ITHEA

International Journal

INFORMATION THEORIES & APPLICATIONS

2007 Volume 14 Number 1

International Journal
INFORMATION THEORIES & APPLICATIONS
Volume 14 / 2007, 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.
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: *info@foibg.com* .
Responsibility for papers published in IJ ITA belongs to authors.

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

International Journal "INFORMATION THEORIES & APPLICATIONS" Vol.14, Number 1, 2007

Printed in Bulgaria

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

Publisher: **Institute of Information Theories and Applications FOI ITHEA**
Sofia, 1000, P.O.B. 775, Bulgaria. www.foibg.com, e-mail: info@foibg.com

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

PREFACE

Verba volant, scripta manent !

The "**International Journal on Information Theory and Applications**" (IJ ITA) has been established in 1993 as independent scientific printed and electronic media. IJ ITA is edited by the *Institute of Information Theories and Applications FOI ITHEA* in collaboration with the leading researchers from the Institute of Cybernetics "V.M.Glushkov", NASU (Ukraine) and Institute of Mathematics and Informatics, BAS (Bulgaria).

During the years, IJ ITA became as well-known international journal. Till now, including this volume, more than 625 papers have been published. IJ ITA authors are widespread in 39 countries all over the world: *Armenia*, *Belarus*, Brazil, Belgium, *Bulgaria*, *Canada*, Czech Republic, Denmark, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, *Israel*, Italy, Japan, Kirghizia, Latvia, Lithuania, Malta, Mexico, Moldavia, Netherlands, Poland, Portugal, Romania, *Russia*, Scotland, Senegal, Serbia and Montenegro, *Spain*, Sultanate of Oman, Turkey, *UK*, *Ukraine*, and USA.

Volume 14/2007 of the IJ ITA contains 60 papers written by 121 authors from 10 countries (marked in italics above), selected from several international conferences, seminars and workshops organized or supported by the Journal.

At the first place, the main source for selection were the ITA 2006 Joint International Events on Information Theories and Applications, (June 20-30, 2006, Varna, Bulgaria):

- XII-th International Conference "Knowledge-Dialogue-Solution" (KDS 2006);
- V-th International Workshop on General Information Theory (GIT 2006);
- IV-th International Conference "Information Research and Applications" (i.TECH 2006);
- IV-th International Workshop on Multimedia Semantics (WMS 2006).

Several papers were selected from the pool of papers directly submitted to IJ ITA.

Main characteristic of ITA 2006 International Conferences was that the papers were combined into thematic sessions. Because of this, the selected papers are published in this volume following the thematic sessions' organisation.

Congratulations to **Juan Castellanos** (Spain) and **Georgi Totkov** (Bulgaria) who were awarded by the International Prize "**ITHEA**" for the year 2006. The "ITHEA" Prize has been established in 1995. It is aimed to mark the achievements in the field of the information theories and applications.

More information about the IJ ITA rules for preparing and submitting the papers as well as how to take out a subscription to the Journal may be obtained from www.foibg.com/ijita.

Krassimir Markov

IJ ITA Founder and Editor in chief

International Prize "ITHEA"

Awarded Scientists till 2006:

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*
2004	*Varna*	*B. Kokinov, T. Vamos*
2005	*Varna*	*L.F. de Mingo, M. Dobreva*
2006	*Varna*	*J. Castellanos, G. Totkov*

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*

Bioinformatics
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
Quality of the Programs
Software Engineering
Theory of Computation

APPLICATIONS

Business Information Systems
Communication Systems
Computer Art and Computer Music
Hyper Technologies
Intelligent Information Systems
Multimedia Systems
Programming Technologies
Program Systems with Artificial Intelligence
Pyramidal Information Systems
Very Large Information Spaces

BASIC STRUCTURE OF THE GENERAL INFORMATION THEORY

Krassimir Markov, Krassimira Ivanova, Ilia Mitov

***Abstract**: The basic structure of the General Information Theory (GIT) is presented in the paper. The main divisions of the GIT are outlined. Some new results are pointed.*

***Keywords**: General Information Theory.*

***ACM Classification Keywords:** A.1 Introductory and Survey*

Introduction

There exist several common theoretical information paradigms in the Information Sciences. May be, the most popular is the approach based on the generalization of the Shannon's Information Theory [Shannon, 1949], [Lu, 1999]. Another approach is the attempt to be synthesized in a common structure the existing mathematical theories, which are applicable for explanation of the information phenomena [Cooman et al, 1995].

Besides of this, we need to point the diligence of the many researchers to give formal or not formal definitions of the concept "information". Unfortunately, although they are quite attractive in some cases, these definitions did not bring to any serious theoretical results [Abdeev, 1994], [Bangov, 1995], [Markov P., 2002], [Tomov, 1991], [Elstner, 1993].

At the end, there exist some works that claim for theoretical generality and aspire to be a new approach in the Information Science, but theirs authors should clear up what they really talk about [Burgin, 1997].

The theoretical base of the informatics needs the philosophical support and substantiation to become wide accepted scientific paradigm. This way, the closely scientific research in the domain of informatics would be able to leap across its boundaries and to become as elements of the scientific view of life.

Discovering the common philosophical base has exceptional importance.

The philosophical rationalizing and substantiating of the information phenomena become as leading goal of the scientific knowledge.

Starting point need to be the consideration that the General Information Theory (GIT) needs to be established as internal non-contradictory logical system of contentions [Markov et al, 1993]. This rule contrasts the understating of the informatics as a mosaic of formal theoretical works and applications.

Basic requirement is that the GIT needs to explain the already created particular theories and paradigms.

The mathematical structures ought to serve as a tool for achievement the precise clearness of the philosophical formulations and establishing the common information language for describing and interpreting the information phenomena and processes.

The second very important requirement is to build the GIT on the base of the inceptive philosophical definition of the concept "information" using as less as possible the primary undefined concepts with maximal degree of philosophical generalization. This requirement follows the consideration that ***the concept "information" is not mathematical concept.*** The behavior, peculiarity and so on could be described by the mathematical structures but this is another problem. In this case, the accent is stressed on the comprehension that the information has purely material determination and that it is a consequence of the interaction between the material objects as well as of the real processes and phenomena occurred in them and with them.

The presented in this paper General Information Theory (GIT) is based only on primary consideration of the world as variety of entities, which are formed by relationships between entities that form lower levels.

The development of GIT had started in the period 1977-1980. The first publication, which represents some elements of GIT, was published in 1984 [Markov, 1984]. The establishment of GIT was not rectilinear.

Occasionally, the influences of other paradigms have disturbed this process and have turned it to the vain effort [Burgin, Markov, 1991].

The fundamental notion of the GIT is the concept "Information". All other concepts are defined based on this definition. In 1988, the not formal definition of the concept of Information was published in [Markov, 1988]. It became as a fundamental definition for the General Information Theory [Markov et al, 1993], [Markov et al, 2003a]. The translation of the philosophical theory into the formal one is a good approach for verification of the scientific ideas [Markov et al, 2003b], [Markov et al, 2004]. Because of this, the basic concepts of the General Information Theory ware presented philosophically and formally.

This paper is aimed to present the internal structure of GIT in its current state. For this purpose we will remember some main results as well as we will discuss the new achievements of GIT.

The GIT is build by three specialized theories:

- Theory of Information,
- Theory of Infos,
- Theory of Inforaction.

Theory of Information

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]. Several attempts to develop a formal definition were introduced during the years [Markov et al, 2003b], [Markov et al, 2004].

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;
- established 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****.*

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 realization 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 realized 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 organization 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 lower levels of the structure of the entity. Such kind of entities has relatively free sub-entities with own behavior 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.

Finally let remark that the reflection could not be detected by the entity that contains it. This is dialectical behavior of the reflection - it is only an internal change caused by the interaction.

Information

The real world contains unlimited number of entities. When an entity contacts another, there exists a great possibility to join third entity in this process. It is clear; the third entity may contact and reflect each of others as well as the process of realization of the interaction between them — the process of realization of the contact is a specific (temporal) forming relationship between entities and during the process of establishing the contact the entities form new (temporal) entity which in the same moment may be reflected by the third entity. So, the third entity may reflect any vestiges of this interaction from both first and second entities.

In the special case when the third entity contains reflections of the first entity received by both 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 (-ies)

then the third entity became as an external relationship between first entity and its reflection in the second entity – it became as ***"reflection evidence"*** of this relationship.

✓ The first entity is called ***reflection source;*** the second entity is called ***reflection recipient***; and the third entity is called ***reflection evidence***.

In this special case, when there exist the triple

"(source, recipient: evidence)",

the reflection of the first entity in the second is called ***information*** in the second for the first entity.

Let point one very important case of the real world - simultaneous contacts of the three entities. Every one of them may be source, recipient and evidence in the same time. There exist six cases which represent the simultaneous contacts of three entities. Therefore, the entities $\mathfrak{A}$, $\mathfrak{B}$ and $\mathfrak{C}$ may be in the next six reflection relations: ($\mathfrak{A}$, $\mathfrak{B}$: $\mathfrak{C}$); ($\mathfrak{B}$, $\mathfrak{C}$: $\mathfrak{A}$); ($\mathfrak{C}$, $\mathfrak{A}$: $\mathfrak{B}$); ($\mathfrak{A}$, $\mathfrak{C}$: $\mathfrak{B}$); ($\mathfrak{C}$, $\mathfrak{B}$: $\mathfrak{A}$); ($\mathfrak{B}$, $\mathfrak{A}$: $\mathfrak{C}$).

All reflection relations are equivalent from point of view of the interrelations between reflection source, reflection recipient and reflection evidence. Because of this we will discuss only the case ($\mathfrak{A}$, $\mathfrak{B}$: $\mathfrak{C}$).

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, iff there exist reflection evidence than 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

"***The information is reflected relationship".***

So, the *reflection* of the first entity in the second one is "***information"*** for the first entity if there is corresponded *reflection evidence*. The generalization of this idea leads to assertion that ***every reflection can be considered as information, iff there exists corresponding reflection evidence.***

General Structure of Information

The entities and theirs relationships form space hierarchies. Every entity contains all entities of its low levels. In this sense, we can say that every relationship contains in itself the relationships of low levels of the entity.

As reflected relationship, the Information is reflected space hierarchy of all relationships of this one. From this point of view, we can say the general structure of information reflects general structure of real relationships.

The information of one level contains space hierarchy of information for low levels. Therefore, the main idea is:

The General Structure of Information is a Space Hierarchy.

Information Elements and Information Memory

The triple

***i* = (source, recipient : evidence)**

defines concrete ***(single) information element.*** The triple "**i**" is called ***"information relationship".***

The ***(information) memory*** of the entity is the set of all information elements, which are reflected in the entity.

It is clear, from point of view of the period of existing of the corresponded reflections; the entity memory may be more *temporal* or more *permanent.*

Information Spaces and Information Environment

The information elements are real reflections in the entities and they exist in the real world. This means that for every contact or interaction as well as for every single entity or set of entities the corresponded sets of information elements may exist.

The set of information elements, which is defined by:

- single source and single recipient, is called ***single information space***;
- many sources and single recipient, is called ***common information space***;
- single source and many recipients, is called ***single information environment,*** which contains many information spaces;
- many sources and many recipients, is called ***common information environment***.

Types of the Information

The information is a result from the interaction. It is a kind of the reflection. Therefore, the information has the corresponding properties.

Especially, we have primary interaction, secondary (transitive) interaction, self-interaction etc.

This way, there exist corresponding types of the reflection and the main types of information are:

- direct information;
- transitive information;
- transitive self-information;

- interactive direct information;
- interactive transitive information;
- interactive transitive self-information.

From other point of view, the interaction may be provided on different levels of the structure of the entities. Therefore, we may talk about corresponded types of information.

The types of information memories as well as the structures of the information environments define corresponded types of information, too.

The Further Investigation in the Theory of Information

The further investigation and development of the Theory of Information may be directed towards investigation the types and characteristic of information in correspondence with the specific of entities and relationships as well as characteristics of the environment.

Contacts and Interactions need to be investigated according different types of entities.

The philosophical support is very important so the research need to take in account the "Theory of reverberation" [Pavlov, 1987] as well as the development and extending of ideas about reflection and self-reflection given in this paper.

The main place needs to take the investigation of the types and possible interconnections of the basic information triple ***i* = (source, recipient: evidence)** as well as the types and the characteristics of the direct, transitive, and interactive information and self-information based of the hypothesis about the general structure of information.

Special attention needs to be paid on the basic types of information in closely correspondence of type of interaction and reflection as well as the different levels of the structure of the entities.

As we have seen, the types and the characteristics of the information memories, the information spaces as well as the information environments are another main theme of this theory.

Theory of Infos

The genesis of the concept of **Infos** started from the understanding that the concept **"Information Subject"** is perceived as human characteristic. It is clear that in the nature there exist many creatures which may be classified to this category. To exclude the misunderstandings we decide to introduce new word to denote all possessors of the characteristics of the Information Subject.

This word is **"*Infos*"**.

Activity and Information Expectation

Every forming relationship as well as every relationship unites the entities and this way it satisfies some theirs possibilities for building the relationship by establishing the contact. In other words, for creating the forming relationship we need:

- entities, from which the new entity is able to built;
- possibilities of the entities for establishing the contact by satisfying of which the forming relationship may be originated.

The forming relationship is the aggregate of the satisfied possibilities for establishing the contact.

It is clear that after establishing the relationship we may have any of two cases:

1) all possibilities of the entities for establishing the contact are satisfied by such possibilities of other entities;
2) there are any free possibilities after finishing the establishment of the new relationship - on the low levels of the entity or, if it is a new entity, on the level of the whole entity. Disintegration of the whole entity or any its part may generate any possibilities too.

In the second case, the entity has any ***"free valences"***, which needs to be satisfied by corresponded contacts with other entities. We may say the entity has ***activity*** generated by the free possibilities for establishing the contacts with the entities from the environment.

The process of interaction is satisfying the possibilities for contact of the entities. From point of view of the entity, the interaction may be external or internal.

During the interaction given entity may be destroyed partially or entirely and only several but not all parts of the destroyed entity may be integrated in the new entity. This means that there exist both constructive and destructive processes in the process of interaction between entities. The determination of the type of the interaction depends on the point of view of given entity. The interaction dialectically contains constructive and destructive sub-processes.

If the entity is a complex, it is possible for it to have an opportunity of self-reflection. In such case, it is able to reflect any reflection, which has been already reflected in it. In this case, because of the new internal changes (self-reflection) the entity may obtain any new ***"secondary activity"***.

The secondary activity is closely connected to the structural level of the entity, which correspond to the level of the self-reflection. This way the secondary activity may be satisfied by internal or external entity from point of view of the given entity. In other words, the ***resolving*** of the secondary activity may be ***internal*** or ***external***.

During the establishment of the information relationship it is possible to be generated any secondary free activity (possibilities on the low levels of the entity or on the level of the whole entity) which needs to be satisfied by corresponded contacts with other entities.

The secondary activity generated by the information relationship is called ***"information activity".***

On given level of complexity of the entities a new quality becomes — the existence of self-reflection and internal activity based on the main possibilities for contact of the sub-entities as well as based on the new (secondary) possibilities created after internal self-reflection.

The internal activity may be resolved by:

- the internal changes which lead to partial internal disintegration of the sub-entities and theirs a posterior internal integration in the new structures;
- the external influence on the environment.

The internal changes may lead to removing of some sub-entities if they have no possibilities for integration with the others, i.e. if they have no free valences to be resolved in the process of integration.

The external influence is the most important. The impact on the entities around the entity is the way to resolve its activity. The destroying of the external entities and including the appropriate theirs parts in itself is the main means to exist and satisfy the free valences.

One special kind of activity is the information one. We assume that the secondary activity needs to be resolved by relevant to the information valences corresponded opposite (information) valences which need to be of the same genesis, i.e. generated by any information relationship. So, not every entity may be used for resolving the secondary activity.

This way, the entity expects a special kind of (information) contacts and (information) interaction for resolving the information activity. Because of this the information activity is called ***"information expectation".***

The Information Witness

Let remember the special case from above when the third entity contains reflections of the first entity received by both 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 (-ies).

In this case the third entity became as an external relationship between first entity and its reflection in the second entity — it became as ***"reflection evidence"*** of this relationship.

In addition, if during establishing the information relationship ***i*** = (source, recipient: evidence) in the reflection evidence is generated information expectation (activity) it is called ***"information witness".***

As the information witness is more complex entity so the information relationship may be more complex. In addition, let remark that the complex reflection is time-depended process. In other hand, the memory and actual context determine the result of the complex reflection.

The Information is a Model

As Marx Wartofsky remarks, the concept "**model**" has been used for denotation of the very large class of phenomena: mechanical, theoretical, linguistic, etc. constructions. He gave a good definition of the model relation and made clear the main characteristics of the model [Wartofsky, 1979]. This definition is as follow:

The model relation is triple M:

$$M: (S, x, y)$$

where "S" is subject for whom "x" represents "y". In other words only in this relation and only for the subject "S" the entity "x" is a model of the entity "y".

As we point above, the interaction between two entities is a specific theirs relationship. If there exist information witness (**W**) of the interaction between two entities as well as of the existence of the information about the first entity in the second entity, **W** became as subject for whom the information in the second entity represents the first one. In other words, there exists relation

$$M: (\mathbf{W}_{BA}, I_{BA}, A),$$

where "A" and "B" are entities, and the $\mathbf{W}_{BA}$ is the information witness, which proofs that the assertion "$I_{BA} \subset B$ is information in B for A" is true.

In the relation ($\mathbf{W}_{BA}$, I_{BA}, A) the information I_{BA} is a model of A**.**

The Information Model

The entities of the world interact continuously in the time. It is possible, after any interaction one another may be realized. 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.

Let A, B and C are entities; A and B interact and after that B interacts with C. If there exist the relations:

- M_{BA}: ($\mathbf{W}_{BA}$, I_{BA}, A), where $\mathbf{W}_{BA}$ is the information witness, which proofs that the assertion "$I_{BA} \subset B$ is information in B for A" is true,
- M_{CB}: ($\mathbf{W}_{CB}$, I_{CB}, B), where $\mathbf{W}_{CB}$ is the information witness, which proofs that the assertion "$I_{CB} \subset C$ is information in C for B" is true,

and if $M_{C(B)A}$: ($\mathbf{W}_{C(B)A}$, $I_{C(B)A}$, A), where $\mathbf{W}_{C(B)A}$ is the information witness, which proofs that the assertion "$I_{C(B)A} \subset C$ is information in C for information in B for A" is true.

In such case, from point of view of the $\mathbf{W}_{C(B)A}$ the information $I_{C(B)A}$ is a model of A. In other hand, because of transitive reflection, $I_{C(B)A}$ is created as reflection of the I_{BA} but not directly of A.

This means that **$I_{C(B)A}$ is a model of the information in B for A**.

In other words the $I_{C(B)A}$ is an ***information model*** in C for A**.**

The collecting of information models for given entity in one resulting entity may exist as a result of the process of interaction between entities. Such process is in the base of the ***Information modeling***.

If an information model **IM** contains information for (reflected from) the two source information models $\mathbf{IM}_1$ and $\mathbf{IM}_2$ than the source information models are "***similar***" in the sense of the model **IM**.

The similarity of the information models causes the establishing the relation of aggregation between them.

The relation of similarity aggregates the similar models in new ***internally determined information model*** in the memory of the information witness.

The aggregation may cause the generating the new information activity, which may be resolved not only in the environment around the information witness. The possibility of self-reflection may cause the generating the new information models in his memory without any external influence and so on.

This process of aggregation and generation of new models is not limited.

The (information) models internally generated via self-reflection are called "***mental (information) models***" of the information witness.

Resolving the Information Expectation

Because of the existing of the information expectation, i.e. the existing of the secondary information activity, the Information Witness "expects" to combine the information valences with any others.

The combining the valences of the information expectation with some others is called ***resolving the information expectation***.

Let "n" is the number of free valences in an information expectation. After the contact some of them are combined as well as the others are not. The new valences, which are generated by the contact, do not belong to the information expectation before contact. They may form new information expectation but the basis for our reasoning will be the starting information expectation.

The normalized by "*n*" number *D'* of the not combined valences is called ***degree of discrepancy*** (**D**) of the incoming reflection to the information expectation, i.e.

$$\mathbf{D} = \frac{D'}{n}$$

The normalized by "*n*" number *C*' of the combined valences is called ***degree of combining*** (**C**) of the incoming reflection to the information expectation, i.e.

$$\mathbf{C} = \frac{C'}{n}$$

There exists the equation: **C + D = 1.**

From point of view of given expectation for contact the number of free valences is fixed. After the contact, as a result of reflection, some of the free valences of the entity may be combined with any new (internal or external) valences. Of course, new free valences may occur. The number "*n*" varies in the process of interaction. Every contact may change it.

