of ontologies in those domains, which are not directly connected with text processing.

The carried out analysis of the electronic text collection ensures:

completeness of concepts covering in reference to the collected corpus;

objectivity of concepts and terms interpretation, as various texts from the collection are analyzed.

"Minimal" relations set

makes possible to begin the ontology construction at once, as soon as a task is set and a domain is determined;

provides a conceptual basis for communicating with experts in the given domain;

provides the initial domain structuring which may be used as a basis for singling out special relations in the domain.

## Acknowledgments

The work has been performed thanks to the support of the Russian Fund of Basic Research, grant № 02-07-90279.

## Bibliography

[Loukachevitch, 2002] Loukachevitch Natalia V., Dobrov Boris V. Evaluation of Thesaurus on Sociopolitical Life as Information Retrieval Tool // Proceedings of Third International Conference on Language Resources and Evaluation (LREC2002) / M.Gonzalez Rodriguez, C. Paz Suarez Araujo (Eds.) – Vol.1 – 2002, Gran Canaria, Spain – p.115-121.

[Loukachevitch N., 2002] Loukachevitch N.V., Dobrov B.V. Thesaurus of the Russian language for automatic processing of large text collections// Computer linguistics and intellectual technologies: Works of the International seminar Dialogue'2002 / edited by. a.S.Narinyani – м.: Nauka – 2002. – Vol.2 - p.338-346. In Russian.

[Nevzorova, 2001] Nevzorova O.A., Fedunov B E. System of analysis of technical texts "Lota": main conceptions and project decisions. // RAS publishing house. Theory and management systems – 2001. – № 3.– pp. 138-149. In Russian.

[Dobrov, 2002] Dobrov B.V., Loukachevitch N.V., Nevzorova O.A. Computer-aided construction of the applied ontology: technological aspects // International conference IEEE Artificial intelligence systems (IEEE AIS'02) Gelendzhik-Divnomorskoe, 5-10 September 2002 года – Text processing and cognitive technologies: Collection (ed. 7) / edited by V.D.Solovyev – Kazan: Otechestvo 2002. – pp.103-109. In Russian.

[Loukachevitch, 1996] Loukachevitch N.V. Computer-aided formation of information searching thesaurus of social and political life in Russia// NTI. ser.2. - 1995. - N 3. - pp.21-24. In Russian.

[Smith, 1998]Smith B. Basic tools of formal ontology. In "Formal Ontology in Information Systems", N. Guarino, ed.

[Guarino, 1998] Guarino N. Some Ontological Principles for Designing Upper Level Lexical Resources. In "Proceedings of First International Conference on Language Resources and Evaluation".

## Author information

Boris Dobrov – Research Computing Centre Moscow State University, Russia, Moscow, Vorobjovy Gory; e-mail: dobroff@mail.cir.ru

Natalia Loukachevitch – Research Computing Centre Moscow State University, Russia, Moscow, Vorobjovy Gory; e-mail: louk@mail.cir.ru

Olga Nevzorova – Chebotarev Institute of Mathematics and Mechanics, Russia, Kazan, ul. Kremlevskaja, 18; e-mail: Olga.Nevzorova@ksu.ru.

# AN ALGORITHM FOR OPTIMAL BIPARTITE PLA FOLDING

*Liudmila Cheremisinova*

*Abstract. This paper presents some results of PLA area optimizing by means of its column and row folding. A more restricted type of PLA simple folding is considered. It is introduced by Egan and Liu and called as bipartite folding. An efficient approach is presented which allows finding an optimal bipartite folding without exhaustive computational efforts.*

*Key Words: Programmable Logic Array, area optimization, PLA folding, bipartite folding*

## Introduction

Programmable Logic Array (PLA) is widespread used hardware form for the structured design of digital VLSI systems due to the regularity of its structure. PLA layout design is easily automated because of its direct correspondence with PLA personality matrix. The price paid for the structural regularity is that much PLA area is unused because a large percentage of the row-column intersections are not personalized. Several techniques have been proposed for reducing the area required.

Two approaches are usually used to reduce the area occupied by the PLA: logic minimization that provides logic expressions with minimal number of products and topological minimization reclaiming unused space. The proposed paper deals with the problem of topological optimizing PLA area by means of its folding [1 – 5]. PLA folding allows reducing needed area without loss of regular structure of the PLA. There exist different types of PLA folding. They are based on merging several columns (and/or rows) of a PLA into a single column (row). The paper focuses on simple column (and/or row) folding that involves merging pairs of columns (and/or rows) into single columns (rows).

Folding a column of a PLA supposes to split that column into two segments so that two inputs or outputs may share the same column of the folded PLA. This means vertical lines are broken into an upper and a lower parts and two variables (pair of inputs or pair of outputs) don't need two lines but only two segments of the only line of the PLA. The task is to find such a permutation of PLA rows, which allows a maximum set of column pairs to be implemented on segments of single lines of the folded PLA.

In [3, 4, 5] a special type of simple column and row PLA folding is considered, in which all of the breaks of the columns (or rows) occur at the same level (Fig. 1, 2). Such a case is referred to as bipartite folding. The single break level of bipartite folding allows speaking of an upper and lower folding regions, which contain the segments of those folded columns that are correspondingly above and below the breaks.

While bipartite folding may theoretically be only 25 percent as effective as regular simple folding for PLA's with column type constraints, this class of folding approaches the effectiveness of column simple folding for sparse PLA's. Some justifications for this approach are offered [3, 4], the most important of them are 1) the folded columns (rows) entering from the top (left) of PLA can be ordered independently of the folded columns entering from the bottom (right) of the PLA, that simplifies the routing signals; 2) the same algorithm can be applied for the row folding a previously column folded PLA, that simplifies subsequent PLA row folding; 3) a bipartite folded PLA allows to use much less additional area required for inclusion of testability features. The bipartite folding can be used to partition a large PLA into smaller PLA's, i.e. bipartite folding can be considered as a special type of PLA decomposition too.