The more valences of the information expectation have been resolved, the more qualitative is the incoming information and vice versa.

The difference **A** between normalized number **C** of resolved valences and normalized number **D** of not resolved valences of the information expectation is called ***adequacy of the reflection to the information expectation***, i.e.

$$\mathbf{A = C - D}$$

It is easy to see that the values of adequacy A are in the interval [-1,1].

The Infos

The resolving of the information activity is ***a goal*** of the information witness.

This goal may be achieved by the establishment and providing (information) contacts and interaction.

The entity, which has possibility for:

- ***(primary) activity*** for external interaction;
- ***information reflection and information memory***, i.e. possibility for collecting the information;
- ***information self-reflection***, i.e. possibility for generating "secondary information";
- ***information expectation*** i.e. the (secondary) information activity for internal or external contact
- ***information modeling and resolving the information expectation***

is called ***Infos***.

The Further Investigation in the Theory of Infos

What gives us the concept “Infos”?

At the fist place, this is the common approach for investigating the natural and artificial information agents.

In other hand, this is the set of common characteristics which are basic for all entities, which we may classify to the category of the Infos.

And, at the end, this is a common philosophical basis for understanding the information subjects.

Our main goal is to provoke the scientists to continue the research in this important area and to make the next step.

The concept **"Infos"** is basic for the General Information Theory [Markov et al. 2003a]. Its definition is only the starting point for further investigations and building the ***Infos Theory***.

The variety of types of Infoses in the real world needs to be investigated and classified in the future research. At the first step, we may propose that may be at least two main types of Infoses exist:

- ***infogens*** - the natural creatures;
- ***infotrons*** - the artificial creatures.

Also, the Infos Theory needs to give answers to many other very important questions, such as:

- What is the nature of the activity of the Infos?
- What is the difference between the living level of the Infos and the non-living one?
- Is it true that the boundary between non-living and living entities is self-reflection and internal activity for satisfying the secondary (information) possibilities for internal or external contact?
- Etc.

It is impossible to answer to all questions in a single article. We may make only the next little step. This is the aim of the present paper.

The concept "Information Model" (IM) is fundamental for the Informatics. There exist many approaches to define this concept. As a rule, every definition is based on those concepts, which the concrete scientific paradigm had given. Every new theoretical approach needs to redefine the concepts it uses in the frame of the corresponded to it new paradigm. This way in different paradigms we may have different definitions of the given concepts [Popper, 1968].

There exists a long list of names of scientists who worked to define more precisely the concept "Model" (and respectively - the information model). It contains the names of N.Wiener and A.Rosenblueth [Rosenblueth, Wiener, 1945], V.M.Glushkov [Glushkov, 1986], M.W.Wartofsky [Wartofsky, 1979] and many others.

For long period, the concept "Information model" has been used to denote one of the main activities in using the computer technique. May be, now it is the most popular understanding of it and many scientists are satisfied of the meaning it contains.

Nevertheless, the definition of the concept of information model may and need to be extended to cover the new scientific paradigms, which come from the current Informatics. This is the goal of the Theory of the Infos.

Presented above so simple and clear definition of the concept "information model" has very great impact on GIT and key role for definition and explanation of all subjective information phenomena in the world. In addition, it may be used as a base concept in the area of Artificial Intelligence research.

The information models initiated inside the Infos form subjective information set. Inside his information set, the Infos can build "information spaces". The information space of the Infos is dynamic as a structure and content. When a new information model is generated the Infos compares it with the information models from context and with the information expectation. This is the starting point for the processes of reasoning.

However, the information modeling is only the first part of the complex process of decision making in usual practical situations. The decision making is at least two-level process:

- Collecting information models for given entity in one resulting entity;
- Analyzing and knowledge discovery on the base of given goal, which results aimed to be used for predicting

of any characteristics of the modeled entity.

In everyday language, the concept "Information modeling" is assumed to denote the whole chain of phases of decision making, which we make to solve the problem [Gladun, 1994].

Theory of Inforaction

Information Objects

When the Infos interact with the entities around in the environment, there exist at least two cases of reverberation:

- the contacts and interaction are casual and all reflections in the Infos as well as in the entities have casual origin;
- the contacts and interactions are determined by the information activity of the Infos.

In the both cases, the contacted entity may reflect any information model from Infos. The possibility for reflection of the information model is greater in the second case.

An entity, in which one or more information models are reflected, is called ***"information object".***

The information objects can have different properties depending on:

- the kind of influence over the entities - by ordering in space and time, by partial or full modifying, etc.,
- the way of influence over the entities - by direct or by indirect influence of the Infos on the object,
- the way of development in time - static or dynamic,

etc.

It clear, the Infoses are information objects.

Information Operations

The information is kind of reflection. Therefore, the only way one to operate with information is to operate with the entity it contains. Every influence on the entity may cause any internal changes in it and this way may change the information already reflected. Another type of influence is to change the location of entity or to provoke any contact between given entity and any other.

The influence over the information object is called "***information operation***".

The information operations may be of two main types:

- the Infos internal operations with the sub-entities that contain information,
- external operations with the information objects that contain information.

Internal Operations

The internal operations with the subentities closely depend of the Infos possibilities for self-reflection and internal interaction of its subentities.

The self-reflection (self-change) of the Infos leads to the creating of new relationships (and corresponding entities) in it. These are subjectively defined relationships, or shortly - ***subjective relationships***. When they are reflected in the memory of the Infos they initiate information model too, but on a higher level. These high-level information models may have not real relationships and real entities that correspond to them.

The possibility for creating the relationships of similarity is a basis for realizing such very high level operations as "comparing elements or substructures of the information models", "searching given substructure or element pattern in the part or in the whole structure of the information model", etc.

It is clear, the Infos is built by entities some of which may be also Infoses, but on the lowest levels. So, the internal operations are determined by the concrete internal level but from the point of view of the higher levels, they are assumed as external operations. Because of this, we will concentrate out attention on the second type of operations.

External Operations

The external operations with information objects may be differed in two main subtypes — basic and service operations.

There are two "***basic information operations***" which are called I-operations:

- I-reflection (reflecting the information object by the Infos, i.e. the origination of a relevant information model in the memory of the Infos).
- I-realisation (creating the information object by the Infos);

In the process of its activity, the Infos reflects (perceives) information from the environment (entities O_i, i=1,2...) by proper subentities (sensitive to video, acoustic, tactile, etc. influences) called "***receptors***" R_i (i=1,2...). Consequently, the Infos may receive some information models. This subjective reflection is called "***I-reflection***".

When necessary, the Infos can materialize (realize) in its environment (entities O'_j, i=1,2...) some of the information models, which are in his memory, using some sub-entities called "***effectors***" M_j (j=1,2...). Consequently, new or modified already existing entities reflect information, relevant to these information models. This subjective realization is called ***"I-realization"***.

There are several operations, which can be realized with the information objects: transfer in space and time, destroying, copying, composition, decomposition, etc. Because of the activity of the Infoses, these operations are different from other events in reality. In this case, the Infos determined operations with information objects are called "***service information operations***".

For example, some of the very high-level service operations are based on the external influence on the information object to change any existing reflection: including and removing an element in and from the structure; copying or moving substructures from one place to an other; building new structure using as a basis one or several others; composing or decomposing of elements or substructures; etc.

Information Processes

Let "O" is a set of real information objects and "M" is a set of information models.

We will consider every set of real information objects as an information object, if the opposite is not stated.

Every set of information models we consider as information model.

The *information operations* are:

- the function **r**: $O \rightarrow M$. (*I-reflection)*

- the function **e**: $M \rightarrow O$. (*I-realization)*

- the function **s**: $O_d \rightarrow O_r$ between two sets of information objects, O_d and O_r may be coincidental (*service operation*).

Let t_1, t_2,..,t_n are information operations. The consequence of information operations P, created using the composition, i.e.

$$P = t_1 \circ t_2 \circ \ldots \circ t_n$$

is called "***information process***".

In particularly an information process can include only one operation.

It is clear, the composition of two or more I-reflections as well as the composition of two or more I-realizations are not allowed.

Information Contact

If an information model from the Infos is reflected in another entity, there exist possibility, during the "a posterior" interactions of the given entity with another Infos, to transfer this reflection in it. This way an information model may be transferred from the Infos to another.

If the second Infos has already established information expectation, the incoming reflection will be perceptible for the Infos. The information expectation will be resolved in some degree and the incoming information model and information in it will be received by the second Infos.

Let S_1 and S_2 are Infoses and O is an arbitrary entity.

The composition of two contacts

$$S_1 \xrightarrow{\Theta^{S_1 O}} O \xrightarrow{\Theta^{O S_2}} S_2$$

is called "***information contact***" between Infos S_1 and Infos S_2 iff during the contacts any information model from S_1 is reflected in the Infos S_2 true the entity O. The Infos S_1 is called "***information donor***", the Infos S_2 is called "***information recipient***", and the entity O is called "***information object***".

In this case, when the donor and the recipient are different Infoses the information contacts between them consist of a composition of at least two information operations - I-realization and I-reflection. For the realization of a direct information contact between two different Infoses is necessary the execution of the composition of these two "basic" operations. All the rest information operations are necessary for supporting the basic ones i.e. they are auxiliary (service) operations.

For the realization of one information contact at least one information object is necessary.

This way the elementary communicative action will be provided.

In general, every information process "k", having as a start domain the set S_d (of information models) and as a final domain the set S_r (again of information models), which may be coincidental, we call "information contact":

$$k: S_d \rightarrow S_r$$

S_d is called "Infos-donor" and S_r - "Infos-recipient".

Information Interaction

The set "R" of all information contacts between two Infoses S_a and S_b

$$R= \{k_i \mid i=1,2..; k_i: S_a \rightarrow S_b\}$$

is called "***information interaction***" or simply "***inforaction***".

When S_a and S_b are coincident, we call it Information interaction with itself (in space and time).

The set "B" of all information objects, used in the information interaction between given Infoses is called "***information base***".

Information Society

The **"Information Group"** (IG) is a set of Infoses, with common Information Base of the information interactions between them.

The **"Information Society"** (IS) is a set of Information Groups, with common Information Base of the information interactions between them.

In the small Information Group the service information operations may be provided by the every Infos when it is necessary.

In the Information Society this is impossible or not optimal. In such case, some Infoses or Information Group became as **"information mediators"** between the others. They start to provide the service information operations.

They realize **"Information Service"**.

The Further Investigation in the Theory of Inforaction

For more than twenty years the Theory of Inforaction has traversed the way from the exotic and unusual concepts such as "information contact" and "information object", presented by the authors in 1983, to actual and world-wide investigated area of informatics.

Nevertheless, there exist many problems for future research.

Note that I-realization is not just reflections of information models in material entity. They include both a reflection of the information models of the Infos and a reflection of the state of the Infos in the moment of I-realization.

This means that here we consider the notion of "information objects" as a more general than the notion of "message".

It is possible that the entity of the information object is not able to keep (save) the whole influence of the I-realization. In other hand, the Infos consciously, by proper actions, restricts the I-realization to reflection of information model only by suppressing the reflection of his condition in the moment of I-realization.

In this case, we are near to the notion of "message" as we use it conventionally.

For example, from the point of view of the notion message, the speech of one politician on the meeting and the same speech printed in the paper are the two equal variants of the message. However, the influence and the result from the perceiving of the speech are different in both cases. In the first one (direct contact) the perceiving one can reflect the condition of the speaker (intonation, pauses, etc.) but in the second case (indirect contact) this is almost impossible. From this point of view, there exists a relation between two different information objects.

The Theory of Inforaction closely depends on the results of information operations. After the execution of some of the information operations, a new information object is possible to be created (for example, after the composition or decomposition). In other cases, the operation may not lead to appearance of new information object but to destroying of a certain existing information object.

The Infos is the only one who can determine whether after the execution of one operation (or a consequence of operations) an information object has appeared. Analogously, the Infos is the witness whether a new information object appears, when in the process of I-realization the Infos acts upon entities, which include some reflection of earlier I-realization. That is why, when there is not exact instruction from the Infos, we suppose that all information operations, with the exception of two - "destroying" and "I-reflection" will produce (one or more) information objects. The operation destroying initiates "empty" information object by destroying the starting one. We suppose that the operation "I-reflection" always initiates information model in the memory of the Infos.

Because of the growing of the communicative aspects of the information service now all over the world the everyday concept is the "Information society".

The growth of the global information society shows that the knowledge turns into important and necessary article of trade. The open environment and the market attitudes of the society lead to arising of the knowledge customers and knowledge sellers, which step-by-step form the "Knowledge Markets". As the other markets, the Knowledge Market is the organized aggregate of participants, which operates in the environment of common rules and principles.

Examination of the market demand for various types of courses and training modules is an essential criterion for effectiveness and high efficiency. Market trends, industry requirements, and companies training needs have to be examined on a regular basis in accordance with the Theory of Information Interaction.

Conclusion

The development of the General Information Theory should not become by the single creative impulse. For a long period, the constructive activity of the many researchers is needed for establishing the new common paradigm.

We all need free scientific look at things, which will permit us to build the general theory without partiality, and aberrations taking in account all information paradigms already created and adopted.

During the years, the investigation in the area of the GIT has showed that the received theoretical results may be used for building the ontology of informatics. Our opinion is that the GIT may be used as main classification scheme. The first step is to describe the main divisions of informatics. The further investigation needs integration with other scientific areas and paradigms.

We have made a little walk toward the establishing the new paradigm. It is synthesized in the table below.

Basic Structure of the General Information Theory

✓ Occurrence	✓ Specificity	✓ Subject	✓ Theory
✓ Reflection	✓ Information Relationship	✓ Evidence	✓ Theory of Information
✓ Activity	✓ Information Expectation	✓ Witness	✓ Theory of Infos
✓ Modeling	✓ Information Modeling	✓ Infos	
✓ Interaction	✓ Information Interaction	✓ Society	✓ Theory of Inforaction

Acknowledgements

This paper is the next step of the process of establishing the GIT as common paradigm. It is based on the ideas considered during very creative discussions at the International Conference "KDS 1997" (September, 1997, Yalta, Ukraine) and at the International Conference "ITA 2000", (September, 2000, Varna, Bulgaria) as well as at the previous scientific meetings organized by the International Workgroup on Data Base Intellectualization (IWDBI). The creative discussion at the KDS 2003 International conference, based on the [Markov et al, 2003a] gives us a great impulse to continue working in this very important scientific area. Authors are very grateful to all participants in the fruitful discussions at KDS and ITA International Conferences and to all members of the International Workgroup on Data Base Intellectualization (IWDBI) and the Association of Developers and Users of Intellectualized Systems (ADUIS) for supporting the advance of the General Information Theory.

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

Bibliography

[Abdeev,1994] R.F.Abdeev. The Philosophy of the Information Civilization. Moscow, VLADOS, 1994. (in Russian)

[Bangov, 1995] I.Bangov. A Graph-Topological Model of Information. Int. J. Information Theories and Applications, 1995, v.3, n.6, pp.3-9.

[Burgin, 1997] M.S.Burgin. General Information Theory. http://www.math.ucla.edu/~mburgin/res/compsc/Site3GTI.htm

[Burgin, Markov, 1991] M.Burgin, Kr.Markov. Formal Definition of the Concept Materialization. Mathematics and Mathematical Education, BAS, Sofia, 1991, pp.175-179.

[Cooman et al, 1995] G.de Cooman, D.Ruan, E.Kerre, Eds. Foundations and Applications of Possibility Theory. World Scientific, Singapore, 1995.

[Elstner, 1993] D.Elstner. About Duality of the Information and Organization. Int. J. Information Theories and Applications, 1993, v.1, n.1, pp. 3-5. (in Russian)

[Gladun, 1994] V.P. Gladun. Processes of New Knowledge Formation. Pedagog 6, Sofia, 1994 (in Russian).

[Glushkov, 1986] V.M.Glushkov. Epistemological Nature of the Information Modeling. In: V.M.Glushkov. Cybernetics, Questions and Answers, (in Russian), Moscow, Science, 1986, pp. 33-41.

[Lu, 1999] C.-G.Lu. A Generalisation of Shannon's Information Theory. Int. J. of General Systems, 28:(6), 1999, pp.453-490.

[Markov, 1984] Kr. Markov. A Multi-domain Access Method. Proc. of Int. Conf. "Computer Based Scientific Research". Plovdiv, 1984. pp. 558-563.

[Markov, 1988] Kr.Markov. From the Past to the Future of 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, 2003a] Kr.Markov, Kr.Ivanova, I.Mitov. General Information Theory. Basic Formulations. FOI-Commerce, Sofia, 2003.

[Markov et al, 2003b] K. Markov, K. Ivanova, I. Mitov, E. Velikova-Bandova. *The Information*. IJ Information Theories and Applications, FOI ITHEA, Sofia, 2003, Vol.10, No.1, pp.5-9.

[Markov et al, 2004] K. Markov, K. Ivanova, I. Mitov, E. Velikova-Bandova. *Formal Definition of the Concept "INFOS"*. Proceedings of the Second International Conference "Information Research, Applications and Education" i.TECH 2004, Varna, Bulgaria. Sofia, FOI-Commerce, 2004, pp. 71-74. Int. Journal "Information Theories and Applications", 2004, Vol.11, No.1, pp.16-19

[Markov P., 2002] P.Markov. Think with Your Mind. Markov College. Sofia, 2002. (in Bulgarian)

[Pavlov, 1987] T.Pavlov. Collection of Selected Works, Vol. II. Theory of Reverberation. Science and Art, Sofia, 1987. (in Bulgarian).

[Rosenblueth, Wiener, 1945] A.Rosenblueth, N.Wiener. Role of Models in Science. Philosophy of Science, Vol.12, No.4, 1945.

[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.

[Tomov, 1991] K.Tomov. The Resomal-Isomorphic Principle. Arges, Sofia, 1991. (in Bulgarian)

[Wartofsky, 1979] M.W.Wartofsky. Models. Representation and the Scientific Understanding. D.Reidel Publishing Company, Dordrecht: Holland /Boston: USA, London: England/, 1979.

[Popper, 1968] K.R.Popper. Conjectures and Refutations: The Growth of Scientific Knowledge. Harper & Row, Publishers, New York and Evanston. 1968.

Authors' Information

Krassimir Markov - Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; Institute of Information Theories and Applications FOI ITHEA, 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

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

TECHNOLOGY FOR ONTOLOGICAL ENGINEERING LIFECYCLE SUPPORT [1]

Vladimir Gorovoy, Tatiana Gavrilova

Annotation*: Presented paper describes software system project ONTOLINGE-KAON that provides technological support for the whole lifecycle of ontological engineering. The main stress is put on the evaluation of maturity and quality of ontologies and on the usage of ontologies with the help of automated generation of knowledge portals, based on ontologies. Possibility of creation of knowledge portals built on top of ontologies can become a big step forward in the field of e-learning. The paper presents advantages provided by knowledge portals based on top on ontologies.*

Keywords*: ontological engineering, knowledge engineering.*

ACM Classification Keywords*: H.0 Information systems – General, I.2.6 Artificial intelligence - Learning*

Introduction

ONTOLONGE-KAON is aiming at providing technological support for the full lifecycle of ontological engineering. At present a great number of components and software products are implementing various tasks in work with

[1] The research is partially supported by Russian Foundation for Basic Research (grants **06-01-81005 and 07-01- 00053)**

ontologies. According to this fact an urgent problem of integration of these components in a common system supporting all stages from creation of an ontology, evaluation of its maturity and quality up to using of it by target users emerges.

Evaluation of quality of developed ontologies and support of the phase of ontology usage is a bottleneck of present instruments providing technological support for working with ontologies. And this is regardless of the fact, that today we can state a broad interest to ontological engineering. Thus there are more than 50 ontological editors:

- Protégé,
- GALEN Case Environment (GCE),
- ICOM
- Integrated Ontology Development Environment
- IsaViz
- JOE
- KAON (including OIModeller)
- KBE -- Knowledge Base Editor (for Zeus AgentBuilding Toolkit)
- LegendBurster Ontology Editor
- LinKFactory Workbench
- Medius Visual Ontology Modeler
- NeoClassic
- OilEd
- OLR3 Schema Editor
- OntoBuilder
- And others.

But all of them don't support development of corporate knowledge portals. The first step in this direction was done in the KAON project (http://kaon.semanticweb.org) [Motik et al., 2002]. One of the developed subsystems (KAON Portal) allowed generating portal providing web-interface for browsing of ontologies created by OI-Modeler. One of the KAON Portal's shortcomings is absence of any information bound to ontology (moreover, binding of information is not supported). Besides KAON Portal generates a portal on the basis of ontologies, kept in an internal format (private expansion of RDFS), not being a conventional standard. Thus, ontologies used for generation of a portal should be created only in KAON OI-Modeler what is certainly an essential restriction.

One more example of the system focused on creation of a portal is PORTO [Gavrilova et al., 2003]. Work in PORTO system consists of the following stages:

1. Creation of a portal ontology by an analyst by means of the visual editor.
2. Creation of design of a portal and binding ontology concepts to representation by the web-designer.
3. Online generation of pages of a portal by server PORTO in reply to the user queries.

One of the disadvantages of PORTO is that it lacks integration with other systems for ontology development. Thus ontology for portal generation can be created only in the offered visual editor that kept its data in own internal format.

Automated generation of knowledge portals on the top of ontologies would become a great step forward in the field of e-learning. For example, automatic creation of a portal the course would be invaluable to students and would release teachers from a part of work for creation of the portal during development of an ontology for a training course or a topic of one lecture.

The fact of portal being built on the top of ontology can help users during its use. In this case it is possible to enhance standard text search on a portal with the reasoning system being able to run queries on ontology. As the interface for the advanced users it is possible to use the form for input of RQL-inquiries [Karvounarakis et al.,

2003] in this case. Executing of such queries will release users from a filtration of superfluous information that he often receives as a result of standard text queries.

System components and their interaction

Components of the system are focused on solving of the following problems:

1. Ontologies producing
2. Ontologies evaluation
3. Ontologies usage

For ontologies producing it is possible to use any ontology editor or some other instrument for ontologies creation allowing saving ontologies in OWL format (http://www.w3.org/TR/2003/CR-owl-features-20030818). Among ontological editors it is worth mentioning Protege [Noy et al., 2001] or SWOOP [Kalyanpur et al., 2004]. For creation of OWL ontologies programmatically Jena (HP labs - http://www.hpl.hp.com/semweb/downloads.htm), KAON2 (http://kaon2.semanticweb.org), IODT (IBM Integrated Ontology Development Toolkit - http://www.alphaworks.ibm.com/tech/semanticstk) or OWL-API (http://owl.man.ac.uk/api.shtml) can be used. These APIs considerably facilitate life for software developers who implement import of their internal ontologies to OWL format. Thus we get an ontology in OWL format as an output of the first component.

Important part of ONTLINGE-KAON system is an evaluation module for created ontology and providing recommendations for its improvement. One can use existing tools for ontologies evaluation (OntoAnalyser, KAON2) as such a component. Unfortunately these instruments for evaluation haven't reached serious level of maturity and are not widely used. This is caused by the fact that recommendations of knowledge engineers in the domain of ontologies forming can hardly be formalized and implemented in software. As a result of this, problem of building new instruments for ontologies meeting needs of the majority of users is topical. Some ideas regarding building of such a module are presented below in the section devoted to evaluation component.

There are the following requirements for evaluation component in ONTOLINGE-KAON system:

1. Ability to work with ontologies in OWL format.

2. Inference of remarks concerning ontology quality and providing suggestions for quality improvement.

Corporate and educational portals, generated on the basis of ontologies, formed by other system components, are targeted at implementation of ontology using. We formulated the following requirements for a portal:

- Generation on top of ontology in OWL format
- Access to created ontology
- Adding relevant information to concept instances for displaying on the page generated by the portal
- Ordinary portal search
- Portal search using that portal is built on the top of ontology

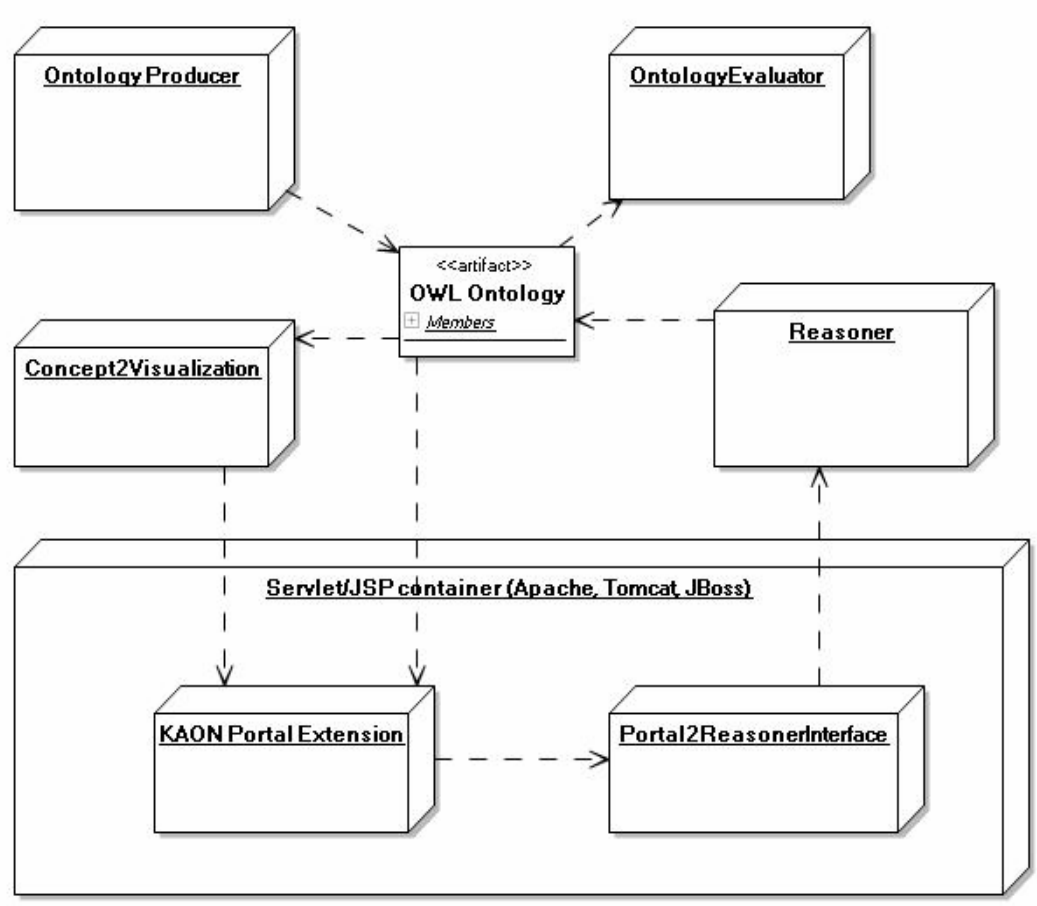
Architecture of the system solving all stated problems is presented on picture 1. Its main components are:

- OntologyProducer – component of ontology forming
- OntologyEvaluator – component for evaluation of quality and maturity of ontologies
- KAON Portal Extension – extension of KAON Portal'a. Mission of this component includes generation of a portal built on ontologies
- Concept2Visualization – module providing possibility of matching presentation with concepts of the ontology (for exampled concept "project OntoWeb" is matched to some visualization on the generated page which contains all necessary information concerning the project)
- Portal2ReasonerInterface – module providing interface for search queries bound to portal ontology

- Reasoner - module for reasoning (for example Pellet OWL Reasoner).
- Servlet/JSP container – server supporting Servlets' and JSPs'. On such a server KAON Portal Extension module can work (for example Apache Tomcat or JBoss).

Thus, in created architecture we develop the following components and integrate them with each other and other components of ONTOLINGE-KAON system:

- KAON Portal Extension
- Concept2Visualization
- Portal2ReasonerInterface
- OntologyEvaluator



Pic. 1. ONTOLINGE-KAON architecture

Component of ontologies evaluation – OntologyEvaluator

Existing instruments for ontologies evaluation (OntoAnalyser, KAON2) can be used as a component of ontologies evaluation. Unfortunately existing evaluation tools haven't reached serious level of maturity and are not widely used. Suggested solution is supposed to eliminate shortcomings peculiar to existing evaluation tools.

Analysis of student works on creating ontologies in simple and well-known domains revealed several primary factors that distinguished good ontologies from bad ones. These laws can be reformulated and made applicable for a practicing knowledge engineer. The main hypothesis can be stated as: "Harmony = conceptual balance + clarity".

At that conceptual balance means that:

- Concepts of the same level are connected with parent concept with the same type of relations (e.g. “class-subclass”, “whole-part”).
- Depth of ontological tree branches should be about the same (±2).
- The whole picture should be pretty symmetric.
- Cross links should be avoided as far as possible

Clarity includes:

- Minimization: Thus maximum number of concepts of the same level or branch depth shouldn’t exceed famous Ingve-Miller number (7±2) [Miller, 1956].
- Clarity for reading. Relations’ type should be as obvious as possible, not to overload ontology scheme with unnecessary information and skip names of relations.

All these laws correspond with some results of Gestalt psychology formulated by Maks Wertheimer as early as 1944 [Wertheimer, 1944]. Thus the main principle of good gestalt (good shape) or Pragnanz law was formulated in such a way:

“Organization of any structure in nature or cognition will be as good (regular, complete, balanced, or symmetrical) as the prevailing conditions allow “

The major part of enumerated factors may be formalized and their verification may be implemented in OntologyEvaluator. Thus using of this component can contribute to create harmonic ontologies.

Ontology using

On this phase our solution suggests generation of knowledge portals based on built ontologies and further using of a portal as an ontology navigation tool, for necessary information search and for querying ontology.

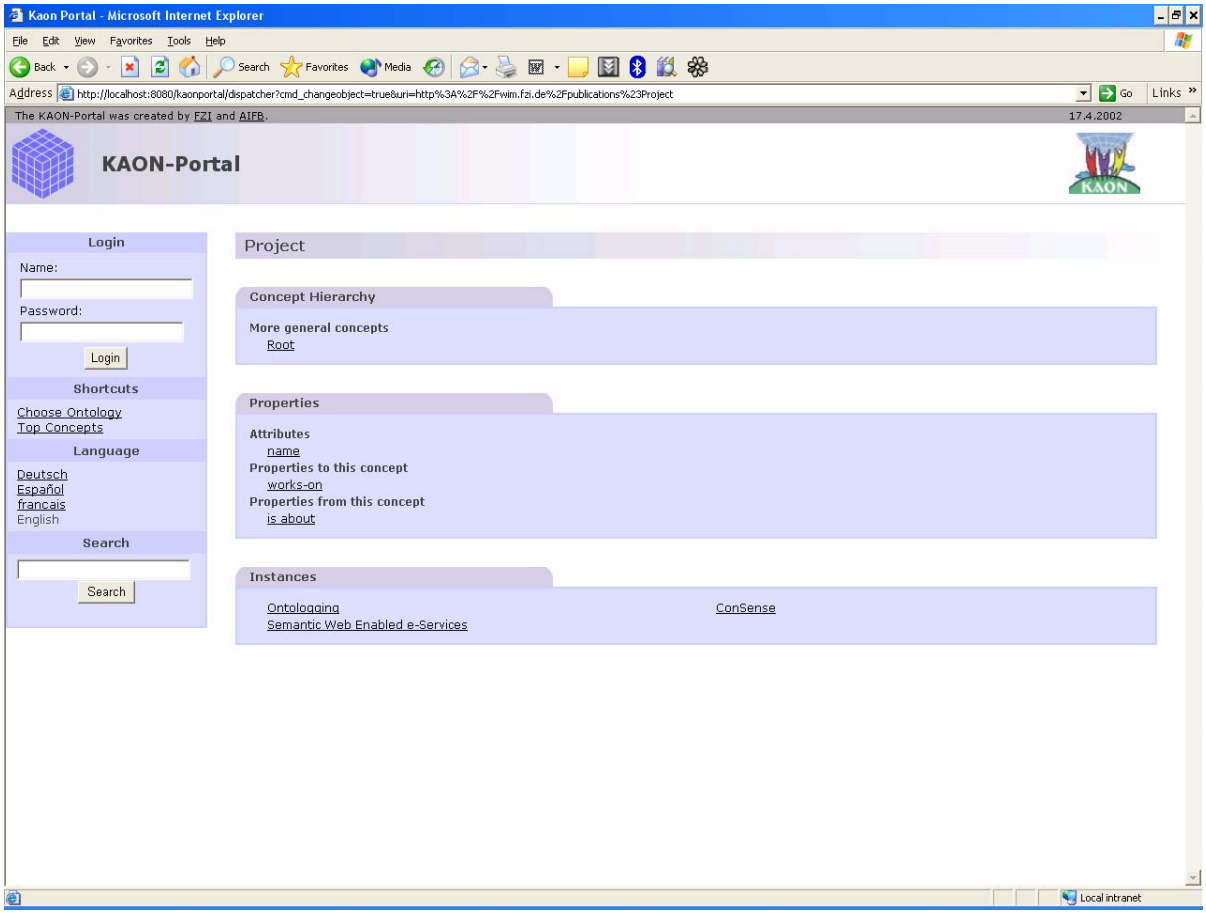
The components below are related to the ontology using:

- KAON Portal Extension
- Concept2Visualization
- Portal2ReasonerInterface
- OntologyEvaluator
- Reasoner
- Servlet/JSP container.

We can use KAON Portal module developed in KAON project as a base prototype for KAON Portal Extension. KAON Portal allows generating portal for navigation of ontology, described on proprietary extension of RDFS (ontology in this format can be produced by OI-Modeler). Screenshot devoted to project concept is presented on pic. 2.

For practicable using of the generated portal as an educational knowledge portal it is necessary to implement in KAON Portal Extension the following features:

- OWL support, because ability to work with conventional standard in the field of ontologies description would contribute better interoperability and integration with other instruments of ontological engineering.
- Ability to bind presentations with ontology concepts. Without this functionality it is impossible to create a usable knowledge portal. This feature is implemented by Concept2Visualization module.
- Interface for search queries of a portal ontology. In the simplest case it can be ability to input RQL-queries providing results on them. Portal2ReasonerInterface implements this functionality.



Pic. 2: Concept of project

Conclusion

Having all the features described above ONTOLINGE-KAON can become a big step forward on the way to using technologies and methodologies of ontological engineering for creating educational knowledge portals. Suggested approach for evaluation of quality and maturity of ontologies seems to be interesting. It can assist in creating high quality harmonic ontologies. Compared to existing systems of ontological engineering new is ability for dynamic generation of portal based on created ontology and containing possibilities for ontology navigation and information constituent related to ontology instances. Proposed solution is flexible and allows to automate reflection of changes made in ontology on generated pages. New is also a feature of portal search leveraging that portal is based on ontology. This functionality radically differs from ordinary search with the help of search systems because it provides only semantically correct results and saves user from the necessity of choosing from many variants much of which are very far related to expected results.

Bibliography

[Gavrilova et al., 2003] T.A. Gavrilova , V. A. Gorovoy. Ontological Engineering for Corporate Knowledge Portal Design // In "Processes and Foundations for Virtual Organizations", Eds. L. Camarinha -Matos and H. Afsarmanesh, Kluwer Academic Publishers, 2003. - p.289-296.

[Kalyanpur et al., 2004] A. Kalyanpur, E. Sirin, B. Parsia, J. Hendler. Hypermedia inspired ontology engineering environment: Swoop. // In Proceedings of 3rd International Semantic Web Conference (ISWC-2004), Japan (Poster).

[Karvounarakis et al., 2003] G. Karvounarakis, A. Magkanaraki, S. Alexaki, V. Christophides, D. Plexousakis, M. Scholl, K. Tolle. Querying the Semantic Web with RQL. // In Computer Networks and ISDN Systems Journal, Vol. 42(5), August 2003, pp. 617-640.

[Miller, 1956] G.A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. // In The Psychological Review, 1956, vol. 63, pp. 81-97

[Motik et al., 2002] Boris Motik, Alexander Maedche, Raphael Volz. A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications. // In Proceedings of the First International Conference on Ontologies, Databases and Application of Semantics (ODBASE-2002). Springer, 2002.

[Noy et al., 2001] N.F. Noy, M. Sintek, S. Decker, M. Crubezy, R.W. Fergerson, M.A. Musen. Creating Semantic Web Contents with Protege-2000. // IEEE Intelligent Systems 16(2), pp. 60-71, 2001

[Wertheimer, 1944] M. Wertheimer. Gestalt theory. // In Social Research, 11, 78-99.

Authors' Information

Vladimir Gorovoy – PhD student**,** Saint-Petersburg State Polytechnic University, Intelligent Computer Technologies Dpt. 195251, Politechnicheskaya 29/9, St. Petersburg, Russia; e-mail: vgorovoy@mail.ru

Tatiana Gavrilova – Professor**,** Saint-Petersburg State Polytechnic University, Intelligent Computer Technologies Dpt. 195251, Politechnicheskaya 29/9, St. Petersburg, Russia; e-mail: gavr_csa@rambler.ru

INTELLIGENT SEARCH AND AUTOMATIC DOCUMENT CLASSIFICATION AND CATALOGING BASED ON ONTOLOGY APPROACH

Vyacheslav Lanin, Lyudmila Lyadova

Abstract*: This paper presents an approach to development of intelligent search system and automatic document classification and cataloging tools for CASE-system based on metadata. The described method uses advantages of ontology approach and traditional approach based on keywords. The method has powerful intelligent means and it can be integrated with existing document search systems.*

Keywords*: electronic document, automatic document classification and cataloging, ontology approach, information system development.*

ACM Classification Keywords*: I.2.7 Artificial Intelligence: Natural Language Processing – Text analysis; D.2.2 Software Engineering: Design Tools and Techniques – Computer-aided software engineering (CASE).*

Introduction

Development tools used for implementing large distributed information systems, which consist of separated subsystems and should be installed in territorially remote organizations, should meet the requirements, providing possibility of its customization on various maintenance conditions and user's requirement during installation and dynamically during maintenance. Organizations has various technical possibilities, organizational and business forms, it makes information system development difficult. Implementation of these requirements provides efficiency of expenses for system creation, a high degree of its adaptability and scalability, robustness of the system.

The CASE-system METAS bases on interpretation of the multilevel metadata. The metadata describes an information system from different points of view and with a various grain size. Opportunities of dynamic system customization are provided by re-structuring of a database, generation and customization of user interface, generation of queries and creation of reports [Lyadova, 2003].

The data domain analysis is the most labour-intensive and important stage in process of information system development by means of CASE-system. Any changes in business operations of organization, for which information system is created, demand the iterated analysis and modification in information system model. Often changes of system maintenance conditions or changes of user's requirements are connected with some normative documents. It can be the normative documents determining business processes of the data domain or the documents of the particular organization. Thus, the analysis of data domain in many respects bases on the analysis of documents, which constitute difficult system. Modifications in model should be grounded on the changes fixed in normative documents.

Complexity of analyst work can be lowered by automation of document analysis process. To solve this task it is necessary to have tools intended for search and keeping document set. Documents can be received from different sources, connected with considerated data domain. Except this for automation of analysis process tools of classification, cataloguing and data mining should be included to the system.

In this paper the problems connected with information processing in an inhomogeneous program and organizational environment, particularly with documents search and their electronic cataloguing, are considered. As an example of such documents we can mention various internal organization documents (orders, contracts, acts and so forth), normative - legal documents, etc. Documents come to a system in a random order from different sources. It is usually semistructured. It reasons the complexity of document processing. The extremely important tasks for implementation in this area are automation of processes of data exchange with various legal informational systems, and the possibility of import texts and documents from files and databases of various formats and documents management systems.

The main problems which prevent fast and high-quality document processing in electronic documents management systems are insufficient structuredness of the information, information redundancy, and presence of great deal of undesirable for user information. The human factor has a significant impact on the efficiency of document search. An average user is not aware of the advance option of a query language and uses simply typical queries.

Development of a specialized software toolkit intended for information systems and electronic document management systems can be effective solution of tasks listed above. Such toolkit should be based on the means and methods of AI.

Problem of document search

Situation when a user searches something in a book or printed document is considered below. The most obvious way is to read the whole book (document), but such process takes a long time. However if a user knows something connected with the data domain, he can use book (document) content and choose appropriating part containing necessary information. Also he can look throw subject index to define page numbers where searched terms are mentioned.

In the example content and index are tools which make search process easier. In case of information system including document management tools documents play a part of searched information and services called subject directory are used as content and subject index. Here under document management tools we mean advanced tool, which functions are not only creating and keeping documents, it also allows to search, import, analyze, categorize and catalogue documents.