In this paper a new bipartite PLA folding technique is presented. It is based on transformations of a Boolean column disjoint matrix that specifies the relation to be disjoint on the column set. That allows the PLA bipartite folding problem to be treated as a maximum unit minor problem. Before searching for a desired maximal unit minor some procedures of the column disjoint matrix reduction are made that allow pruning some rows and columns. Some of suggested results could reduce the search space for algorithm from [4], the other allow yielding optimal solutions.

## Definitions

The overall combinational PLA is a standard two level NOR-NOR structure. Vertical lines of a standard combinational PLA are assigned with the input variables (and their complements) and output variables. Inputs

run vertically through the first part of a PLA matrix, called as the PLA AND plane. It generates signals on its rows, which are used as inputs to the second part of a PLA matrix called as the PLA OR plane.

An example of the PLA AND plane (that is the example PLA from [4]) is shown in Fig.1. In this figure, columns are associated with complemented and uncomplemented inputs. Each horizontal line of the PLA carries a product term. A dot means placing a transistor on crosspoint of vertical and horizontal lines. This PLA AND plane will be used further throughout the paper. Here the OR plane is not shown but either AND plane or both AND and OR planes together can be described in symbolic form by a Boolean matrix. And the only difference is that inputs can share the columns with inputs only and outputs share the columns with outputs.

The area of a PLA is proportional to the total number of its columns times the number of rows. The area occupied by the PLA AND plane in Fig.1 is 273.

Before we formulate a mathematically tractable definition of PLA folding problem and its solution we have to give some definitions.

Each PLA column $c_i$ implies the set $R(c_i)$ of rows, which are populated on it. Any two columns $c_i$ and $c_j$ are *disjoint* if $R(c_i) \cap R(c_j) = \varnothing$. A pair of disjoint columns is defined to be a column *folding pair*, and only those two columns of a PLA can be folded together.

For example PLA, columns $c_1$ and $c_6$ are disjoint, since $R(c_1) \cap R(c_6) = \{r_{12}, r_{21}\} \cap \{r_{14}, r_{16}, r_{17}\} = \varnothing$. But $R(c_5) \cap R(c_6) = \{r_{14}, r_{17}\}$, hence columns $c_5$ and $c_6$ do not generate folding pair.

The square Boolean matrix that depicts when the PLA columns are disjoint will be called a *column disjoint matrix D*. This matrix has as many rows and columns as the number of the PLA columns. The element $d^j_i \in D$ is 1 if columns $c_i$ and $c_j$ are disjoint, otherwise $d^j_i = 0$.

The following column disjoint matrix corresponds to the example PLA AND plane of Fig. 1:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 5 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 6 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 8 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 9 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 12 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| weight | 7 | 5 | 1 | 8 | 8 | 8 | 10 | 9 | 7 | 7 | 5 | 7 | 0 |

The relation to be disjoint on the column set is symmetric and irreflexive, so the column disjoint matrix *D* is symmetric too and has all 0's on the leading diagonal.

Bipartite folding is a column or row folding in which all of the breaks of the columns (or rows) occur at the same level (Fig. 2). The single break level of bipartite folding allows speaking of an upper and lower folding regions, which contain the segments of those folded columns that are correspondingly above and below the breaks.

In bipartite folding all column breaks used to facilitate folding are made between the same two rows. Thus all columns, which are folded and placed at the top of the PLA must be disjoint from all columns folded and placed to the bottom of the PLA. Let the PLA columns of the set $C^u$ belong to the upper folding region and the columns of the set $C^l$ belong to the lower one.

Thus the *necessary and sufficient condition*, the pair $C^u$, $C^l$ of the column sets involves bipartite folding, is the columns of $C^u$ to be disjoint from each column of $C^l$ (and symmetrically vice versa). So, the rows of the column disjoint matrix *D* corresponding to the PLA columns of the set $C^l$ have 1's in all columns of the set $C^u$.

Such a pair of equinumerous sets is called a *bipartite folding pair of sets*. The cardinality of these sets defines the size of the bipartite folding pair of sets and the size of induced PLA bipartite folding.

It is clear that a pair $C^u$, $C^l$ of column sets contains all information needed to fold a PLA i.e. it specifies the pairs of columns to be folded and their relative position (top or bottom).

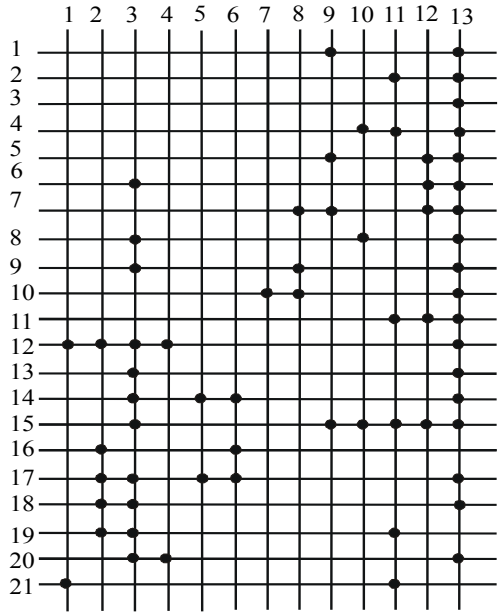



Fig.1. An example PLA

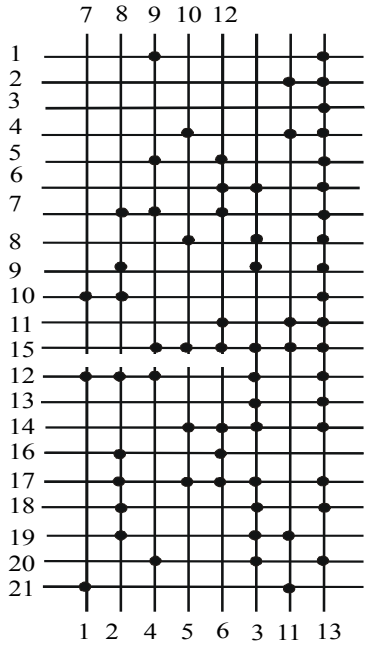



Fig.2. A bipartite folded example PLA

## Transforming bipartite PLA folding into unit minor identification

The column disjoint submatrix, their columns and rows correspond to the PLA columns of the set $C^u$ and the rows correspond to the PLA columns of the set $C^l$, is called a folding matrix $F$ if the pair $C^u$, $C^l$ involves bipartite folding. Such a matrix satisfies the following properties:

– it is square, $m \times m$ matrix, where $m$ is the number of folding pairs of PLA columns;

– columns of the matrix correspond to the elements of the upper folding region, and rows of the matrix correspond to the elements of the lower folding region;

– $C^u \cap C^l = \varnothing$;

– all matrix elements are 1's.

Thus a folding matrix is a unit square minor of the column disjoint matrix of a PLA. It implies from this that bipartite folding of size $m$ exists if and only if there is a folding matrix of size $m$.

A folding matrix $F$ of the bipartite folded PLA in Fig.2 is specified by the pair $C^u = \{c_7, c_8, c_9, c_{10}, c_{12}\}$ and $C^l = \{c_1, c_2, c_4, c_5, c_6\}$ and is highlighted above in the column disjoint matrix $D$ in thick print.

Identifying the optimal column bipartite folding set is to find a bipartite folding pair of sets of the greatest size, in other words the task is to find unit square minor of the PLA column disjoint matrix $D$ with the greatest number of rows (or columns). That minor will be folding matrix required.

It is evident that the greatest size $m$ of the folding matrix does not exceed $n/2$, where $n$ is the number of PLA columns. The example PLA (Fig. 1) has 13 columns, so it allows for at most 6 pairs of folded columns.

Let the *weight* of the column $c_i$ (row $r_i$) of the column disjoint matrix $D$ be the number of 1's in it. The last row of the matrix $D$ given above shows weights of its columns.

So, if a PLA permits bipartite folding pair of sets of size $m$ then there is at least $2m$ columns having weights greater than or equal to $m$ in the column disjoint matrix [4]. This statement follows from folding matrix definition. In other words, if looking for a bipartite folding of size $m$ all candidate columns and rows of the column disjoint matrix should have weights greater than or equal to $m$.

The example PLA (Fig. 1) has only 9 columns with weights at least 6, so there is no folding matrix of size 6. But the example PLA has 11 columns with weights greater than or equal to 5. Thus a folding matrix is upper-bounded by 5.

## Reducing the search space for bipartite folding

The optimal bipartite folding problem was shown to be NP-complete [3]. The classification of a problem as NP-complete is not sufficient reason to develop only heuristic optimizing methods. In many cases it is

interesting to get just an optimal solution. Some results of reduction of exhaustive computational efforts on the search for the optimal PLA bipartite folding have been described.

The proposed PLA bipartite folding algorithm is organized such a manner to give an optimal solution during an exhaustive search but to allow finding "good" solution at the first its iteration. This algorithm starts from the pruned column disjoint matrix *D*. And then after each algorithm step pruning techniques are made.

The evident way to find a bipartite folding pair of sets of the greatest size is to find a unit minor of the column disjoint matrix *D* of the greatest size. PLA columns that correspond to the unit minor columns can be referred to the upper folding region, and the PLA columns that correspond to the unit minor rows can be referred to the lower folding region. These two PLA column sets $C^u$ and $C^l$ are independent. Before the algorithm functioning the set $C^u$ contains all the PLA columns and the set $C^l$ is empty. It should be noted that the same PLA column is never placed into the unit minor row and column sets simultaneously (both in $C^u$ and in $C^l$) owing to the column disjoint matrix *D* has 0's on the leading diagonal. In the process of the algorithm functioning, some PLA columns will be referred to $C^l$ and eliminated from $C^u$. Thus corresponding rows (or columns) of the PLA column disjoint matrix *D* should be eliminated too.

Before seeking for the minors we can reduce search space, i.e. reduce the matrix *D*. First, for a dense PLA some columns have small weights.

1. It is trivial that columns (and rows) with weights equalled to 0 should be eliminated [4]. They are useless for folding. As well the columns (and rows) with weights equalled to 1 are superfluous: the bipartite folding pair of sets of size 1 can be found trivially.

For example, there exist two superfluous columns (and rows) in the example column disjoint matrix *D*: 3 and 13, having weights 1 and 0.

When the column disjoint matrix *D* has at least 2*m* columns having weights greater than or equal to *m* there can exist a folding matrix of size *m*. It can be found after reducing search space by means of elimination of all columns with weights less than *m*. But if the search fails we will have to decrease the value of *m* and repeat the search as it is proposed in [4]. So, we should choose such a value of *m* that there would be substantially greater than 2*m* columns having weights greater than or equal to *m* to refrain from repetition of the search for folding matrix.

2. The identical rows of matrix *D* should be united. These rows are not disjoint. Otherwise they would not be equal. And they are in the same relation with the rest PLA columns. So they will be in the same folding region in any permissible solution of the folding problem.

After removing superfluous columns (and rows) from the example column disjoint matrix *D* there are three groups of identical rows (and columns): 5 and 6; 7 and 8; 9, 10 and 12. Thus they can be united as it is shown below.

| | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5,6 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7,8 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 9,10,12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| weight | 7 | 5 | 8 | 8 | 8 | 9 | 9 | 7 | 7 | 5 | 7 |

Here weights take into account repetition factor of the corresponding rows.

3. Further, the PLA column (and row) that is disjoint from every other can be added both to the set $C^u$, identifying the upper folding region, and to the set $C^l$, identifying the lower folding region. The column (and row) of the matrix *D*, corresponding to such a PLA column, has all 1's but the only 0 (on the leading diagonal). The column weight is equal to *n*–1. So the column and the row can participate in no process of minor seeking and can be eliminated from the matrix *D*. After desired minor has been found the column can be included either in $C^u$ or in $C^l$, depending on what of them has the less cardinality. Thus we should seek not necessarily a square unit minor but a unit minor of the greatest area.