For example user operates with a computer system and he needs to gather information about Perm city. Some search stages can be marked out. Necessary to find something appears, in other words information need appears. Then user has to formalize his information need somehow. In traditional systems formalization comes to choosing concepts and key words and their relations. Chosen set of key words with fixed relations between it is called query. Next user enters the query by means of search system interface. The system extracts documents which match user's query from document set called information search field according to customizable search criteria and then forms a result. The documents being found divide into two groups according to its content (Fig. 1): documents matching user's needs and documents, which do not match user's needs but match user's query (information noise). In the example documents where Perm is not a city name are information noise.

Measure of correspondence between system response and user information needs is called semantic relevance, and measure of correspondence between system response and user query is called formal relevance. Usually presence of query keywords in the text of a document is a criterion of document formal relevance. If we use search based on keywords usually some of the documents matching user's information needs do not get into result set. For example if keyword is "Perm" documents where it is used in phrase "Perm region capital " instead of "Perm city" may not be found.

The main problem of information search is the result of the fact, that the majority of information search systems are based on using keywords and "word" doesn't have meaning and semantics.

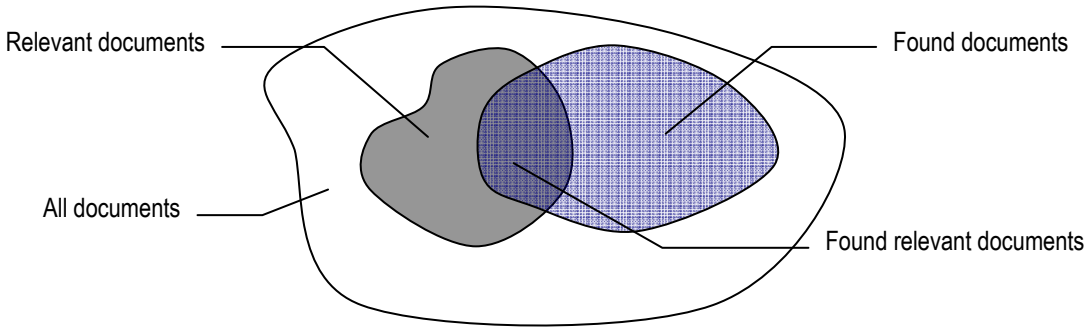


Fig. 1. Document space

The majority of technologies for document processing are oriented on handy work with information. Often principles of processing electronic information just copy principles of processing printed information. There are various means for text formatting, that help to represent information in a convenient form, in a text editor, but there are not means for semantics reflection in it. In most cases computer is used as typewriter or calculator, which goal is automatic examination of answer alternatives. For effective search it is necessary to extend traditional document concept: document should be related with knowledge, which make possible document content interpretation and processing.

Usually artificial intelligence methods are used for solving tasks, which are easy and obvious for people, but it is difficult to formalize them and implement their algorithms. One of these tasks is working with documents in information systems. It includes information search, cataloging, analysis and data mining.

There are different methods, models and languages oriented on integrated data and knowledge declaration. The most perspective and universal approach is using ontologies.

Ontology definition

The concept ontology is one of the most used concepts. The term ontology is used in different contexts, and different meanings are ascribed to it. According to implementing tasks we let that ontology is an exact specification of a data domain, which includes terms dictionary and a set of relations between them (like "instance of", "whole-part"). Relations between terms shows how this concepts are correlated with each other in particular data domain. In fact such ontology definition means that ontology is hierarchical entity base of current data domain for which information system is developed.

It is difficult to find appropriate ontology, this process takes much time. So sometimes it is impossible to find ontology among developed ones, that is why a new ontology creating may be defensible. Then ontology takes into account particular task specific. Except this using developed ontologies has some disadvantages more. In particular knowledge of different people can be represented by different ontologies. At the same time we can not state that one ontology is better than another. Some different ontologies representing various aspects of data domain and solving tasks can be developed for he same organization, in which information system is installed.

A lot of languages and systems for declarating ontologies and operating with it exist. The most perspective is visual method, which allows experts to draw ontologies evidently. This helps to formulate and explain appearance nature and structure. Visual (graph) models have especial cognitive force.

Search of documents based on ontologies

According to the approach [Chuprina, 2004] information search is carried out using an ontology either representing data domain of information system or specially developed by user. Generally document content interpretation is extremely difficult task, but document and ontology matching mechanism is necessary only for intelligent document search.

Document search process based on ontology is described below.

The process is started with search of basic ontology concepts in the document. If all concepts have been found in the document we make decision that the ontology describes the document. In case a concept has not been found the system begins to search its synonyms.

If synonyms of searched concept have not been found the system tries to gather concept by parts according to relation called “part of”. If we do not get a result even after this operation the system can use relation “class-subclass”. It allows to take into account stricter or more general concept.

So, recursive ontology search mechanism has been posed. In contrast of traditional search system method mentioned above has more powerful semantics. It allows to find concepts implicitly contained in the document.

The main advantages of ontology search are

- systematic *viewpoint (ontology represents entire data domain);*
- *uniformity (knowledge is presented in standard form);*
- *completeness (*ontology allows to reconstruct not mentioned relations).

Documents found in outer sources can be imported to information system for classification and cataloging, analysis and extracting useful information.

Ontological document classification and cataloging

To organize document cataloging process user have to correlate each document category with an ontology.

When a new document gets to the system it is sequentially compared with ontology of each category. If comparison is successful the document falls within this category. Each document may match several ontologies, so it can be attributed to several categories.

ORDER # 1

Use since 01.12.2005 following list of members of staff:

Assistant	4 000
Associate professor	10 000
Head of	15 000

Head of a university J. Green

01.11.2005

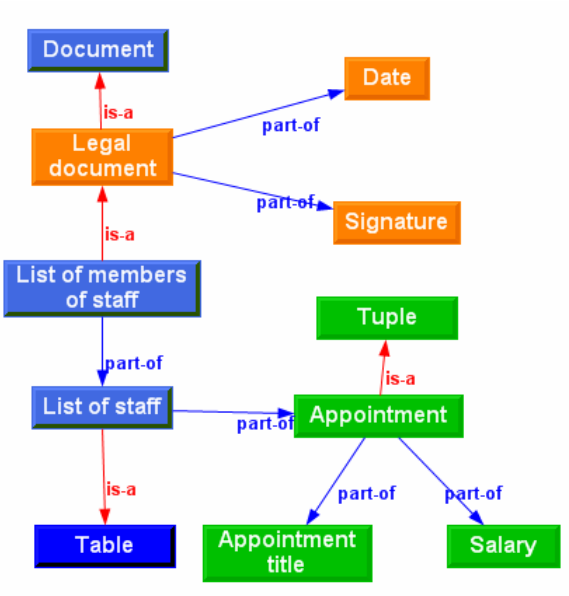


Fig.2. Documents and its ontology

It is convenient to represent a category system as a tree (Fig. 2). Therefore corresponding ontologies constitute an hierarchy. Within such approach to ontology representation child nodes qualify parent nodes ontologies.

For example top level nodes can match small ontologies which represent administrative documents, contracts etc., and nodes of other levels can match ontologies which specify these document types.

Possibility to operate with various document formats is a feature of the approach.

It is necessary to interpret terms describing format and structure of particular document to match ontologies and semistructured documents). Such terms are “table”, ”tuple”, ”date”, ”number” and so on. So special components which provide unificated access to documents with different format should be included to system architecture. Such functionality can be realized by using “installable format driver”. It is components implementing predetermined interface, which allows access to document of a specific type. Implementation of such driver can be based on using patterns, samples, which make possible document structure recognition.

Conclusion

The features of the developed approach are universality of its using, capability to integrate with existing document search systems, powerful intelligent capabilities. Mechanism of ontological clusterization together with functionality of particular information system makes possible effective document management both for documents which are generated in the system and documents which are imported from outer heterogeneous sources.

CASE-technology METAS developed by “Computing institute” makes possible effective using of described tools both during information system development process and when users operate with information system. Also these tools can be used in process of document interpretation to adapt a system dynamically.

The technology gives to user not only customization tools. It also provides components which are used to navigate through information objects, representing entities of data domain, and its relations. Object explorer displays object tree, which can be customized by users due to their information necessaries. Object tree is used not only for object preview, appropriate business operations can be run from the node too. Each document created in the system is represented in the database as an entity. So it is displayed in a tree node, which corresponds to the entity, and user can see it. A document can be reached in tree-walk by different ways according to user’s tasks.

User can include to the tree nodes intended for document clusterization. Such node can be associated with an ontology. So the same mechanism is used for working with documents imported from outer sources and clusterizated in the system.

Adding to CASE-system tools for analysis of semistructured documents, which are categorized on the base of ontologies created by developers and users, essentially reduce labor intensiveness of maintenance and customization.

Bibliography

[Лядова, 2003] Л.Н. Лядова, С.А. Рыжков. CASE-технология METAS. В кн.: Математика программных систем. Пермский государственный университет, Пермь, 2003. С. 4-18.

[Chuprina, 2004] S. Chuprina, V. Lanin, D. Borisova, S. Khaeva. Internet Intelligent Search System SmartFinder. In: Proc. of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology. Knowledge-Based Media Analysis for Self-Adaptive and Agile Multimedia Technology / The Royal Statistical Society, November 25-26, 2004, London, U.K. P. 151-156.

Authors' Information

Vyacheslav Lanin – Perm state university, student of computer since department; Russia, 614990, Perm city, 15, Bukireva st.; e-mail: lanin@perm.ru

Ludmila Lyadova – Institute of Computing, Deputy Director; 19/2-38, Podlesnaya St., Perm, Russia; e-mail: e-mail: lnlyadova@mail.ru

ONTOLOGICAL MULTILEVEL MODELING LANGUAGE

Sergey Shavrin

***Abstract**: This paper presents ontological multilevel modeling language O_2ML, aimed at using with metadata driven information systems. The first part of this paper briefly surveys existing modeling languages and approaches, while the last part proposes a new language to combine their benefits.*

***Keywords**: Metamodeling, information systems, modeling languages.*

***ACM Classification Keywords**: H.0 Information Systems - General.*

Introduction

Information systems development comprises a diversity of artifacts creation, e.g. domain model, users guide, code, set of tests, etc. Short term company productivity depends on availability of tools that can ease or automate the process of artifacts creation and usage. However, medium and long term productivity in many respects depends on universality of these artifacts.

Rising of abstraction level is a common way of universalization and therefore a way of artifacts life prolonging. However, abstracting increases semantic gap between an artifact and a machine, thus leading to translation necessity. As is well known there are two types of translators – compilers and interpreters. Overwhelming majority of contemporary CASE-tools utilize compiler approach. Benefits are obvious: translation process executes once, before system exploitation, thus saving target machine resources. On the other hand, interpreter-based systems exhibit great flexibility. The last property appears to be more valuable in the modern circumstances.

Given the interpreter-based information system, domain model is the natural candidate for the role of "control program". In this case system must have capabilities to understand and execute models described in some modeling language. The most widespread modeling language nowadays is UML [7]. At the moment, OMG (Object Management Group) is working on the second version of the language and concomitant standards. Not all specifications had been published yet, but we can already say that a huge amount of work had been done to formalize UML semantics. Completely formalized semantics sets the stage for building unified UML virtual machine. Figure 1 shows a UML-model example.

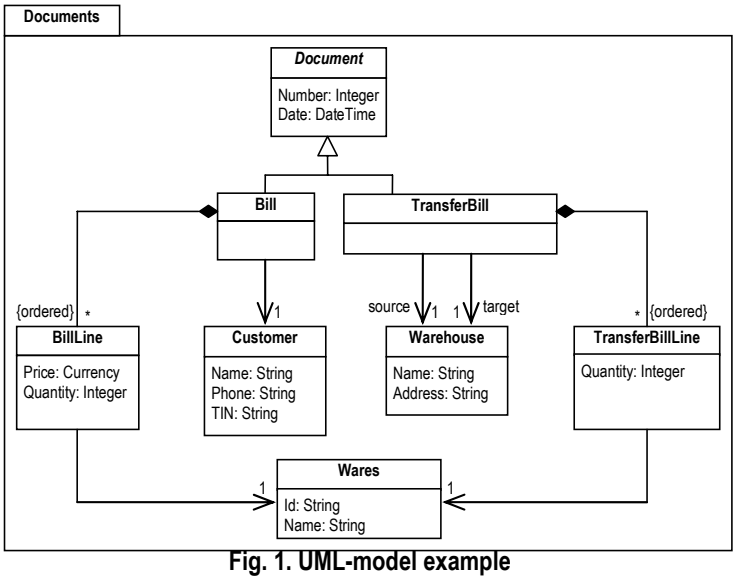


Fig. 1. UML-model example

Given an appropriate tool support, exploiting domain model as a primary artifact significantly increases short term company productivity. However, domain model is prone to become out of date. On the other hand, adjacent domains can be described using similar models differing in details. In this case company can increase its medium and long term productivity by exploiting metamodels which describe more stable metaaspects, common to a set of domains.

UML offers far from complete metamodeling capabilities. These capabilities include stereotypes and tagged values. Powerful metamodeling language has to be able to operate with full-value metaentities at arbitrary number of metalevels.

OMEGA Project

OMEGA [4] – Ontological Metamodeling Extension for Generative Architectures – is a MOF [6] (Meta Object Facility – UML metamodel) extension that introduces ontological metamodeling. OMEGA is aimed at code generation.

OMEGA project introduces a series of notions that enable full-value ontological metamodeling. These notions include metaclasses, metaattributes and metaassociations. It is essential that metaattribute in this case isn't just a metaclass attribute, but a full-value metaentity. An instance of metaattribute is a conventional attribute. This allows one to model such domain features as "Document of each type has exactly one numeric attribute (document number), not less than one date attributes and several property attributes" (Fig. 2 and Fig. 3).

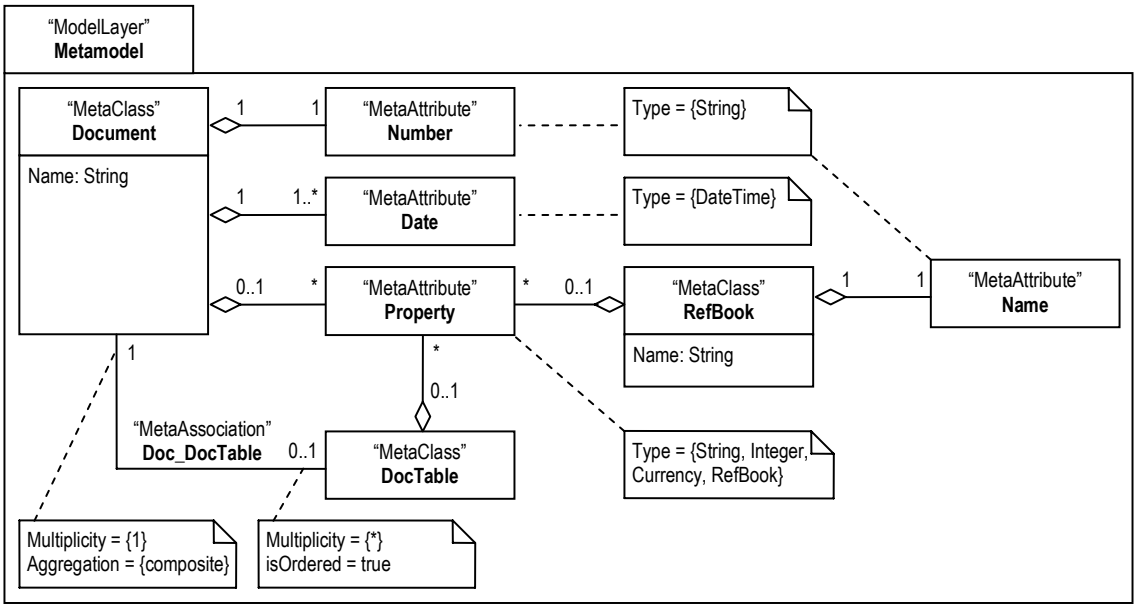


Fig. 2. OMEGA-metamodel example

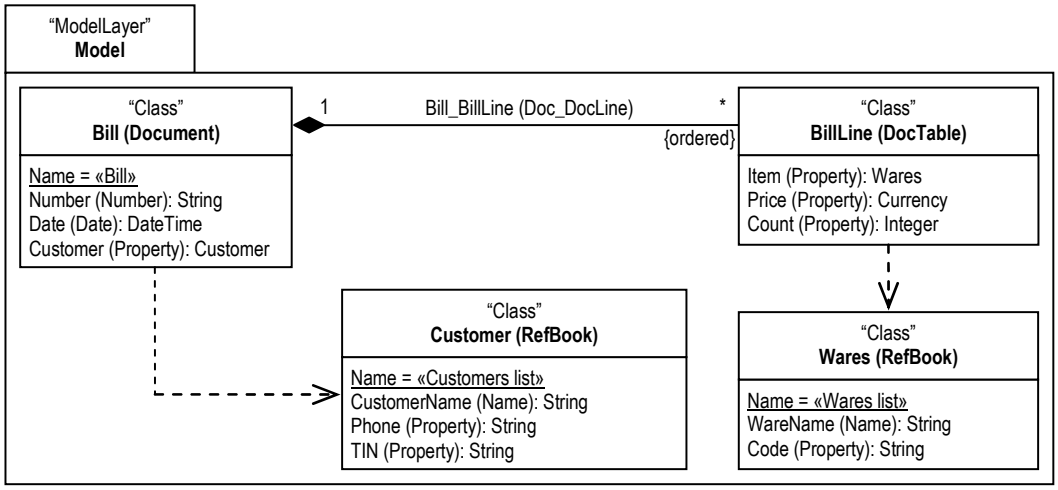


Fig. 3. OMEGA-model example

However OMEGA has two disadvantages. First of all, OMEGA is based on MOF and therefore inherits all its features. In particular MOF is aimed at describing languages like UML and CWM [5] doesn't have some capabilities that are useful in information systems' domain modeling. Namely MOF (and therefore OMEGA) doesn't support plural multidimensional classification – a very convenient modeling tool in author's opinion.

The second disadvantage concerns OMEGA semantics - its description is mainly informal. This fact complicates OMEGA virtual machine creation.

Deep Instantiation

Speaking about instantiation one usually have in mind shallow instantiation. This implies that an instance is created in accordance with its class definition. In other words defining a class we make assertions about its immediate instances. Obviously, it is the only possible interpretation of instantiation in two-level "class-instance" model. However, exploitation of this notion in multi-level case can result in a series of problems. In particular, ambiguous classification and replication of concepts arise [1, 2].

In order to solve shallow instantiation problems Atkinson and Kühne proposed to use a new notion of deep instantiation [2]. This notion allows one to make assertions not only about immediate instances, but instances of instances and so on. This capability is gained by introduction of potency notion – a number that defines allowed instantiations quantity. For example, an instance of class with potency 2 (metaclass) is a class with potency 1 (ordinary class). And an instance of class with potency 1 is a class with potency 0 (object). Similarly, an attribute with potency 2 becomes an attribute with potency 1 (ordinary attribute) that, in its turn, becomes an attribute with potency 0 (slot). Figure 4 shows an example of potency exploitation.

Aside from potency, Atkinson and Kühne introduced a *dual field* notion – an object possessing attribute and slot semantics [2]. In terms of potencies dual field is a slot with non-zero potency. Figure 4 shows some dual field examples, namely "EntityName" and "Description".