For column disjoint matrix *D*, depicted above, the group of rows (columns) 7 and 8 is disjoint from all others. So that group can be eliminated from *D*. Having 1, 2, 3 in mind we get the following reduced column disjoint matrix *D*:

| | 1 | 2 | 4 | 5 | 6 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5,6 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| 9,10,12 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| weight | 5 | 3 | 6 | 6 | 6 | 5 | 5 | 3 | 5 |

For dense PLAs, most columns have small weights, resulting in a small size of the column disjoint matrix. For sparse PLAs, most columns have large weights, resulting in a large number of identical rows (and columns). Thus in both cases it is expected that reduced column disjoint matrix *D* to be solved will have moderate size.

## The search algorithm

The maximal unit minor is sought in the process of traversing the search tree in which each node represents a pair of two sets $C^u$ and $C^l$ of unit minor columns and rows. The rows for including to the set $C^l$ are chosen from the set *R* of row-candidates corresponding to the node. $C^u$ includes all columns that correspond to intersection of rows from the set $C^l$. At the first step the set $C^u$ contains all the columns of the reduced column disjoint matrix *D*, the set $C^l$ of minor rows is empty and the set of row-candidates for $C^l$ contains all rows of the reduced column disjoint matrix *D*. The size of a folding matrix, associated with the unit minor described by the sets $C^u$ and $C^l$, is the lesser of the cardinalities of these two sets.

The transition to a son of any node of the search tree consists in adding one more new row-candidate from *R* to the set $C^l$, thus decreasing, in general case the number of minor columns. The optimal solution is found during an exhaustive search of the tree reduced by means of some pruning techniques.

We are going to build successively one by one the unit minors of the Boolean matrix *D*. After getting a new better solution we store it and its size *k*. At each next step we need to consider only those minors of greater size than that early-found ($i > k$). If we get a new better solution we can reduce the Boolean matrix *D* at the sacrifice of elimination of all its columns and rows with weights less than or equal *k*.

*The search tree*. The set of all possible unit minors is organized in the form of a search tree in which each node represents two sets: $C^u$ and $C^l$, presenting columns and rows of the corresponding unit minor and the set *R* of row-candidates for including in $C^l$. The sons of a node represent those minors that contain one additional row in $C^l$. By traversing this search tree all possible unit minors of increasing size should be systematically examined. The process continues until a solution is found or all possible choices have exhausted.

*Pruning the search tree*. Traversal of the search tree is done in a depth-first manner, backtracking from a node whenever the sub-tree rooted by a node has been completely explored. The size of such search tree grows exponentially. However the suggested branch and bound algorithm restricts the exploration to within only a small subset of the nodes in the tree by means of pruning the search tree. A sub-tree can be pruned only if the algorithm can determine that this sub-tree contains no unit minor of size greater than that has been found so far.

A node of the search tree is called viable if the size of the corresponding unit minor is greater than the size of a minor found when traversing the tree before getting to this node. Only viable nodes need to be examined during an exhaustive search of the tree. When reaching a viable node, the rows with weights, less than or equal to the size of the unit minor corresponding to that node, should not be considered. This allows reducing the search space. If in any step a unit minor of the greatest possible size *m* is found then it is the desired maximal unit minor and the algorithm ends.

*Backtracking* is made if not viable node is reached. In that case a row, last included in $C^l$, is discarded and another one, not considered earlier, is selected.

*Heuristic for choosing the traversing path of the tree*. The algorithm suggested employs a simple heuristic to determine an order in which sub-trees, rooted in a node, are traversed so that a near optimal solution can be discovered quickly. Finding a near-optimal solution quickly is important since the sooner such a pair of sets $C^u$ and $C^l$ is found the sooner it can be used for pruning purposes.

The selection steps are important to produce a “good” solution the sooner the better, so to find a solution at the first branch of the search tree as close as possible to the maximum. It appears that justified strategy is greedy one. Therefore, a good heuristic is to select, at the first step, a row with maximum weight as candidate

to be placed in the set $C^l$. Then a row of the matrix $D$ is chosen that most intersects the set $C^u$: in other words it has the greatest number 1's in the columns of the set $C^u$.

Starting with reduced column disjoint matrix $D$ depicted above, the algorithm finds maximal unit minor with $C^u$ = {9, 10, 12} and $C^l$ = {1, 2, 4, 5, 6}. It is highlighted in the matrix $D$ in thick print. Having in mind that PLA columns 7 and 8 may be included either in $C^u$ or $C^l$, we insert them in the set $C^u$ with less cardinality. Thus the following folding matrix $F$ is associated with found maximal unit minor:

| | 7 | 8 | 9 | 10 | 12 |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 1 |
| 6 | 1 | 1 | 1 | 1 | 1 |

The resulting bipartite folded PLA is given in Fig. 2. The area occupied by the bipartite folded PLA is 168 instead of 273 (Fig. 1).

## Additional chance to reduce the area of bipartite folded PLA

After completion of the algorithm for bipartite PLA folding we can get a unit minor that is defined by a pair of sets $C^u$ and $C^l$ of different cardinalities. In this case we have some extra (for the bipartite folding) PLA columns. That is another possibility to reduce the area of the PLA. Such extra members of $C^u$ or $C^l$ can be folded with any PLA columns that are disjoint from them. It is possible to find unit minors with row (or column) set containing only some of the extra PLA columns and with column (or row) set containing PLA columns not included in $C^u \cup C^l$.

An example of such a case of double folding is shown in Fig. 3 (that is example PLA from [4] with modified 13-th column) and Fig. 4. The area occupied by the double bipartite folded PLA is reduced to 147 instead of 168 (Fig. 2).