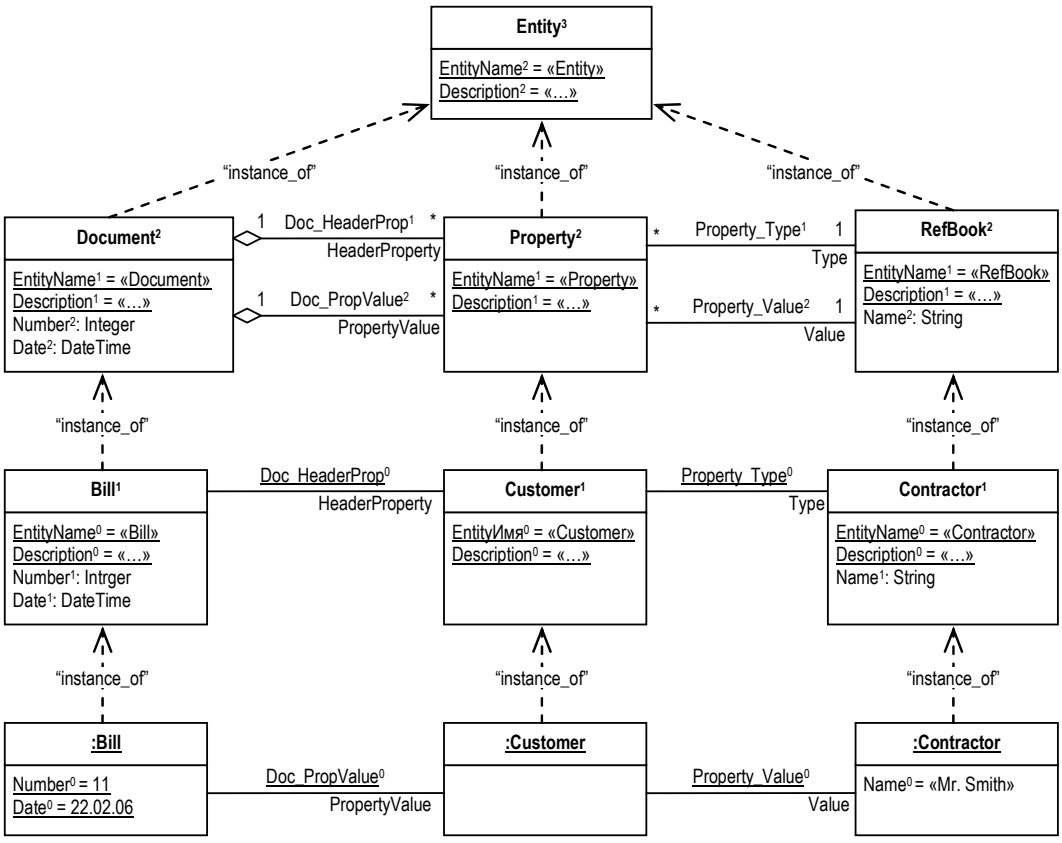


Fig. 4. Potency usage example

In spite of the fact that potency allows to avoid multilevel modeling problems mentioned above, it is obviously insufficient to solve real-world problems. The language with potency support proposed in [2] is too simple to be used in practice. There is a need in additional metamodeling tools like metaattributes and metaassociations.

O_2ML

This part proposes a new modeling language – O_2ML (Ontological Multi-Level Modeling Language). This language combines the best modeling features considered above. Namely, O_2ML is based on:

- UML – reach intra-level modeling capabilities (plural inheritance, plural multi-dimensional classification);
- OMEGA – reach inter-level modeling capabilities (metaclasses, metaattributes, metaassociations);
- Deep Instantiation – multi-level modeling support (potency values).

Figures 5 to 7 show O_2ML usage example. One can see on these figures that potency values allow reducing metaattributes quantity. This leads to more simple and compact models. Formally, an attribute with potency value of $n > 1$ is a metaattribute with potency value of $n - 1$ that satisfies following constrains:

- set of allowed types is constrained to only one type;
- instance quantity in each owner-class instance is precisely one;
- instances names replicate their parent name.

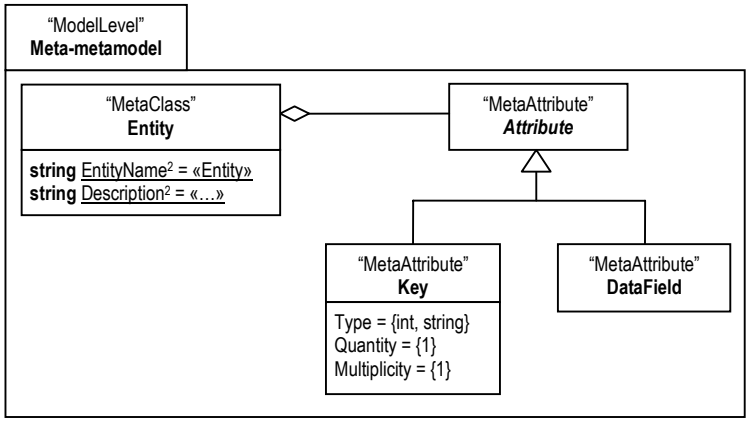


Fig. 5. O_2ML-meta-metamodel example

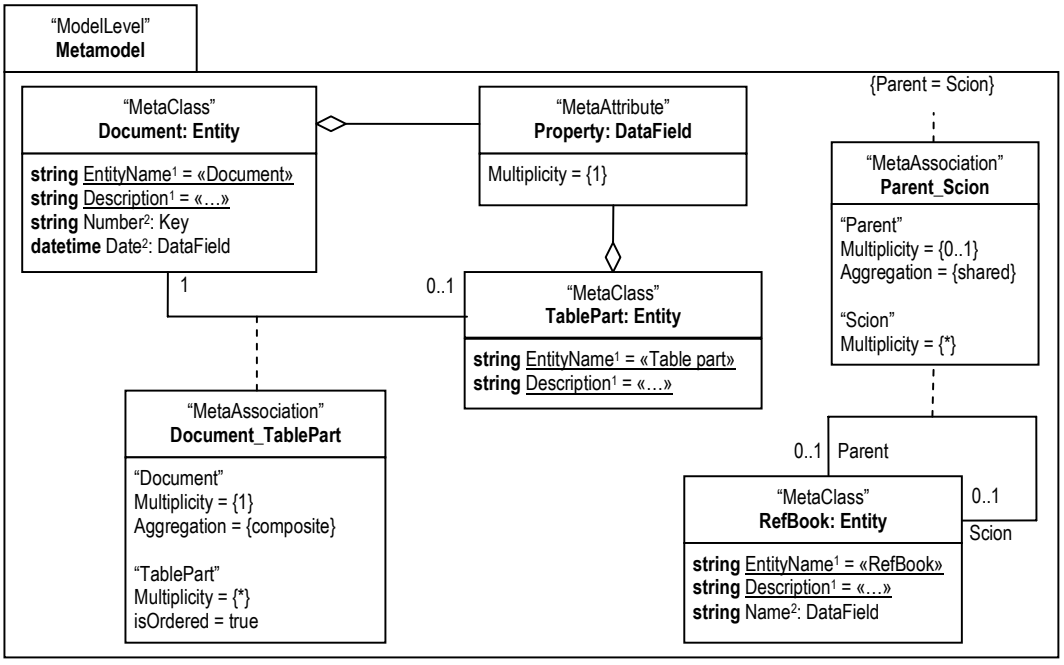


Fig. 6. O_2ML-metamodel example

O_2ML graphical notation uses attribute definition syntax that differs from the one used in UML. The syntax is as follows: <Type> <Attribute Name><Potency value>: <Metaattribute Name>. This approach conforms to the fact that metaattribute is a classifier for corresponding attributes.

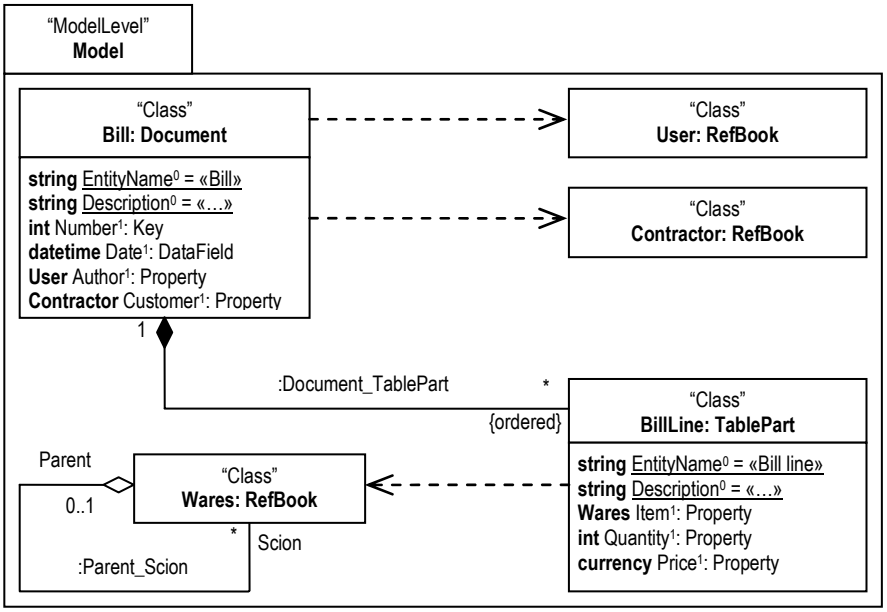


Fig. 7. O_2ML-model example

Important O_2ML feature is that its semantics is described formally. This fact eases O_2ML virtual machine creation. The semantics is described using XOCL (eXecutable OCL) – an extension of OMG's OCL and a part of XMF (eXecutable Metamodeling Framework) [3].

Bibliography

[1] Atkinson C., Kühne T. Rearchitecting the UML Infrastructure. ACM Transactions on Modeling and Computer Simulation, Vol. 12, No. 4, October 2002.

[2] Atkinson C., Kühne T. The essence of multi-level metamodeling. In Proceedings of the Fourth International Conference on the Unified Modeling Language, M. Gogolla, C. Kobryn, Eds., Lecture Notes in Computer Science, vol. 2185, 19–33, 2001.

[3] Clark T., Evans E., Sammut P., Willans J. Applied Metamodelling: A Foundation for Language Driven Development http://albini.xactium.com/web/downloads/b1a35960appliedMetamodelling.pdf, 2004.

[4] Gitzel R., Ott I., Schader M. Ontological Metamodel Extension for Generative Architectures (OMEGA), Working Paper, University of Mannheim, Department of Information Systems III, http://www.bwl.uni-mannheim.de/Schader/_files/gitzel-omega.pdf, June, 2004.

[5] Object Management Group, Common Warehouse Metamodel, http://www.omg.org/technology/cwm, 2001.

[6] Object Management Group, Meta Object Facility Core v2.0, http://www.omg.org/cgi-bin/doc?formal/2006-01-01, January 2006.

[7] Object Management Group, UML Superstructure Specification v2.0, http://www.omg.org/cgi-bin/doc?formal/05-07-04, July 2005.

Author's Information

Sergey Shavrin – Perm state university, computer science department, senior lecturer; e-mail: shavrin@gmail.com

MATHEMATICAL MODELS OF DOMAIN ONTOLOGIES[1]

Alexander Kleshchev, Irene Artemjeva

Abstract: *In this article the notion of a mathematical model of domain ontology is introduced. The mathematical apparatus (unenriched logical relationship systems) is essentially used. The representation of various elements of domain ontology in its model is considered. These elements are terms for situation description and situations themselves, knowledge and terms for knowledge description, mathematical terms and constructions, auxiliary terms and ontological agreements. The notion of a domain model is discussed. The notions of a precise ontology and precise conceptualization are introduced. The structures of situations and knowledge and also their properties are considered. Merits and demerits of various classes of the domain ontology models are discussed.*

Keywords*: Domain ontology, domain ontology model, ontology language specification, kernel of extendable language of applied logic, unenriched logical relationship systems, enriched logical relationship systems, enrichment of logical relationship system.*

ACM Classification Keywords*: I.2.4 Knowledge Representation Formalisms and Methods, F.4.1. Mathematical Logic*

Introduction

A few different definitions for the notion of domain ontology have been suggested by now. But every definition has certain flaws. Because different interpretations of the notion of a domain ontology are used when different problems related to domain ontologies are solved, it may be deduced that now there is no universally accepted definition of the notion. This article suggests another definition of the notion of domain ontology. As this takes place, the mathematical apparatus (unenriched logical relationship systems) introduced in [1-3] is essentially used.

A Mathematical Model of a Domain Ontology

An unenriched logical relationship system [3] can be considered as a domain ontology model, if each of its logical relationship has a meaningful interpretation that a community of the domain agrees with, and the whole system is an explicit representation of a conceptualization of the domain understood both as a set of intended situations and as a set of intended knowledge systems of the domain. Some examples of unenriched logical relationship systems and their meaningful interpretations as models of simplified domain ontologies were given in [1-2]. Models of ontologies for medicine close to real notions of the domain were described in [5]. Models of ontologies for physical and organic chemistry and also for roentgen fluorescent analysis were described in [6-9]. Model of ontology for classical optimizing transformations is described in [10-12].

Information concerning a finite (real or imaginary) fragment of a real or imaginary reality (the fragment may be related to a finite part of the space and to a finite time lapse) will be called a situation (a state of affairs in terms of [4]), if this fragment contains a finite set of objects and a finite set of relations among them.

Objects and relations (including unary ones) among them depending on situations are designated by special domain terms which will be called terms for situation description. Objects in situation models can be represented: by elementary mathematical objects (numbers and so on); by names having neither sort nor value [1] (such a name is a designation of an object); by structural mathematical objects (sets, n-tuples, and so on) constructed of elementary or structural mathematical objects or names having neither sort nor value by composition rules defined in the language of applied logic.

[1] This paper was made according to the program № 14 of fundamental scientific research of the Presidium of the Russian Academy of Sciences, the project "Intellectual Systems Based on Multilevel Domain Models".

The set of names having neither sort nor value and used as designations of objects (and their components) in situation models can be determined explicitly or implicitly in a domain ontology model. In the former case, all these names appear in sort descriptions for unknowns. In the latter case, all these names are constituents of parameter values. In the domain ontology model the names of these parameters are used for describing sorts of unknowns. If a domain ontology model determines some names having neither sort nor value then these names have the same meaning in every situation of the domain. A domain ontology model can determine only some of the names having neither sort nor value and used in situations for designating objects (and their components). In this case these names are determined by a model of situation and may have different meaning in different situations.

Unknowns represent relations among objects depending on situations. In different situations the relations corresponding to the same unknown can be different. Every objective unknown designates a role that in each situation an (unique) object of the situation plays, and also in every situation there is its own object playing the role. Every functional unknown designates a set of functional relations. For each situation this functional relation is the one among objects of the situation. For different situations these relations corresponding to the same unknown can be different. Analogously, every predicative unknown designates a set of nonfunctional relations. For each situation this nonfunctional relation (it may be empty) is the one among objects of the situation. For different situations these relations corresponding to the same unknown can be different.

Thus, every unknown can be considered as a designation of a one-to-one correspondence between situations and the values of the unknown in these situations.

The sort description for an unknown determines the set of value models for the unknown. In any (real or imaginary) situation only an element of this set can be a value of the unknown. Thereby, the sort description for an unknown determines a model of the capacity for the concept designated by the unknown. A model of the capacity for a concept can be both a finite and infinite set.

A model of a (real or imaginary) situation is a set of values of the unknowns for the unenriched logical relationship system representing a domain ontology model. A model of a situation can be represented by a set of value descriptions for the unknowns.

Knowledge Models and Terms for Knowledge Description

If an unenriched logical relationship system is a model of a domain ontology then any of its enrichments is a model of a knowledge system for the domain. If a model of a domain ontology is an unenriched logical relationship system O without parameters, then the ontology model introduces all the terms for description of the domain. In this case any enrichment k of the system O is a set of logical relationships – restrictions on the interpretation of names representing empirical or other laws of the domain. Since this enrichment does not introduce any new names it cannot contain any sort descriptions for names [3].

If a model of a domain ontology is an unenriched logical relationship system with parameters, then the parameters of the system are the domain terms which are used for knowledge description.

If a model of domain ontology is a pure unenriched logical relationship system O with parameters then any enrichment k of the system O is a set α_P of the parameter values for the system O [3]. A value of an objective parameter determines a feature of the domain, a set of names for situation description, or a set of parameter names. Every enrichment (a knowledge base) can introduce new names as compared with the ontology – terms for situation and knowledge description. Functional and predicative parameters represent empirical or other laws of the domain. The value of every functional or predicative parameter is some relation among terms and/or domain constants. In this case domain knowledge is described at a higher level of abstraction than in the case when a domain ontology model is an unenriched logical relationship system without parameters. The values of parameters can be represented by a set of propositions – value descriptions for names.

If a model of a domain ontology is a mixed unenriched logical relationship system O with parameters, then any enrichment k of the system O is a pair $<\Phi', \alpha P>$, where Φ' is a set of logical relationships (restrictions on the interpretation of names) representing a part of empirical or other domain laws, and αP is a set of parameter values for the system O representing the other domain laws [3]. In this case domain knowledge is represented at

two levels of abstraction: as logical relationships among unknowns of the system O and as relations among terms of the domain (as parameter values of the system O).

The sort description for a parameter determines the set of value models for the parameter. In any knowledge model only an element of this set can be a value of the parameter. Thereby, the sort description for a parameter determines a model of the capacity for the concept designated by the parameter. A model of the capacity for a concept can be both a finite and infinite set.

Mathematical Terms and Constructions. Auxiliary Terms

The language of applied logic [1] determines mathematical terms and constructions used for domain description in that the unenriched logical relationship system which is an ontology model for the domain is represented. The kernel of the applied logic language [1] determines a minimal set of logical means for domain description. The standard extension of the language [1] apart from additional logical means introduces arithmetic and set-theoretic constants, operations and relations. Every specialized extension [2] of the language gives us a possibility to define both additional logical means and constants, operations and relations of other divisions of mathematics. The specialized extensions Intervals and Mathematical quantors of language [2] introduce integer-valued and real-valued intervals, and also mathematical quantifiers. Other examples of mathematical terms which can be introduced by specialized extensions are operations of differentiation and integration, predicates of optimization, and the like.

Mathematical objects (names, numbers, sets, n-tuples, and the like) serve to represent models of elementary and combined domain objects. Mathematical functions and relations represent the properties of domain objects which are kept with mathematical models in place of domain objects. In every domain a specific mathematical apparatus is used, as a rule. This property of domains is represented by specialized extensions of the language in domain ontology models. At the same time, the practice shows that the same mathematical apparatus can be used for description of different domains. In this case, for description of ontology models of these domains the same specialized extensions of the applied logic language given by the names of these extensions can be used.

Thus, mathematical terms and constructions have more or less universally accepted designations, syntax and semantics. They are separated from a domain ontology by their definition in the applied logic language (in its kernel and extensions) rather than in the unenriched logical relationship system representing the ontology model. They are associated with the domain ontology by the fact that the name of the logical theory representing the set of logical relationships contains the names of all the extensions used for description of this theory. Using mathematical terms and constructions with this interpretation does not constrain the possibility of unenriched logical relationship systems application for representation ontologies of different domains, and mathematics is among them. In the latter case mathematical terms and constructions play the role of elements of the metalanguage with completely defined syntax and semantics, and the other terms play the role of terms of (the domain) mathematics, their semantics being defined by an ontology.

Auxiliary terms are introduced to make a domain ontology description more compact. A value of an auxiliary term is defined by the values of other domain terms: of mathematical terms, terms for situation descriptions, terms for knowledge descriptions and other auxiliary terms. The definitions of auxiliary terms are represented by a set of value descriptions for names in a domain ontology model.

Ontological Agreements

Ontological agreements about a domain are represented by a set of restrictions on the interpretation of names of the unenriched logical relationship system which is an ontology model of the domain. Ontological agreements are explicitly formulated agreements about restrictions on the meanings of the terms in which the domain is described (additional restrictions on capacity of the concepts designated by these terms).

If a domain ontology model is an unenriched logical relationship system without parameters, then all the ontological agreements are only constraints of situation models. The set of ontological agreements, in this case, can be empty, too. If a domain ontology model is an unenriched logical relationship system with parameters then the set of ontological agreements can be divided into three nonintersecting groups: constraints of situation models, i.e. the agreements restricting the meanings of terms for situation description; constraints of knowledge

models, i.e. the agreements restricting the meanings of terms for knowledge description; agreements setting up a correspondence between models of knowledge and situations, i.e. the agreements setting up a correspondence between the meanings of terms for situation and knowledge description. Every proposition of the first group must contain at least one unknown or a variable whose values are unknowns and cannot contain any parameters; every proposition of the second group must contain at least one parameter or a variable whose values are parameters and cannot contain any unknowns; every proposition of the third group must contain at least one parameter or a variable whose values are parameters and at least one unknown or a variable whose values are unknowns. In doing so, the definitions of auxiliary terms should be taken into account.

Now let us define the informal notion of domain ontology using the formal notion of a domain ontology model. The part of information about a domain, which is represented by an ontology model of the domain, will be called an ontology of the domain. It immediately follows that a domain ontology contains a set of capacity concept definitions for situations description (it cannot be empty), a set of capacity concept definitions for knowledge description (it can be empty), characteristics of mathematical apparatus for domain description, a set of auxiliary term definitions (it can be empty), a set of restrictions on the meaning of terms for situation description (it can be empty), a set of restrictions on the meaning of terms for knowledge description (it can be empty), and a set of agreements setting up a correspondence between meanings of terms for situation description and for knowledge description (it can be empty).