The following column disjoint matrix corresponds to the example PLA AND plane of Fig. 3:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 8 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 9 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 10 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 12 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 13 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| weight | 8 | 6 | 1 | 9 | 9 | 9 | 10 | 9 | 7 | 7 | 5 | 7 | 5 |

There exists the only superfluous column (and row) in the above column disjoint matrix $D$ – 3, having the weight 1. After its removing there are three groups of identical rows (and columns): 5 and 6; 7 and 8; 9, 10 and 12, they are united. Thus we have the following reduced column disjoint matrix $D$:

| | 1 | 2 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 4 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 5,6 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 7,8 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 9,10,12 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| 13 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| weight | 7 | 6 | 9 | 9 | 8 | 9 | 9 | 7 | 7 | 5 | 7 | 5 |

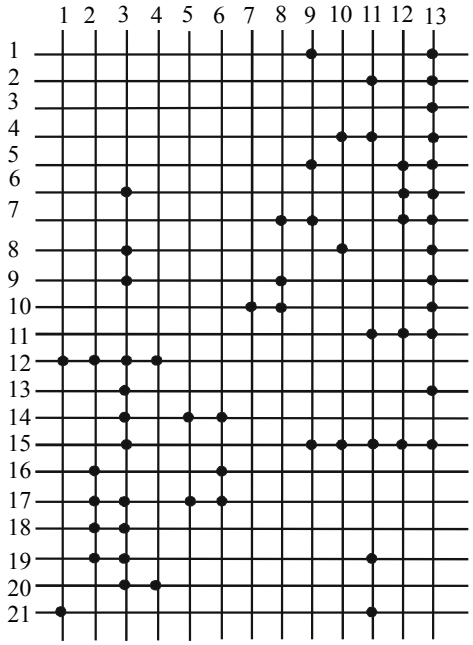



Fig. 3. The second example PLA

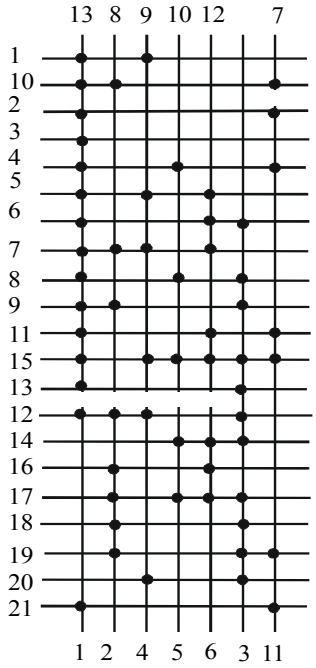



Fig. 4. A bipartite folded second example PLA

The bipartite folding (Fig. 4) is induced from the maximal unit minor with $C^u$ = {7, 8, 9, 10, 12, 13} and $C^l$ = {1, 2, 4, 5, 6}. It is highlighted in the above PLA column disjoint matrix *D* in thick print. Extra PLA columns 7, 8 of the set $C^u$ are disjoint from the PLA column 11 constituting one of the possible unit minors of size 2.

## Conclusion

In this paper a new bipartite folding technique is presented. Compared with the other bipartite folding methods the suggested method has the following new features. The problem of bipartite folding is reduced to a search for a maximal unit minor of a Boolean matrix. The method contains some new procedures of reduction of Boolean column disjoint matrix that allow reducing the search space before functioning the basic bipartite folding algorithm. The approach presented in the paper may lead to the optimal bipartite folding without much wasteful computation.

## References

1. G.D. Hachtel, A.R. Newton and A.L. Sangiovanni-Vincentelli, “An Algorithm for optimal PLA Folding”, IEEE Trans. Computer-Aided Design of Integrated Circuit Syst., vol. CAD–1, no 2, pp. 63–77, 1982.
2. G. DeMicheli and A. Sangiovanni-Vincentelli, “Multiple Constrained Folding of Programmable Logic Arrays: Theory and Applications”, IEEE Trans. Computer-Aided Design, vol. CAD-2, no 3, pp. 151–167, 1983.
3. J.R. Egan and C.L. Liu, “Bipartite folding and partitioning of a PLA”, IEEE Trans. Computer-Aided Design, vol. CAD-3, no. 3, pp. 191–199, 1984.
4. Chun-Yeh Liu and Kewal K. Saluja, “An efficient algorithm for bipartite PLA folding”, IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, vol. 12, no 12, pp. 1839–1847, 1993.
5. J. E. Lecky, O.J. Murphy and R.G. Absher, “Graph theoretic algorithms for the PLA folding problem”, IEEE Trans. Computer-Aided Design, vol. 8, no 9, pp. 1014–1021, 1989.

## Author information

Liudmila Cheremisinova, The United Institute of Informatics Problems of National Academy of Sciences of Belarus, Surganov str., 6, Minsk, 220012, Belarus, Tel.: (10-375-17) 284-20-76,
e-mail: cld@newman.bas-net.by

# QUALITY ASSURANCE IN EXTREME PROGRAMMING

*Plamen Balkanski*

*Abstract: Our previous research about possible quality improvements in Extreme Programming (XP) led us to a conclusion that XP supports many good engineering practices but there is still place for refinements. Our proposal was to add dedicated Quality Assurance (QA) measures, which should be sufficiently effective and at the same time simpler enough in the context of XP. This paper intends to analyze the possibilities for an effective way for applying approved quality assurance practices to XP. The last should not affect negatively to the process and in the meantime must lead to better quality assurance. We aim to make changes to XP that even if would slow down a bit the development process, will make it more suitable for widest range of projects including large and very large projects as well as life critical and highly reliable systems.*

## INTRODUCTION

By our means XP suggests some very good practices proposed almost by every proven software development method but at the same time skips most of the documentation, rejects code reusability and relies on some circumstances which can be considered as not always completely realizable. We think that XP might be extended in a way that it will become more suitable for large and very large projects and teams and at the same time we will be able to keep it extreme. The first step in this direction will be adding the QA. In this paper we will try to propose a convenient way to implement well-known and effective QA practices in the XP process. XP is considered as a defined and disciplined process. Having in mind this we will make comparisons and look for relations using ISO standard 9000-3 which defines Quality Management System relevant to software development and Capability Maturity Model for Software (CMM), which defines at its level 2 the QA practices. The goal of this research will be achieved by examining both sides' practices, comparing activities and analyzing possibilities for implementation of the QA proposed by the above mentioned documents in the XP process.

## SOFTWARE QUALITY ASSURANCE PROPOSED BY CMM

As is well known the CMM describes an evolutionary improvement path from an immature process to a mature disciplined process. CMM defines key practices to improve the ability of the organization to meet goals for cost, functionality and quality.

The QA activities are defined at level 2. According to CMM the purpose of Software Quality Assurance (SQA) is to provide the management with appropriate visibility into the process being used by the software project and of the products being built. It is required that the project follows a written organizational policy for implementing the SQA. The project should have assigned a SQA group that is responsible for coordinating and implementing SQA for the project. This groups should be provided with adequate resources and funding and should include members that are trained to perform their SQA activities.

CMM defines eight activities to be performed as follows:

- A SQA plan is prepared for the software project according to a documented procedure. This plan is developed among with the overall project planning and is reviewed by the affected groups and individuals (managers at different levels, client representative and any other involved) The SQA plan should be managed and controlled.
- The SQA Group's activities are performed in accordance with the SQA plan which includes:
    - Responsibilities and authority of the SQA group
    - Resource requirements for the SQA group
    - Schedule and funding of the project's SQA activities
    - Participation in establishing the software development plan (SDP), standards and procedures for the project
    - Evaluations to be performed
    - Audits and reviews to be conducted
    - Project standards and procedures to be used as the basis for the SQA group's reviews

  - Procedures for documenting and tracking noncompliance issues
  - Documentation to produce
  - Method and frequency to providing feedback to other related groups
- The SQA group participates in the preparation and review of the project's software development plan, standards and procedures. The SQA group provides consultation and reviews on: compliance to organizational policy and to external standards and requirements, standards that are appropriate for use by the project, topics that should be addressed in the SDP and other areas assigned by the project. The SQA group verifies that the plans standards and procedures can be used to review and audit the software project
- The SQA group reviews the software engineering activities to verify compliance
- The SQA group audits designated software work products to verify compliance
  - Software products are evaluated before they are delivered to the customer
  - Software is evaluated against the designated software standards and procedures
  - Deviations are identified and tracked and Corrections are verified
- The SQA group periodically reports the result of its activities to the software engineering group
- Deviations identified in the software activities and software work products are documented and handled according to a documented procedure.
  - Deviations are documented and resolved if possible
  - Not resolvable items are periodically reviewed by senior management until they are resolved
- The SQA group conducts periodic reviews of its activities and findings with customer's SQA personnel as appropriate

At the same time when performing these activities, measurement and analysis are made to be used to find out the cost and schedule status of the SQA activities. These measures include:

- Completion of milestones for the SQA activities compared to the plan
- Work completed, effort expended and funds expended compared to the plan
- Numbers of products audits and activity reviews compared to the plan

CMM also proposes verification of the SQA activities made by 3 different instances.

- The SQA activities are reviewed by the senior management on a periodic basis to provide awareness of and insight into software process activities at an appropriate level of abstraction
- The SQA activities are reviewed with the project manager on both a periodic and event driven basis
- Independent experts periodically review SQA activities and software work products of the SQA group.

## QUALITY ASSURANE PROPOSED BY ISO 9000-3

ISO 9000-3 is the standard of the ISO 9000 series that is most relevant to software development and maintenance. Organizations typically use ISO 9000 standards to regulate their internal quality systems and assure the quality systems of their suppliers. ISO proposes a quality assurance manual that consists of management responsibilities, a set of measurements, analysis and improvement activities, and required documentation.

An ISO 9000 organization should have implemented a Quality Management System (QMS) that is continuously maintained for effectiveness and process improvement. The effectiveness of the quality management system should be improved by the use of Quality Policy, quality objectives, audit results, analysis of data, corrective and preventive actions and management reviews.

The organization defines and documents its Policy, which provides the overall objectives for an effective QMS. The Quality Policy should be relevant to the organization's goals and expectations of its customers.

ISO 9000 requires an organization to plan and perform audits. The results of audits are communicated to management and any deficiencies found are corrected.

ISO 9000 states that organizations must establish adequate statistical techniques and use them to verify the acceptability of process capability – this is also called measurement. According to ISO 9000-3 "there are currently no universally accepted measures of software quality". The auditors can accept the use of statistical tools or any consistently collected and used data.

The organizations should implement and maintain documented procedure to initiate corrective and preventive actions. Corrective action procedures define the requirements for:

- Reviewing non conformities including customer complaints
- Determining causes of non-conformities
- Evaluating the need for action to ensure that non-conformities do not recur
- Determining and implementing the action needed
- Records of the results of action implemented
- Review of corrective action implemented

The QA manager is responsible for Corrective and Preventive actions and a feedback system should be used to provide early warnings of quality problems. Preventive action procedures define requirements for:

- Determining potential non-conformities and their causes
- Evaluating the need for action to prevent occurrence of non-conformities
- Determining and implementing the action needed
- Records of the results of action implemented
- Reviewing preventive action implemented

The QMS documentation structure can be described at five levels. "Level 1" is maintained in the form of Quality Policy. "Level 2" documentation is maintained in the form of the Quality Assurance Manual. "Level 3" consists of quality procedure; "Level 4" contains work instructions. "Level 5" documentation is maintained as records/reports.

## QA IN XP AND WAYS OF IMPROVEMENT

Adding QA seems to be the easiest part on the way to develop XP version suitable for large teams and projects. If we study the XP practices carefully we will notice that there already is some form of QA. It just need to be structured and controlled as well as somehow documented in order to use its results for analyzes. Bellow we will try to find out which of the above practices, measures and recommendations are suitable for the XP process, what roles are needed to conduct the new actions and how the collected information can be used. All the changes must be made in a way that would not slow down the development but still reaching the desired QA process.

We will go through the QA practices proposed by CMM and ISO and then will review the situation with XP and will propose our solution where appropriate for actions to be taken, roles to be added or other changes.

a) QA

- CMM defines the SQA aim as: providing the management with appropriate visibility into the process.
- ISO proposes a quality assurance manual that consists of management responsibilities, measurements, analysis and improvement activities, and required documentation.