A Domain Model

If an unenriched logical relationship system O is a domain ontology model and $k \in En(O)$ is a knowledge model of the domain, then the enriched logical relationship system <O, k> [3] is a model of the domain. In this case the set of solutions A(<O,k>) is a model of the domain reality. Thus, the domain ontology model O determines a class of the domain models $\{<O,k>|k \in En(O)\}$. Every domain model consists of two parts: an ontology model O that is the same for the whole class and a knowledge model k, which is specific for a particular domain model <O,k>.

The set of all possible situations in a domain which have ever taken place in the past, are taking place now and will take place in the future will be called the reality of the domain. Thus, the reality has the property that the persons studying the domain, the developers of its conceptualization and its models do not know the reality completely. Only a finite subset of situations forming the reality and having taken place in the past is known (although the information forming these situations also can be not completely known). We will suggest that relative to any conceptualization of a domain the hypothesis on its adequacy is true: the reality is a subset of the set of intended situations. In view of the reality definition it is evident that this hypothesis cannot be verified. Hence, every adequate conceptualization imposes certain limitations on the notion of the reality.

A(<O, k>) is an approximation of the unknown set of models of all situations which are members of the domain reality. It is apparent that the better A(<O, k>) approximates the reality the more adequate the domain model <O, k> is. A model of a domain is adequate to the domain, if the set of models of all the situations forming the domain reality is equal to the solution set of the enriched logical relationship system which is a model of the domain, i.e. the reality approximation is precise.

We will consider only such domain ontology models O that there is the adequate model <O, k> of the domain in the class of models of the domain determined by the ontology model O (the hypothesis on existence of the adequate domain model). The hypothesis on existence of the adequate domain model is stronger than the hypothesis on conceptualization adequacy. The first hypothesis states that there is such a knowledge base (an element of the set En(O)) that A(<O, k>) is the same as the set of models of all the situations of the domain reality. The second one states only that the latter set is a subset of the set of models of all the intended situations. Inasmuch as the reality is not completely known (not all the situations which took place in the past and take place at present are known, and no future situation is known either), it is unknown for any domain model how well the reality model approximates the reality. Thus, it is impermissible to hold about any domain model that it is an adequate model of the domain. At the same time, a criterion of inadequacy can be formulated: a model of a domain represented by an enriched logical relationship system is an inadequate model of the domain, if such a situation is known which took place in the reality that its model is not a solution of the logical relationship system. If a domain ontology model O is given, and inadequacy of a domain model <O, k> is revealed, then experts

usually look for some other model of the domain knowledge k' $\in$ En(O), so that the domain model <O, k'> won't be inadequate with respect to the available data (known situations). If in the process of storing empirical data (extending the set of known situations) it becomes clear that inadequacy of the current domain model is sufficiently often found, and that the model has to be permanently modified, and that this process leads to constant increasing of the number of empirical laws and/or to constant growth of complexity of the knowledge model, then an aspiration may arise for finding another conceptualization of the domain and an ontology representing it (changing the paradigm) and for finding an adequate model of the domain within the restrictions of the new conceptualization.

A Precise Ontology and Conceptualization

A domain ontology will be called precise, if the set of situation models forming the conceptualization represented by the ontology is equal to the set $\bigcup_{k\in En(O)} A(< O,k >)$, where O is an unenriched logical relationship system that is a model of the ontology, i.e. the approximation of the conceptualization determined by the ontology is precise.

A conceptualization will be called precise, if it is the same as the domain reality. It is apparent that precise conceptualizations are impossible for the domains related to the real world. But conceptualizations are possible for theoretical (imaginary) domains (mathematics, theoretical mechanics, theoretical physics and so on) for which their precision is postulated.

If an ontology and conceptualization are precise, then the unenriched logical relationship system O being a model of this ontology must have the following property: if <O, k> is the adequate model of the domain where k $\in$ En(O), then A(<O, k'>) $\subseteq$ A(<O, k>) for any k' $\in$ En(O). If O is an unenriched logical relationship system without parameters, then the empty set of propositions is this k.

The question arises of whether in the case of precise conceptualization the empty knowledge base is always consistent with the adequate domain model. Let us discuss this question using the example of an ontology of mathematics. An ontology of mathematics (or any one of its branches) consists of definitions and axioms. Any conceptualization of mathematics is assumed to be precise. At the same time, mathematical knowledge consists of theorems (lemmas, corollaries and so on) and their proofs. Since in mathematics any theorem is a logical consequence of the ontology, the theorems impose no additional restrictions on the reality model. Thus, both the empty knowledge base and a knowledge base containing any set of theorems determine adequate (and equivalent [1]) models of mathematics. The role of theorems is to make explicit the properties implicitly given by the ontology, and the role of proofs is to make evident the truth of theorems. Some theorems can have inflexible form (identities, inequalities and so on). So a mixed unenriched logical relationship system with parameters can be a natural ontology model for mathematics where terms identities, inequalities and others describe knowledge.

The Structure of Situations and Knowledge

The set of the unknowns whose values form a model of a situation will be called the structure of the situation model. We will say that models of two situations have the same structures, if the sets of the unknowns forming the structures of these situations are the same. From this point of view, the models of all the situations belonging to the reality model of any domain model have the same structures, if this domain model is an enriched logical relationship system. As for the structures of intended situation models determined by a domain ontology model that is an unenriched logical relationship system, three cases are possible.

1. A domain ontology model is an unenriched logical relationship system without parameters. In this case all intended situation models have the same structures.

2. A domain ontology model is an unenriched logical relationship system with parameters, none of parameter values being able to contain unknowns. In this case all intended situation models also have the same structures.

3. A domain ontology model is an unenriched logical relationship system with parameters, values of some parameters being able to contain unknowns. In this case the models of the situations belonging to the reality

models of different models of the domain (consistent with different knowledge models) can have different structures depending on knowledge models.

The structures of all the situations determined by the ontology model of example 1 of article [3] are the same. They are formed by the unknowns diagnosis, partition for a sign, moments of examination, blood pressure, strain of abdomen muscles, and daily diuresis. The structures of all the situations determined by the ontology model of example 6 of article [3] also are the same. They are formed by the unknowns cubes, balls, rectangular parallelepipeds, length of an edge, volume, substance, and mass.

The parameter signs in example 2 of article [2] contains unknowns (see propositions 2.2.1 and 2.2.13 in [2]). Thus, situation models determined by this ontology model can have different structures. In [2] an example of a knowledge model for this ontology model was given (see example 3, propositions from 3.1.1 to 3.1.9). The structure of situation models corresponding to that knowledge model is formed by the unknowns diagnosis, partition for a sign, moments of examination, strain of abdomen muscles, blood pressure and daily diuresis. If in another knowledge model of the same ontology model the parameter signs has the different value

signs ≡ {pain, temperature, discharge},

and the other parameters have some proper values, then the structure of situation models corresponding to this knowledge model is formed by the unknowns diagnosis, partition for a sign, moments of examination, pain, temperature and discharge, i.e. these structures differ from one another.

Using parameters whose values contain unknowns makes it possible “to hide” some terms used for situation description in domain ontology model description. At the same time, the meanings of these unknowns are completely determined by the propositions describing the sorts of these unknowns (see proposition 2.2.13 of example 2 of [2]): models of concepts designated by these unknowns are determined, for any unknown its meaning in a situation is determined (either the unknown is a name of a role, a functional relation or an unfunctional one), for every name of relation the number of its arguments, the sorts of its arguments and the sort of its result are determined.

The set of parameters of the unenriched logical relationship system being a domain ontology model will be called the structure of domain knowledge model. It follows from this definition that if an unenriched logical relationship system without parameters is a domain ontology model, then any knowledge model of this domain has no structure. If a mixed unenriched logical relationship system with parameters is a domain ontology model, then a part of any knowledge model has a structure but the other its part has no structure. If a pure unenriched logical relationship system with parameters is a domain ontology model, then all parts of any knowledge model of the domain have a structure. Let a domain ontology model be an unenriched logical relationship system with parameters. If no parameter value in its turn contains a parameter, then all domain knowledge models for this conceptualization have the same structures. If values of some parameters in their turn contain parameters, then different knowledge models of the domain can have different structures.

Using parameters whose values contain parameters makes it possible “to hide” some terms used for knowledge description in domain ontology model description. At the same time, the meanings of these terms are completely determined by the propositions describing the sorts of these terms.

A Comparison between Different Ontology Model Classes

Now let us discuss the question about capabilities of domain models and domain ontology models, which are enriched and unenriched logical relationship systems of different classes.

Let us consider several aspects of the term “domain ontology”.

1. If a conceptualization contains intended situations of different structures, then any ontology representing this conceptualization cannot have a model in the class of unenriched logical relationship systems without parameters but can have a model in the class of the systems with parameters.

2. If a conceptualization contains concepts designated by terms for knowledge description, then no ontology representing this conceptualization can have a model in the class of unenriched logical relationship systems without parameters, but it can have a model in the class of the systems with parameters.

3. If a conceptualization contains concept classes and determines properties of the concepts belonging to these classes, and concepts themselves are introduced by domain knowledge, then no ontology representing this conceptualization can have a model in the class of unenriched logical relationship systems without parameters but can have a model in the class of the systems with parameters.

4. If in a conceptualization some restrictions on meanings of terms for situation description depend on the meaning of terms for knowledge description, then any ontology representing this conceptualization cannot have a model in the class of unenriched logical relationship systems without parameters, but it can have a model in the class of the systems with parameters.

5. The more compactly and clearly domain ontology models of a class describe agreements about domains, the better the class is. In this regard unenriched logical relationship systems without parameters require for every term for situation description to appear explicitly in these agreements. For real domains (such as medicine) the models of their ontologies turn out immense because of large number of these terms. At the same time, the systems with parameters describe agreements about domains for groups of terms, rather then only for isolated terms through using terms for knowledge description. In doing so the majority of the terms for situation description and some terms for knowledge description do not appear explicitly in agreement descriptions (they are replaced by the variables whose values are terms from appropriate groups). As a result, a model of agreements becomes compact and agreements themselves become more general.

6. The more understandable knowledge bases represented in terms of an ontology are for domain specialists, the better the class of domain ontology models is. In this respect unenriched logical relationship systems without parameters represent knowledge bases as sets of arbitrary logical formulas. The more complex these formulas are, the more difficult it is to understand them. At the same time, the systems with parameters introduce special terms for knowledge description. The meanings of these terms are determined by ontological agreements, and their connection with terms for situation description among them. In real domains these terms are commonly used to ease mutual understanding and to make communication among domain specialists economical. The meanings of these terms are, as a rule, understood equally by all domain specialists. The role of these terms is to represent domain knowledge as relation tables (sets of atomic formulas, of simple facts). It is considerably easier for domain specialists to understand the meanings of these simple facts than the meanings of arbitrary formulas.

7. The more precise approximation of a conceptualization model a class of domain ontology models assumes, the better it is.

First, let us remark that it follows from the theorem about eliminating parameters of enriched logical relationship systems [3] that if there is a domain model represented by an enriched logical relationship system with parameters which determines an approximation of the domain reality, then there is a model of the domain represented by an enriched logical relationship system without parameters which determines the same approximation of the domain reality. In this regard domain models represented by enriched logical relationship systems with parameters offer no advantages over domain models represented by enriched systems without parameters.

As for domain ontology models, every one represented by an unenriched logical relationship system determines some approximations for both the set of intended domain situation models and for the set of intended domain knowledge models. If a model O_P of a domain ontology represented by an unenriched logical relationship system with parameters determines an approximation $\bigcup_{k\in En(O_P)} A(<O_P, k>)$ of the set of intended domain situation models , then the unenriched logical relationship system O_X without parameters quasiequivalent to O_P determines the approximation $\bigcup_{k\in En(O_X)} A(<O_X, k>)$ of the same set of intended situation models [1]. Let h : En(OP) $\rightarrow$ En(O_X) be the map defined by the theorem about eliminating parameters of unenriched logical relationship systems and H = {h(k) | k $\in$ En(O_P)}. Then $\bigcup_{k\in En(O_X)} A(<O_X, k>) = \bigcup_{k\in En(O_P)} A(<O_X, h(k)>) \cup \bigcup_{k\in En(O_X)\setminus H} A(<O_X, k>)$; but by the theorem about eliminating parameters of enriched logical relationship systems

$\bigcup_{k\in En(O_P)} A(<O_X, h(k)>)$ = $\bigcup_{k\in En(O_P)} A(<O_P, k>)$, i.e. $\bigcup_{k\in En(O_X)} A(<O_X, k>)$ = $\bigcup_{k\in En(O_P)} A(<O_P, k>)$ $\cup$ $\bigcup_{k\in En(O_X)\setminus H} A(<O_X, k>)$. Thus, the approximation of the set of intended situation models determined by the system O_X, is less precise than the approximation represented by the system O_P.

If a model O_P of a domain ontology represented by an unenriched logical relationship system with parameters determines an approximation $En(O_P)$ of the set of intended domain knowledge models, then the unenriched logical relationship system O_X without parameters determines an approximation $En(O_X)$ of the same set of intended knowledge models. In this case H is a subset of $En(O_X)$, i.e. the approximation of the set of intended knowledge models determined by the system O_X also is less precise than the approximation determined by the system O_P. In what follows we show some reasons of this fact.

Let us consider the case when a domain ontology model is a pure unenriched logical relationship system O_P with parameters. First, the constraints of knowledge models represented by O_P determine the set $En(O_P)$ as a proper subset of the set of all possible interpretations of the system O_P's parameters, whereas, if the system O_X without parameters is a domain ontology model, then this ontology model contains practically no restrictions on the set $En(O_X)$. Second, for the theorem about eliminating parameters of enriched logical relationship systems a set of formulas representing empirical and other domain laws can be deduced from every proposition setting up a correspondence between knowledge models and situation models and from parameter values. These formulas contain no parameters. It is obvious that the forms of these formulas are restricted and determined by the forms of propositions setting up a correspondence between knowledge models and situation models. At the same time, if a domain ontology is an unenriched logical relationship system O_X without parameters, then this system imposes no restrictions on the form of formulas entering its enrichments.

Let us consider the case when a domain ontology is a mixed unenriched logical relationship system $O_P = <\Phi, P>$ with parameters. In this case, if $k \in En(O_P)$, then $h(k) = \Phi' \cup \Phi''$ where the propositions belonging to Φ' are deduced from every proposition of Φ setting up a correspondence between knowledge models and situation models and from parameter values (taking into account the parameter constraints) and Φ'' is such a set of propositions that $\Phi_X \cup \Phi' \cup \Phi''$ is a semantically correct applied logical theory where Φ_X is the set of all the propositions of Φ which contain no parameters, i.e. $H \subset En(O_X)$.

Domain ontology models represented by unenriched logical relationship systems with parameters are thus seen to offer certain advantages over domain ontology models represented by unenriched logical relationship systems without parameters (see also [13]).

Conclusions

In the article a notion “a mathematical model of a domain ontology” has been introduced, the representation of different elements of a domain ontology in this model – of terms for situation description and situations themselves; of knowledge and terms for knowledge description; of mathematical terms and constructions; of auxiliary terms, and ontological agreements has been considered. The structures of situations and knowledge and their properties have been considered. The notion "a domain model” has been discussed. Definitions of the notions “precise ontology” and “precise conceptualization” have been presented. Some merits and demerits of different domain ontology model classes have been discussed in details.

References

Kleshchev A.S., Artemjeva I.L. A mathematical apparatus for domain ontology simulation. An extendable language of applied logic // Int. Journal on Inf. Theories and Appl., 2005, vol 12, № 2. PP. 149-157. – ISSN 1310-0513.

Kleshchev A.S., Artemjeva I.L. A mathematical apparatus for ontology simulation. Specialized extensions of the extendable language of applied logic // Int. Journal on Inf. Theories and Appl., 2005, vol 12, № 3. PP. 265-271. – ISSN 1310-0513.

Kleshchev A.S., Artemjeva I.L. A mathematical apparatus for domain ontology simulation. Logical relationship systems // Int. Journal on Inf. Theories and Appl., 2005, vol 12, № 4. PP. 343-351. – ISSN 1310-0513.

Guarino N. Formal Ontology and Information Systems. In Proceeding of International Conference on Formal Ontology in Information Systems (FOIS'98), N. Guarino (ed.), Trento, Italy, June 6-8, 1998. Amsterdam, IOS Press, pp. 3- 15/

Kleshchev A.S., Moskalenko Ph. M., Chernyakhovskaya M.Yu. Medical diagnostics domain ontology model. Part 1. An informal description and basic terms definitions. In Scientific and Technical Information, Series 2, 2005, №12. PP. 1-7.

Artemjeva I.L., Tsvetnikov V.A. The fragment of the physical chemistry domain ontology and its model. In Investigated in Russia, 2002, 5, pp.454-474. http://zhurnal.ape.relarn.ru/articles/2002/042.pdf

Artemjeva I.L., Visotsky V.A., Restanenko N.V. Domain ontology model for organic chemistry. In Scientific and technical information, 2005, №8, pp. 19-27.

Artemjeva I.L., Restanenko N.V. Modular ontology model for organic chemistry. In Information Science and Control Systems, 2004, №2, pp. 98-108. – ISSN 1814-2400.

Artemjeva I.L., Miroshnichenko N.L. Ontology model for roentgen fluorescent analysis. In Information Science and Control Systems, 2005, №2, pp. 78-88. – ISSN 1814-2400.

Artemjeva I. L., Knyazeva M.A., Kupnevich O.A. Processing of knowledge about optimization of classical optimizing transformations // International Journal on Information Theories and Applications. 2003. Vol. 10, №2. PP.126-131. – ISSN 1310-0513.

Artemjeva I. L., Knyazeva M.A., Kupnevich O.A. A Model of a Domain Ontology for "Optimization of Sequential Computer Programs". The Terms for the Description of the Optimization Object. In Scientific and Technical Information, Series 2, 2002, № 12, pp. 23-28.(see also http://www.iacp.dvo.ru/es/)

Artemjeva I. L., Knyazeva M.A., Kupnevich O.A. A Model of a Domain Ontology for "Optimization of Sequential Computer Programs". Terms for Optimization Process Description. In Scientific and Technical Information, Series 2, 2003, № 1, pp. 22-29. (see also http://www.iacp.dvo.ru/es/)

Kleshchev A.S., Artemjeva I.L. Domain Ontologies and Knowledge Processing. Technical Report 7-99, Vladivostok: Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of Sciences, 1999. 25p. (see also http://www.iacp.dvo.ru/es/).

Authors' Information

Alexander S. Kleshchev – kleschev@iacp.dvo.ru

Irene L. Artemjeva – artemeva@iacp.dvo.ru

Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of Sciences
5 Radio Street, Vladivostok, Russia

A METHOD OF ESTIMATING USABILITY OF A USER INTERFACE BASED ON ITS MODEL

Valeriya Gribova

Abstract. *The article presents a new method to estimating usability of a user interface based on its model. The principal features of the method are: creation of an expandable knowledge base of usability defects, detection defects based on the interface model, within the design phase, and information to the developer not only about existence of defects but also advice on their elimination.*

Keywords*: Ontology, defects, interface model, user interface development*

ACM Classification Keywords*: I.2.2 Artificial intelligence: automatic programming*

Introduction

Quality and speed of software development are traditionally considered as a compromise where one of them is paid more attention than the other. However, to remain a competitive company developing software should not

only increase speed but also improve quality of its software. To achieve this aim, a lot of efforts of developers are required. According to the Cnews channel in 2001 defects in software cost the world business 175 billion US dollars.

A user interface is an integral part of most software so quality of its development is of critical importance. In addition to general criteria of software quality the user interface has an additional one, namely, usability. The user estimates the whole application program based on its user interface.

Estimating usability is an expensive task in terms of time and labor. This problem is usually solved by increasing the number of testers or by automation of the process.

In this article an additional component of automated detection of usability defects to a tool for user interface development is proposed. The main task of this component is to detect defects of usability in a user interface based on its model and to give advice on their elimination. The paper demonstrates urgency of the problem, the basic idea of the method, and an ontology of defects.

Urgency of the problem

Usability is the measure of the quality of a user's experience when interacting with an application program. It is also a combination of factors that affect the user's experience with the application program, including easiness of learning, efficiency of using, memorability, error frequency and severity, and subjective satisfaction [http://www.usability.gov].

Every year the number of interface elements and their properties is increasing. There are criteria for design of each interface element, their groups and individual characteristics depending on the user's profile (age, experience, specific requirements, etc.), the structure of a domain, a field of using an application program, a type of an application program, and so on. However, all criteria of usability are described in articles, textbooks and manuals informally, as sets of recommendations. The developer must know all these criteria. This fact requires high qualification of developers, their expertise in usability principles, and more evaluators. As a result, cost and time of development increase. To make an application program reliable and to improve its quality, it is suggested to provide the process of user interface development with a system of automated detection of usability defects.

Automation of this process has several potential advantages over non-automated methods, such as [1]:

-Reducing the cost of usability evaluation;
-Increasing consistency of the errors uncovered;
-Predicting time and error costs across an entire design;
-Reducing the need for evaluation expertise among individual evaluators.
-Increasing the coverage of evaluated features.
-Enabling comparisons between alternative designs.
-Incorporating evaluation within the design phase of user interface development.

At present only a few model-based tools for user interface development have facilities for evaluation of a user interface. However, all of them are built into a tool for development and cannot be expanded. These tools quickly become out of day because interface elements are modified, requirements to their design are changed, and new standards are established. So an expandable system of automated detection of usability defects is a problem of urgency.

The Basic Idea of the Method

The principal requirements to a system of automated detection of usability defects are expandability of the system, informing the developer about defects, and giving advice on its elimination.

The author has described a conception of user interface development based on ontologies in [2]. The main idea of this conception is to form an interface model using universal ontology models which describe features of every

component of the model and then, based on this high-level specification, generate a code of the user interface. Components of the interface model are a domain model, a presentation model, a model of linking to an application program and a model of a dialog scenario. Every component of the interface model is formed by a structural or graphical editor managed by a domain-independent ontology model.

Similarly, a presentation model is formed by a graphical editor managed by a graphical user interface (GUI) ontology model. The GUI ontology model describes knowledge required for designing WIMP (windows, icons, menus, and pointing devices) interfaces. It consists of two basic groups of elements (windows and widgets) and three additional groups (control panels, menus and extra elements). Windows are main elements in a user interface since they make up its structure. Other elements are constituents of windows. Widgets (push and radio buttons, checkboxes, lists, etc.) manage an application program and specify properties of objects. Control panels are used to get quick access to commands.

Thus, the GUI ontology model describes interface elements, their properties and interconnections. It is platform-independent and expandable.

Example 1 shows a fragment of the GUI ontology for a text element of a menu.

Example 1. A fragment of the GUI ontology

The example shows the hierarchy of menu elements (see Fig. 1) and description of a text menu element.

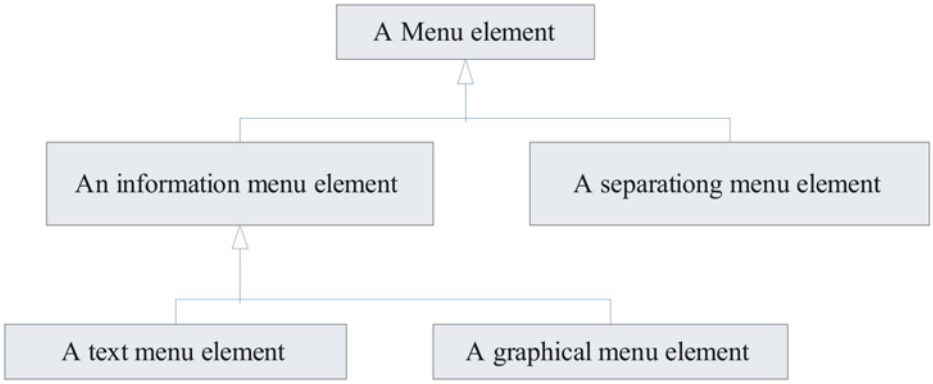


Fig. 1 The hierarchy of menu elements

A text menu element

Description: a class for presenting menu elements with verbal information.

Superclass: an information element of a menu.

Parameters:

Text: describes name to a menu element [type: String]

Prefix: describes prefix of the element [type: String]

Postfix: describes postfix of the element [type: String]

Font: describes font of the element [type: Font parameters]

Background: describes background color of the element [type: color]

A particular presentation component of a user interface model is a subset of the GUI ontology model. It means that to form a presentation component of the user interface model the developer is to determine values of properties of the GUI ontology model. This process requires that the developer should have expertise in usability principles; otherwise a presentation component a user interface model would have defects.

To detect these defects a knowledge base of interface defects has been made. Every element of this knowledge base is linked to elements of the GUI ontology model. It should be noted that since the GUI ontology model is

expandable, when a new element is added to this ontology model, description of a defect in the knowledge base could be modified or a new description of a defect could be added.

There can be two ways to detect defects in the interface model. The first one is detecting a defect in designing an interface element, e.g., when the developer forms a string (a component of an interface element). If the length of this string exceeds a maximal length, the developer can be informed immediately. The second one is checking a set of properties in different interface elements. It is possible only after a fragment of an interface has been designed. For example, to detect a defect of an interface element arrangement in a window it is necessary to design this window first and then to check it. Therefore, the system of automated detection of usability defects is to work in two modes. Fig. 2 shows the basic architecture of the system.

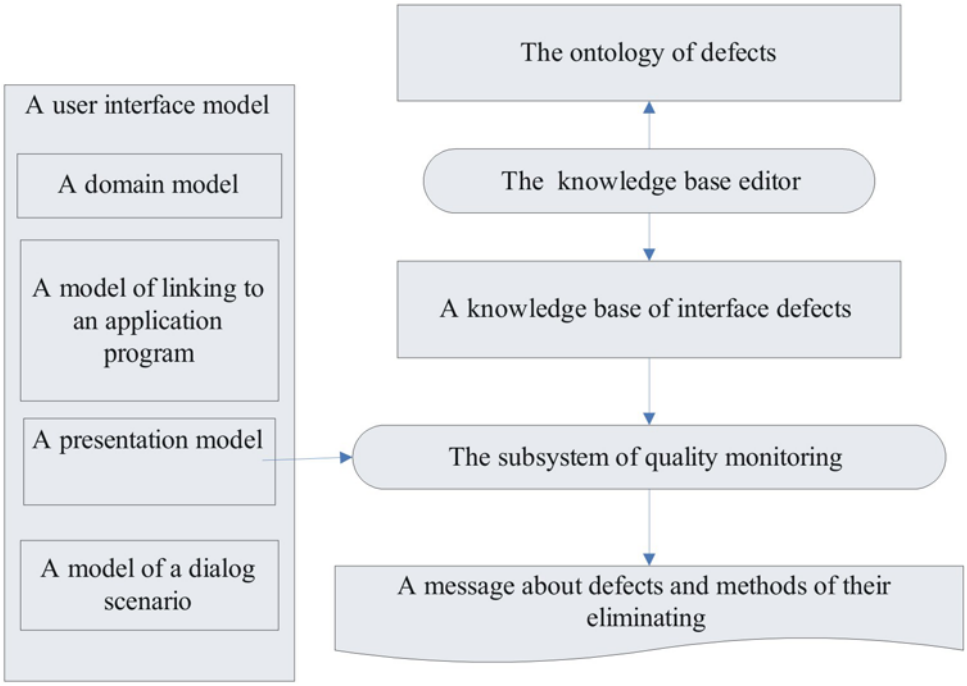


Fig. 2 The basic architecture of the system of automated detection of usability defects.

Ontology of Defects

A defect (fault) is detected in software when the developer makes a mistake due to a typo, poor understanding of some processes, principles, and so on. A defect is a coded mistake of the developer. To detect defects in software it is necessary to accurately classify them. The following ontology of defects is proposed.

1. ***Name of a defect***.

2 ***Type of a defect***. There may be two defect types, namely, presentation element defects or composition defects. The former occurs in designing an interface element; the latter is found after designing a set of different interface elements.

3. ***Name of a class***. It is a metaterm of the GUI ontology model. It indicates a class name of the interface element whose defect is described. This item can contain some classes. On the one hand, an interface element can consist of some classes; on the other hand, when a composition defect is described we must include all classes involved in detecting a defect.

4. ***Superclass***. It is also a metaterm of the GUI ontology model indicating a name of a parent class.

5. ***Parameters***. There are the parameters that are used for detecting a defect. They correspond to parameters of a class from the GUI ontology model.

6. ***Method of detecting***. It is an algorithm of detecting a defect.

7. ***Advice***. It is a message to the developer on eliminating a defect.

To illustrate the above, let's consider the following descriptions of defects from the knowledge base based on the ontology of defects.

Name of the defect: too many menu elements.

Type of the defect: a presentation element defect.

Name of the class: a top-level menu.

Superclass: menus.

Parameters: the number of menu elements.

Method of detecting: the number of menu elements>9

Advice. This menu consists of more than 9 elements. It will be difficult for the user to perceive. The number of menu elements should be decreased.

Name of the defect: the window has no name.

Type of the defect: a presentation element defect.

Name of the class: a window.

Superclass: an element of the GUI ontology model.

Parameters: the name (type: Boolean).

Method of detecting: the name=0

Advice. The window has no name.

Summary

In this article an approach to automated detection of usability defects is proposed. The basic idea of the approach is to add a system of automated detection of usability defects to the tool for user interface development operated by a knowledge base of interface defects. The main task of the system is to detect defects in a user interface model within the design phase and to give advice to the developer on their elimination.

At present a prototype of the system has been developed at the Intellectual Systems Department of the Institute for Automation and Control Processes, the Far Eastern Branch, the Russian Academy of Science.

The GUI ontology model and a knowledge base of interface defects corresponding with this ontology are used at the Mathematics and Computer Science Institute of the Far Eastern National University within the course "User Interfaces".

Acknowledgements

The research was supported by the Far Eastern Branch of Russian Academy of Science, the grant «An Expandable System for Quality Monitoring».

Bibliography

1. Ivory, M.Y., Hearst, M.A.: State of the Art in Automating Usability Evaluation of User Interfaces. ACM Computing Surveys, 33 (December 2001) 1–47. Accessible at http://webtango.berkeley.edu/papers/ue-survey/ue-survey.pdf .
2. Gribova V., Kleshchev A. From an Ontology-oriented Approach Conception to User Interface Development. International Journal "Information Theories & Applications". 2003. vol. 10, num.1, p. 87-94

Author's Information

Gribova Valeriya – Ph.D. Senior Researcher of the Intellectual System Department, Institute for Automation & Control Processes, Far Eastern Branch of the Russian Academy of the Sciences: Vladivostok, +7 (4323) 314001; e-mail: gribova@iacp.dvo.ru; http://www.iacp.dvo.ru/es.

SEMANTIC SEARCH OF INTERNET INFORMATION RESOURCES ON BASE OF ONTOLOGIES AND MULTILINGUISTIC THESAURUSES

Anatoly Gladun, Julia Rogushina

***Abstract**: the approaches to the analysis of various information resources pertinent to user requirements at a semantic level are determined by the thesauruses of the appropriate subject domains. The algorithms of formation and normalization of the multilinguistic thesaurus, and also methods of their comparison are given.*

***Key words:** an information resource, ontology, thesaurus, informational retrieval.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing, I.2.4 Knowledge Representation Formalisms and Methods (F.4.1).*

Introduction

During last years the Internet becomes one of the main means of the information publication. It is dynamical distributed environment, and the information resources (IR), presented in it, are heterogeneous. The effective retrieval of Internet IR by expand of network amount and complexity becomes more and more difficult and laborious. Thus critical is not the search time but selection of IR that satisfy to real information needs of users.

The quality estimation of information retrieval systems (IRS) is a complex question [1]. The problem concerns with parameters of IRS estimation. A lot of existing techniques analyze such IRS parameters as relevance, completeness, accuracy and their various combinations. Relevance is a thematic correspondence of the information, received as a result of search, to request. The completeness of search is a ratio of the correctly found documents amount to the total relevant documents known to IRS. Accuracy of search is a ratio of correctly found documents amount to the total amount of the documents given by IRS in reply to request.

However it is necessary to take into account, that the formal request to IRS is the user attempt to formalize his/her information need that, unfortunately, not always really reflects this need. It results in degradation of Internet use. Therefore more important such parameter of IRS quality estimation as pertinence – a ratio of amount of the information interesting for user to total amount of the received information. To increase the pertinence of informational retrieval IRS requires information about area of the user interests. This information applies by IRS for choose among accessible resources what are interesting to user and not only formally correspond to request. Such information should be submitted in the form suitable for automatic processing and reuse, and their formation must be automatized.

Internet Informational Resources

Among IR, potentially accessible to the Internet users, still prevails the textual information mainly in HTML and XML formats however it's share constantly decreases due to multimedia IR increase. The subject domain that is characterized by these IR can be represented by two ways: 1) analyzing textual information and 2) considering metadata of these IR.

Metadata contains machine-readable information about the document, which can be automatically processed by computer. Now the most perspective and common metadata model is RDF (Resource Description Framework) based on XML. With the help of RDF one can describe both structure of a site and connected with appropriate domain. RDF describes informational resources in oriented marked graph form - each IR can have properties,

which in turn also can be IR or their collections. Most widespread set of elements for metadata specification is Dublin Core Metadata Elements. Metadata can be built in IR or be stored and updated independently of resources.

Multimedia data. Recently Internet IR along with the textual information includes the graphic elements, video, sound etc. There is a great deal of the widespread formats for a storing of audio and video information, 3D-scripts and images. The multimedia resources are accessible for indexation much worse than textual information. If the information about multimedia IR is not submitted by their provider explicitly in any format known for indexing mean, it is a necessity to apply the complex and laborious operations (image recognition, speech recognition etc.). Now MPEG group develops a number of standards for representation of multimedia information metadata (for example, MPEG7 and MPEG21). In spite of significant differences between multimedia and textual IR, most acceptable for realization of information retrieval (taking into account time of its fulfilment and data level of index BD) is their description with the help of the same means, as textual IR: key words, file size and date of it's creation etc.

Web-services. Initially World Wide Web technology was focused on work with static hypertext documents represented in the Internet. But then sites offering to the clients not only the documents, but also service (for example, sites of e-commerce) began to occur. Many such sites use application servers, which not only return the document but can process the data entered by the user (queries, completed form etc.) and dynamically generate the documents depending on the parameters, specified by the user. Such dynamic component of the Internet grows much faster then static one and requires application of more complex information technologies. In this connection it is possible to consider a separate class of IR - Web-services.

Web-service is a set of logically connected and program-accessible through the Internet functions. There is the program identified on UR. It's interface can be determined by XML structures. Web-services are based on three basic Web-standard: SOAP (Simple Object Access Protocol) - the protocol for sending of messages by the HTTP and other Internets protocols; WSDL (Web Services Description Language) - language for the description of program interfaces of Web-services; UDDI (Universal Description, Discovery and Integration) - indexing standard of Web-services.

Statement of Problem

For effective search of the information that user needs (textual and multimedia documents, information services etc.) there is necessary to generate the model of user interests domain (for example, as ontology) and use this model when IRS fulfils the user's query.

Thesauruses and ontologies as means of domain knowledge representation

Every domain has phenomena that people allocate as conceptual or physical objects, connections and situations. With the help of various language mechanisms such phenomena contacts to the certain descriptors (for example, names, noun phrases).

For the successful solution of an informational retrieval task it is necessary to present user knowledge about domain of her/his interests in some form suitable for computer processing. The specifications of high-level domain are formed by integration of the domain structures of low-level domains. It is important to achieve an interoperability of domain knowledge representation. Ontological approach is an appropriate tool for solution of this task. Ontology is an agreement about common use of concepts that contains means of representation of subject knowledge and agreements on methods of reasons. It can be considered as the certain description of the

views on the world in some specific sphere of interests. Ontology consists of: 1) a set of the terms; 2) a set of rules of their use that limit their meanings in the context of concrete domain [2].

The ontology is knowledge base of a special kind with the semantic information about some domain. It is a set of definitions in some formal language of declarative knowledge fragment focused on joint repeated use by the various users in the applications.

Ontological commitments are the agreements aimed at coordination and consistent use of the common dictionary. The agents (human beings or software agents) that jointly use the dictionary do not feel necessity of common) knowledge base: one agent can know something that don't know the other ones, and the agent that handles the ontology is not required the answers to all questions that can be formulated with the help of the common dictionary.

Every domain with the certain subject of research has it's own terminology, original dictionary used for discussion of typical objects and processes of this domain. The library, for example, involves the dictionary relating to the books, references, bibliographies, magazines etc. Thus, pattern of domain is discovered by its dictionary - the set of words that are used in this domain. Clearly, however, that the specificity of domain is shown not only in the appropriate dictionary. Besides, it is necessary: (i) to provide strict definitions of grammar managing of combining the dictionary terms into the statements, and (ii) to clear logic connections between such statements. Only when this additional information is accessible, it is possible to understand both nature of domain objects and important relations established between them. Ontology - structured representation of this information [3].

The formal model of domain ontology O is an ordered triple $O = < X, R, F >$, where X - finite set of subject domain concepts that represents ontology O; R - finite set of the relations between concepts of the given subject domain; F - finite set of interpretation functions of given on concepts and relations of ontology O.

Until recently term "thesaurus" was used as a synonym of ontology, however now in IT with the help of the thesauruses frequently describe domain lexicon in a semantic projection, and ontology apply to modelling semantics and pragmatists in a projection to representation language [4]. The models either of ontologies or of thesauruses include as the basic concept the terms and connections between these terms.

The term "thesaurus" for the first time was used still in XIII century by B.Datiny as the name of the encyclopaedia. In translation from Greek "thesaurus" means treasure, riches. The thesaurus is the complete systematized data set about some field of knowledge allowing the human or the computer to orient in it.

The thesaurus is a dictionary where the descriptors of the certain field of knowledge with ordering of their hierarchical and correlative relations are represented. The descriptors are given in alphabetic order but they are grouped semantically; the search is carried out from concept to a word. Collection of the domain terms with indication of the semantic relations between them is a domain thesaurus. The thesaurus can be considered as a special case of ontology. The thesaurus is a pair $Th = < T, R >$, where T - finite set of the terms; and R - finite set of the relations between these terms.

The multilingual thesaurus is a coordinated set of the monolingual thesauruses containing equivalent descriptors on languages-components necessary and sufficient for interlingual exchange, and including means for the indication of their equivalence. At an recognizing of equivalence of descriptors of the various monolingual versions it is necessary to distinguish on different languages-components the following degrees of equivalence of the terms: 1) complete; 2) incomplete; 3) partial; 4) absence of the equivalent term. Incomplete equivalents are the terms, for which the volumes of concepts, expressed by them, are crossed. Partial equivalents are the terms, for which volume of concept expressed by one equivalent, is included into volume of concept expressed by other equivalent. One way to the recognizing of equivalence to a various degree bases on appropriate domain ontology

use: every word from the monolingual thesauruses refers to one of the ontology terms that helps to make connection between words of the various thesauruses. If some words from thesaurus refer to one ontology term then they are equivalent. If some words refer to ontological terms being a subclass one another then these words are in relation of incomplete equivalence.

Use of thesauruses for IR retrieval

For taking into account semantics of area of user interests in process of retrieval of IR satisfying his/her informational need o it is necessary (fig. 1):

1. to generate the domain thesaurus corresponding to information needs of the user (by analysis of IR that this user considers relevant to this domain [5];
2. to construct the thesaurus for every IR known to IRS (simple dictionary without stop-words);
3. to compare the thesauruses of IR relevant to user query to IRS with the domain thesaurus and to find those ones that contain the maximum number of words in intersection.

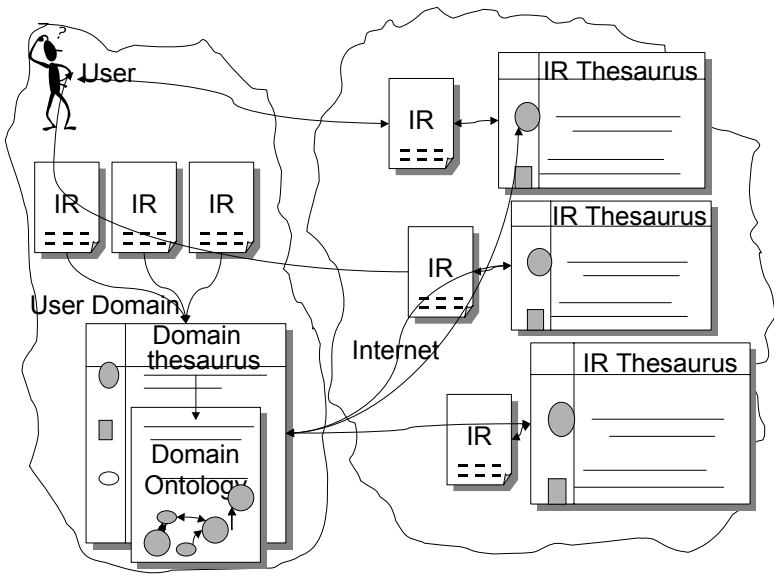


Fig.1. Informational retrieval on base of thesauruses

At thesaurus construction it is necessary to use ontologies of the appropriate areas (with higher level in comparison with user domain to normalize the multilingual thesauruses). Normalization procedure is similar to stemming and provides for integrated processing of words in different morphologic forms and multilingual representations. Normalised thesaurus contains relation between equivalent terms in different languages. As every thesaurus is constructed from the user point of view (which is reflected in user domain ontology), therefore it's forming is the user task.