XP:

- In XP the QA is not even mentioned but a lot of QA practices are presented. These are:
    - The automated acceptance tests to prove that a feature is implemented correctly. We can add here the *test-first development XP practice, which also relies on the acceptance tests.*
    - Pair programming which assures that the entire source code is reviewed all the time.
    - Refactoring practice, which means removing duplication, increasing code integration and lowering mixture of the code.
    - Collective ownership, which proposes code reviewing a common coding standard.

We see that XP uses QA practices but the difference is that these practices are only oriented to achieve the direct goal. For example Automated Tests intent to assure that the current version is good enough for a release, but does not keep any data for the results and does not makes any decisions. The same is for Pair Programming – both programmers care for the same code and make fewer errors but the errors are not recorded.

b) QA group

- CMM: It is required that the project follows a written organizational policy for implementing SQA. The project should have assigned a SQA group that is responsible for coordinating and implementing SQA for the project
- ISO: The organization should have implemented a QMS that is continuously maintained for effectiveness and process improvement. The effectiveness of the quality management system should be improved by the use of Quality Policy, quality objectives, audit results, analysis of data, corrective and preventive actions and management reviews.

XP:

- According to XP every project is different so it needs a different approach. However it is obvious that every team, which is working on more than one project gradually, builds its own software development process. So the next time everyone in the team understands better his responsibilities. The QA is in many ways independent from every single project system, which can be easily achieved in an XP team by adding another role – the QA Manager. As all other roles in the team individual that already has a role but not a programmer also can handle the QA role. The QA Manager will be involved in creating the plans and gathering any analytical information as well as taking decisions about any changes provoked by the QA results.

c) Activities

- CMM defines eight activities as listed above.
- ISO suggests that the organizations implement and maintain documented procedure to initiate corrective and preventive actions also described above.

XP:

Which of the proposed actions already exist in XP and which are a must? Both standards propose a documented procedure (ISO) or an SQA plan (CMM) that is missing in XP. We propose that the QA Manager must develop a documented procedure for the QA of an XP team. The QA process is very similar in most of the projects. Thus the QA manager can use a pattern, which can be tuned as appropriate for every project and of course developed through the process. Using such patterns or a generalized QA plan for the XP projects we think will save time or at least will not lead to serious delays.

In the context of XP we guess that it is better to use only one person for the QA instead of a group. This is needed for a number of reasons including keeping a small team and funding reasons. The QA Manager will take actions defined here and some automated registration programs (ARP) will help him with gathering the information. At the meantime as we cannot reject Code reviews then a lot of QA will be done by programmers but here come the ARPs which will present the QA manager with relevant information which can be used directly. The participation of the QA manager in establishing any plan or procedure or standard is automatically guaranteed by the fact that the team works together. Using common standards is related with the practice Collective Ownership that requires such common standards in every XP team.

The QA Manager should define a QA schedule. However we guess that this should be in relation to the versions issued by the XP team. This is because on every version there is a lot of testing and in this moment the integration can be also assessed.

Evaluations, tracking and reporting of the non-conformities are related to the used measures and will be discussed bellow. What we must note here is that the QA Manager should take part in defining the automated tests, as this is a good approach to evaluate clients as well as project internal quality measures.

d) Measures

- CMM proposes using measures which include:
  - Completion of milestones for the SQA activities compared to the plan
  - Work completed, effort expended and funds expended compared to the plan
  - Numbers of products audits and activity reviews compared to the plan
- ISO states that organizations must establish adequate statistical techniques and use them to verify the acceptability of process capability. These can be the use of statistical tools or any consistently collected and used data.

XP:

- The first thing to say when writing about XP and measures is that in the initial process there are no measures. There are no analytical data as the only aim is to complete each project quickly and then to start over again with the next.

However we will need measures, as we would like to use the results for management decisions. As we have already mentioned above we propose using a number of ARPs which will record most of the measures like code compilation failures, syntax or functional changes, acceptance tests results or even programmers effective time. Such automatic registration can also be easily applied to some popular metrics as Lines of Code or Functional Points. All other tests as logical program behavior or program usability must be planned and assigned for failures registration by the QA Manager. Such failure registrations we propose must be added to the same database used by the ARPs.

Reporting of failures must be delivered to the QA Manager who can re-assign the problems to the team or send the lists to the Team Coach for re-assigning. Reporting of measures which do not directly affect the program functioning are delivered to the QA Manager, which can discuss them with other team members with management responsibilities and propose changes regarding the process.

## CONCLUSION

By using the experience of ISO 9000-3 standard and CMM for Software regarding QA we made an analyze and conclude that adding QA to XP will be harmless for the process and will lead to the desired results – measuring which will provide the managers with relevant information and thus they will be able to make changes to improve their process. The main changes proposed by us are adding a new role – the QA manager that is responsible for the QA plan development and conducting and the use of ARP – small agents to record a number of measures which will be additionally proposed.

As was stated in the paper the QA manager on its own or by assigning tasks to the developers and using the ARPs fulfills nearly completely the requirements proposed by both ISQ 9000-3 and CMM. At the same time it is clear that we will need a list of measures, which will be appropriate and can give the exact information that will be most useful and proper for an XP team. This is the first important step on the way developing an XP version, which can be used as a process for development of large and very large projects.

## Literature

1. A.Eskenazi, P.Balkanski - Quality Improvement in XP Development
2. Beck K., Jeffries R., Anderson A., Hendrickson C., Jeffries Ronald E. Extreme Programming Installed
3. SW-CMM, Software Quality Assurance
4. ISO 9000-3, Quality management and quality assurance standards
5. Software metrics, SEI Curriculum Module SEI-CM-12-1.1

## Author Information

Plamen Balkanski - I.T. Consultant in Global Industries UNISYS; Email - Plamen.Balkanski@unisys.com

*About...*

# THE JOURNAL OF RUSSIAN ASSOCIATION FOR ARTIFICIAL INTELLIGENCE "NEWS OF ARTIFICIAL INTELLIGENCE"

Beginning from 1991, Russian (initially Soviet) Association for Artificial Intelligence (RAAI) publishes the own journal 'News of Artificial Intelligence'. The journal is founded on the initiative of the famous specialist in the field of Artificial Intelligence (AI), the first president of Soviet Association for Artificial Intelligence, the academician of Russian Academy of Natural Science (RANS), doctor of technical sciences (d.t.s.), professor D.A. Pospelov, which from 1991 up to 2001 was its main editor. From 2001 the main editor is the academician of RANS, d.t.s., professor E.V. Popov. In the staff of the Editorial Council the great specialists in the theoretical and applied problems of informatics and AI was included: president of RAAI, corr.-member of RANS, doctor of physic-mathematical sciences, professor G.S. Osipov, the chairman of Scientific Council of RAAI, the academician of RANS, d.t.s., professor O.P. Kuznetsov, the academician of Russian Academy of Sciences O.I. Larichev, the academicians and corr.-members of RANS, d.t.s., professors T.A. Gavrilova, V.V. Golenkov, V.V. Emelyanov, A.P. Eremeev, V.V. Kureichik, D.A. Pospelov, V.L. Stefanuk, I.B. Fominykh, V.K. Finn, V.F. Khoroschevsky, A.I. Erlich, A.E. Yankovskaya.

In the journal there was issued more than 60 numbers including thematic issues and special issues dedicated to the memory of famous specialists in the field of informatics and AI — A.F. Blishun (1991), R.Ch. Zaripov (1995), M.G. Gaase-Rapoport (1997), A.N. Melichov (1997), G.S. Pospelov (1998), O.I. Larichev (2003), and also jubilees issues — to V.K. Finn's 60-years (1993), to V.L. Stefanuk's 60-years (2000), to L.A. Zadeh's 80-years (2001), to D.A. Pospelov's 70-years (2002), to V.K. Finn's 70-years (2003).

In the journal scientific overviews, discussions of specialists, materials on the history and developing AI are published. There are also discussed adjacent scientific disciplines in Russia and abroad, official messages and plans of RAAI activity, information on conferences and seminars devoted to theoretical and applied problems of AI. The important place in the journal structure presents the sections: 'From unpublished books', 'Developing AI ideas', 'Intelligent data analysis and knowledge engineering', 'Practical problems of AI', in particularity, ' Artificial intelligence and business', 'Intelligent decision support systems', 'Intelligent (expert) systems in medicine', 'Intelligent tutoring systems', 'Dynamic intelligent systems', 'Knowledge management systems', 'Natural languages systems' and so on. There are published remembering about scientific schools in the area of informatics and AI, information about the staff preparation on AI in the leading universities of Russia and abroad, advertisements of new information and communication technologies.

From 2001 in the connection with a great popularity of the journal its periodicity is increased up to 6 numbers per year and the number quantity — up to 1000 with using more modern polygraphic base. The journal distribution is organized through subscription (index in ROSPECHAT 81216) and retail selling.

The Editorial Council invites specialists on the preparation of materials, issue, journal distribution and also advertisement distributors. With suggestions You can addressed to the Editorial Council or the RAAI Scientific Council. Information on the journal, possibility of its selling and other RAAI activity directions can be obtained at RAAI site on the address www.raai.botik.ru and on the addresses www.ainews.ru, www.anakharsis.ru.

The Main Editor of the journal
'News of Artificial Intelligence'

E.V. Popov

# ABOUT MANUSCRIPTS FOR IJ ITA

*Kr. Markov, Kr. Ivanova, I. Mitov*

*Abstract: The rules for the papers in the IJ ITA are outlined.*

*Keywords: IJ ITA, formatting rules.*

## Introduction

IJ ITA welcomes scientific papers connected with any information theory or its application. Original and non-standard ideas will be published with preferences. Responsibility for papers published in IJ ITA belongs to authors.

This text is a sample for preparing the articles for publishing in IJ ITA. All styles needed for formatting the papers are included. IJ ITA rules for preparing the manuscripts are compulsory.

*Papers must be written in English.* Please *get permission to reprint* any copyrighted material before you send it to IJ ITA. The camera-ready copy of the paper should be received by e-mail: foi@nlcv.net

## Instructions for Manuscripts Preparation

The authors are hoped to prepare manuscripts in close accordance with the instructions given below.

The articles should written in RTF or DOC format, eight (8) pages, A4 paper (210 x 297mm). The typing area is 160 x 242mm. Margins of the paper sheet are: top - 30mm; bottom - 20mm; left, right - 25mm.

Typing styles: The paper should begin with the title of the paper (use the style "Title") and the name(s) of the author(s), (use the style "Authors"). Use "Normal" style - Arial Narrow or similar font; 11pt; single-spaced text; 3pt before each paragraph; *without special indents; left and right justification*.

Use style "Subtitle" for the titles of the separated sections of the article.

Figures and tables: these should be positioned in the body of the text, as close as possible to the relevant text. Color figures cannot be reproduced.

Abstract: A brief decryption of the contents of the article should be placed of the beginning of the text. Use style "Abstract" for formatting the abstract and keywords.

References: References in the text should be keyed with parenthesized Arabic numbers or with the name(s) and year of the referred material - for instance [Shannon, 1949]. After the text of the article, put list of Bibliography using the style "Bibliography".

Author information: Finish the article with the author information: names of the author(s), organization(s), post and e-mail address(es).

## Conclusion

This exemplar is meant to be a model for manuscript format. Please make your manuscript look as much like this exemplar as possible. Except for formatting and the inclusion of an abstract, authors have complete freedom to structure their papers as they wish; the section headings need not be as given in this exemplar. In the case of serious deviations from the format, the paper cannot be published.

## Bibliography

[Shannon, 1949] C.E.Shannon. The Mathematical theory of communication. In: The Mathematical Theory of Communication. Ed. C.E.Shannon and W.Weaver. University of Illinois Press, Urbana, 1949.

## Author information

Krassimir Markov - IJ ITA, Editor in chief; P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: foi@nlcv.net

Krassimira Ivanova - Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; e-mail: foi@nlcv.net

Ilia Mitov - ITHEA - FOI Institute of Information Theories and Applications, P.O.Box: 775, Sofia-1090, Bulgaria; e-mail: foi@nlcv.net

# TABLE OF CONTENTS