Constructing of domain thesaurus

At first user should independently select the set of IR that he/she considers relevant to domain of his/her interests. Every IR is described by not empty set of the textual documents connected with this IR - text of content, metadescriptions, results of indexing etc. The domain thesaurus is formed as a result of the automated analysis

of these documents (the user actions are reduced to constructing of semantic bunches - by linking of each word of the formed thesaurus with some term of domain ontology. Algorithm of domain thesaurus construction consists from the following steps:

1. Formation of initial set of the textual documents relevant to domain. At the input of algorithm the set A of the textual documents describing chosen IR comes (each of documents from A can have the coefficient of importance and the coefficient of IR relevance *И*Р that allows defining differently weight of words from these documents for the IR description).

2. Creation of domain information space. For every document from A $a_i \in A, i = \overline{1,n}$ the IR thesaurus $T(a_i)$ - dictionary that contains all words occurred in the document a_i - is constructed. The IR thesaurus is formed as union of the thesauruses a_i: $T_{IR} = \bigcup_{i=1}^{n} T(a_i)$, and domain thesaurus - as association of the IR thesauruses.

3. Clearing of the thesauruses. User should specify dictionary for every $a_i \in A, i = \overline{1,n}$ containing a stop-words (for example, prepositions and conjunctions of language of the document are stop-words for it but prepositions and conjunctions of other language used as examples do not concern to them) $s_j, s_j \in Voc$. It is necessary to remove words contained in $s_j, s_j \in Voc$ from the thesauruses. Then all service information is rejected (for hypertext, for example, there are marking tags). The cleared thesauruses $T`(a_i), \forall p \in T(a_i) \Rightarrow p \in T`(a_i) \vee p \in s_j, T`(a_i) \cap s_j =$ thus are formed. The cleared thesaurus *И*Р is under construction as association of the cleared thesauruses $a_i$: $T_{ИР} = \bigcup_{i=1}^{n} T(a_i)$ $T`_{ИР} = \bigcup_{i=1}^{n} T`_{ИР}(a_i)$, and cleared domain thesaurus - as association of the IR thesauruses.

4. Linking of thesaurus with domain ontology. To integrate processing of words with equivalent semantics (for example, synonyms, translations of the term on different languages, various kinds of a spelling) the domain thesaurus is associated with some domain ontology (the user can form it himself, use ready ontology or it's modification).

Each word from the thesaurus it is necessary to link with one of the ontological terms. If the relation is lacking the word is considered as a stop-word or marking element (for example, HTML tag) and should be rejected. $\forall p \in T`(a_i) \exists t = Term(p,O) \in T_O$. The group of the IR thesaurus words terms connected with one ontological term named *the semantic bunch* $R_j, j = \overline{1,n}$ is considered as a single unit. $\forall p \in T`_{ИР} \exists R_j = \{r : r \in T`_{ИР}, Term(p,O) = Term(r,O)\}$. It allows to integrate processing of semantics of the documents written on various languages and, thus, to ensure the multilinguistic analysis of the Internet IR.

5. Extension of ontology. If the IR thesaurus contains words that can't be linked with ontological terms but user considers that these words are significant than it is necessary to add the appropriate terms to domain ontology, specify their connection with other terms of ontology and return to step 4.

6. Construction of the normalized domain thesaurus , i.e. association of all terms of domain ontology that are connected with words from the normalized IR thesaurus (fig. 2):

The normalized thesaurus is a projection of set of the IR thesaurus words on set of the domain ontology terms. $L_{ИР} = \{t : p \in T`(a_i), i = \overline{1,n}, t = Term(p,O) \in T_O\}$, and normalized domain thesaurus is a union of the normalized IR thesauruses (fig. 3).

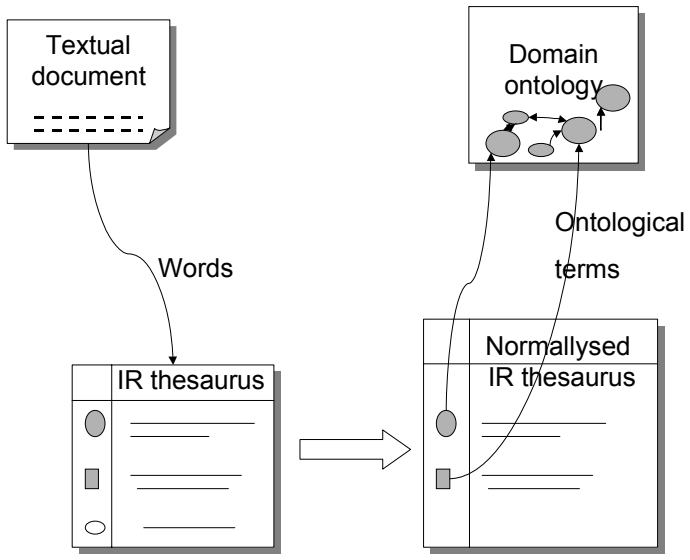


Fig.2. Building of normalized IR thesaurus

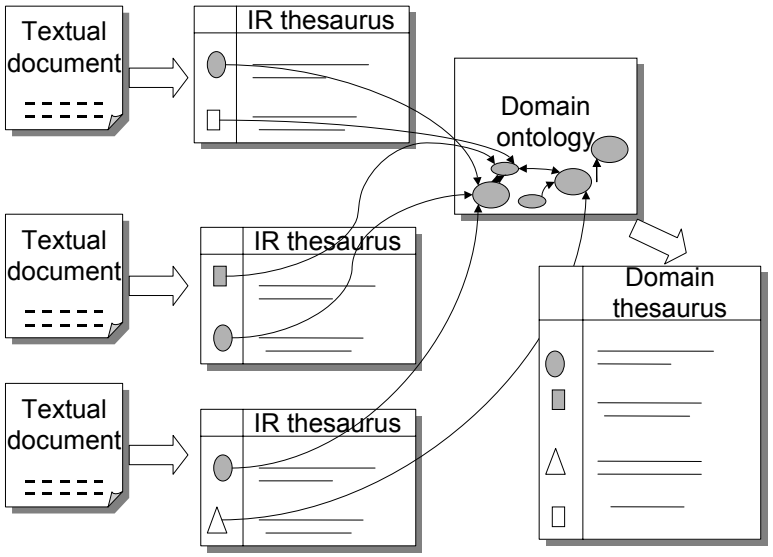


Fig.3. Building of domain thesaurus

Building of IR thesaurus

The thesaurus of IR found by IRS as a result of the user query execution is simple a dictionary that does not contain the relations between words (discovery of such connections from the text is rather difficult and in this case is not justified).

The algorithm of the IR thesaurus building consists of the following steps:

1. Formation of the initial IR set U, $U = \{IR_j, j = \overline{1,m}\}$.
2. Formation of the IR thesauruses from U. For each IR a thesaurus is formed and cleared.
3. Construction of the normalized IR thesauruses: for normalization the semantically bunches generated by the user during formation of the domain thesaurus are used .

Algorithm of domain and IR thesaurusus comparison

The normalized IR thesauruses L_{IR} and domain thesaurus L_{domain} are the subsets of the domain ontology terms O chosen by the user: $L_{IR} \subseteq Term(O)$, $L_{domain} \subseteq Term(O)$. If IR description contains more words linked with terms of domain interest for user (that is reflected in the normalized domain thesaurus) then it is possible to suppose that this IR can satisfy informational needs of the user with higher probability than other IR relevant to same formal query. Thus, it is necessary to find IR q satisfied the conditions $f(q, L_{domain}) = \max f(L_{IR}, L_{domain})$ where the function f is defined as number of elements in crossing of sets L_{IR} and L_{domain}: $f(A,B) = |A \cap B|$. If the various terms of the normalized thesauruses have for the user different importance it is possible to use the appropriate weight coefficients w_j that take into account their importance. In that case the criterion function is $f(A,B) = \sum_{j=1}^{z} y(t_j)$, where the function y is determined for all terms of domain ontology and $y(t_j) = \begin{cases} 0, t_j \notin A \vee t_j \notin B \\ w_j, t_j \in A \wedge t_j \in B \end{cases}$.

Conclusion

The proposed approach to use of domain ontology for creation and normalization of the IR thesauruses allows fulfilling informational retrieval at a semantic level abstracting from language of the IR description. The application of thesaurus measure of the information allows to offer to the user only understandable to him/her items of information that provides pertinence of information retrieval.

Bibliography

1. S. Bechofer and C. Goble. Thesaurus construction through knowledge representation. Data & Knowledge Engineering, 37:25–45, 2001.
2. Gruber T.R. A translation approach to portable ontologies // Knowledge Acquisition, N 5 (2), 1993. – P.199-220..
3. IDEF5 Method Report. Knowledge Based Systems, Inc.1408 University Drive East College Station, Texas 77840, 1994. – 175 pp.
4. Takeda H., Takaai M., & Nishida T. Collaborative development and Use of Ontologies for Design // Proceedings of the Tenth International IFIP WG 5.2/5.3 Conference PROLAMAT 98, September 9 – 10 – 11, 12, Trento, Italy, 1998.
5. Gladun A., Rogushina J., Shtonda V. Ontological Approach to Domain Knowledge Representation for Informational Retrieval in Multiagent Systems // International Jornal "Information Theories and Applications", V.13, N.4, 2006. – P.354-362.

Authors' Information

Anatoly Gladun – PhD, since 1997 works as Senior Researcher in International Research and Training Centre of Information Technologies and Systems, National Academy of Sciences and Ministry of Education of Ukraine, 44 Glushkov Pr., Kiev, 03680, Ukraine, e-mail: glanat@yahoo.com

Julia Rogushina – PhD, since 1997 works as Senior Researcher in Institute of Software Systems, National Academy of Sciences of Ukraine, 44 Glushkov Pr., Kiev, 03680, Ukraine, e-mail: jjj_@cybergal.com

UNCERTAINTY AND FUZZY SETS: CLASSIFYING THE SITUATION

Volodymyr Donchenko

Abstract: *The so called "Plural Uncertainty Model" is considered, in which statistical, maxmin, interval and Fuzzy model of uncertainty are embedded. For the last case external and internal contradictions of the theory are investigated and the modified definition of the Fuzzy Sets is proposed to overcome the troubles of the classical variant of Fuzzy Subsets by L. Zadeh. The general variants of logit- and probit- regression are the model of the modified Fuzzy Sets. It is possible to say about observations within the modification of the theory. The conception of the "situation" is proposed within modified Fuzzy Theory and the classifying problem is considered. The algorithm of the classification for the situation is proposed being the analogue of the statistical MLM(maximum likelihood method). The example related possible observing the distribution from the collection of distribution is considered*

Keywords: *Uncertainty, Fuzzy subset, membership function, classification, clusterization.*

ACM Classification keywords: *I.5.1.Pattern Recognition: Models Fuzzy sets; G.3. Probability and Statistics: Stochastic processes; H.1.m. Models and Principles: miscellaneous*

Introduction

Classical conception of Fuzzy Subsets (ClasFsS) proposed Lotfi Zadeh [Zadeh, 1965] (see methodological view in [Kaufmann, 1982]) seemed to propose the practice a new method to manipulate with uncertainty. This method from the very beginning considered to be alternative to the ones already had been in use that time: statistical, maxmin, interval. The proposition to take into account and formalize the idea of the intermediate, transitional domains between "crisp" alternatives was the essence of the ClasFsS.

As it was mentioned earlier from the very beginning the ClasFsS considered by it founders on one hand as being alternative to those had been being in use by the moment of birth of the theory("isolationism"), and on the other hand as the theory alternative to the classical, "crisp", set theory. Particularly, ClasFsS considered having nothing mutual with the statistical methods. Both of these pretensions seem to be symptoms of the theory coming into being.

Indeed, as relating set theory ClasFsS is in the "naive" faze and nothing like "crisp" set theory axioms there exit. As to the last, in the [Donchenko, 2005] (see also [Donchenko, 1998-3], [Donchenko, 1998-4], also [Donchenko, 2004]) the attempt is represented to propose some form of the abstraction axiom. Also, the full absence of somewhat that one may be nominated to be logical calculus characterizes the situation in ClasFsS now, though there are attempts to say about "Fuzzy logic". But the one called "Fuzzy logic" is simply analogue of "crisp" propositions calculus [Kaufmann, 1982] see also [Kaufmann and Gupta, 1991]). It is reasonably to note opportunely, that functions of this "fuzzy calculus" are not full in the space of all function on "fuzzy propositions" as it is in Boolean algebra.

Also, pretensions of ClasFsS to be alternative to statistical method which namely is the way to investigate uncertainty through the frequencies of results, are not fruitful for the ClasFsS because all the power of statistical methodology including interpretation turned out to be cut off. Statistical interpretation is natural for a membership function [Donchenko, 2005]. Besides, the considerations of that work demonstrating the gap in the object of uncertainty in the definition of ClasFsS pointed out the way to make the definition of ClasFsS correct in this aspest. This modification of the ClasFsS embodied in conception of Modified Fuzzy Sets (MoFS) in the paper has been already referred to [Donchenko, 2005] and earlier publications [Donchenko, 1998-3], [Donchenko, 1998-4], also [Donchenko, 2004]). Namely, in MoFS classical membership function $\mu(e), e \in E$ becomes to be a function of two arguments: from $e \in E$ and from "crisp" predicate $P \in \wp$ – or correspondent "crisp" set. In such variant of

determination "property $P \in \wp$ is described fuzzy way by the modified membership function $\mu^{(P)}(e), e \in E, P \in \wp$" with preserving general restriction $\mu^{(P)}(e) \in [0,1], e \in E, P \in \wp$, P being a parameter.

Definition. By the Modified Fuzzy Set (MoFS) we will mean the pair (E, $\mu^{(P)}(e)$, $P \in \wp$), where E – abstract set (supporter) and $P \in \wp$ with the $\wp$ - the set of "crisp" predicate on abstract "crisp" universal set U_P (or correspondent "crisp" subset of U_P).

MoFS Model Example: generalized variants of logit- and probit-regression

Logit- and probit- regressions and its Generalized variants are the best example of MoFS. As it is known, in these variants of the regressions are the dependence of the Bernoulli distribution parameter on the vector $e \in E = R^m$ is considered. This dependence has the next $\beta \in R^m$ parameterization:

$$P_e\{Y = 1\} = G(e^T\beta),$$

where Y - Bernoulli - random variable, G(z) appropriate known distribution function or the tail of the distribution.

Parameter $\beta \in R^m$ is to be estimated via observations (sample) $(e_1, y_1),...,(e_n, y_n)$:

$e_i \in R^m, i = \overline{1,n}$ - non-random,

$y_i \in \{0,1\}, i = \overline{1,n}$ - the values (realizations) of independent Bernoulli - random variable $Y_i, i = \overline{1,n}$ with the parameters, correspondingly:

$$P_{e_i}\{Y_i = 1\} = G(e_i^T\beta), i = \overline{1,n}.$$

In this model case:

- $E = R^m$;
- P as a predicate defined by the relation: $P = \{Y = 1\}$;
- $\mu^{(P)}(e) = G(e^T\beta)$.

Evidently, $\mu^{(P)}(e)$ is parameterized by $\beta \in R^m$.

General character of the MoFS model example: statistical interpretation of the MoFS

In the paper [Donchenko, 2005] (see also [Donchenko, 1998-3], [Donchenko, 1998-4]) two theorems have been proved made possible statistical interpretation of the ClasFsS and MoFS . These theorems, formulated for discrete and non discrete supporters E are the next.

Theorem 1. For any finite collection of the ClasFsS $(E, \mu_{\underline{A_i}}(e))$, $i = \overline{1,n}$ with the one and the same supporter E one may find discrete probability space (Ω, B_Ω, P), collection of the evens $A_i \in B_\Omega, i = \overline{1,n}$ and complete collection of the events He : $H_e \in B_\Omega$, e∈E, – within this probability space, that all of the membership functions $\mu_{\underline{A_i}}$, $i = \overline{1,n}$ may be represented as the systems of conditional probabilities in the next way:

$$\mu_{\underline{A_i}}(e) = P\{A \mid H_e\}, \text{ for any } e \in E, \ i = \overline{1,n}.$$

Theorem 2. Given the:

- (E, $\Im$, m)- the space with a measure;
- $(E, \mu_{\underline{A_i}}(e))$, $i = \overline{1,n}$ $\mu_i(e)$, i > 0 a collection on Fuzzy subsets with the equal supporters E;
- all of the membership functions $\mu_{\underline{A_i}}(e)$, $i = \overline{1,n}$ are $\Im$, £, measurable (£– Borel σ -algebra on R^1),

then:

- exist probability space (Ω, B_Ω, P),
- exist ξ discrete random S_p – valued random variable on (Ω, B_Ω, P) with S_p is n-element set with elemens say S_i, $i = \overline{1,n}$;
- exist η random E – valued random variable on (Ω, B_Ω, P)

such, that for any $i = \overline{1,n}$

$$\mu_{\underline{A_i}}(e) = P\{ \xi = S_i \mid \eta = e \},$$

where $P\{\xi = S_i \mid \eta \}$ – conditional distribution of r.v. ξ respectively r.v.. η.

Remark 1. Both of the theorems demonstrate, that for ClasFsS exist latent "crisp" predicates-events (or correspondent "crisp"sets): $A_i \in B_\Omega$, $i = \overline{1,n}$ for the first case and $\{\xi_i = S_i\} \in B_\Omega$, $i = \overline{1,n}$, - which are characterized in a Fuzzy way. For the MoFS variants, as it has been demonstrated by the logit-, probit- model example, these predicates are presented in the modified definition patently.

MoFS and MoFS model example in the context of plural models of uncertainties

When saying about ClasFsS or MoFS role in uncertainty description it is interesting to create "general platform" in which basic theory for uncertainty manipulating can take their own places. It is likely the so call "plural" model of uncertainties to be such platform.

"Plural" uncertainties model start from the conception of "observation" and "observation situation".

When saying "observation situation" we mean "conditions" (denoted by κ) plus "observation" by itself. In its turn "conditions" in the "observation situation" consist of "varying part" (denoted by x) and on default part (denoted by f). As to "observation" then compulsory part of them is the response y on the conditions κ. But there is no precise meaning of the "observation". To put it more precisely "observation" may be interpreted in the next three variants:

$$\text{"observation"} = \begin{cases} y \\ (x, y) \\ (\kappa, y) \end{cases}.$$

It is necessary to say, that standard meaning of a sequence of "observations"(real or virtual) is $(x_1, y_1),...,(x_N, y_N)$ while it is necessarily to be $(\kappa_1, y_1),...,(\kappa_N, y_N)$.

Definition. "Plural" model of uncertainties is the model, based on response y plurality in the sequence of "observations": real or virtual.

So, when saying about uncertainties one have to answer himself what content of "observation" is in use.

Indeed, plurality in y take place in deterministic case when "conditions" are of the form $\kappa_1 = (x_1, f),..., \kappa_N = (x_N, f)$, but "observations" treat to be $y_1,..., y_N$ .This case may be interpreted as the "latent" parameter case, as it qualified in physics.

Plurality in y under just and the same $\kappa \equiv \kappa_i, i = \overline{1,N}$ is the object of application of statistical method. As it is well known all result observing or may be observed in observations are described and the frequencies (may be its limits: probabilities) of the results or collections of the results (events) are considered.

Classical regressions illustrates the application of the statistical methods when there is plurality in y in observations with common κ , when response y is real and variability in x take place. Application of Least Square Method (LS) or its modifications are common in this case.

Generalized variants of ligit- and probit-regressions illustrates the application of the statistical methods when there is plurality in y in observations with common κ , when response y i is binary: $y \in \{0,1\}$ - and variability in x

take place. Application of MLM(Maximum Likelihood Method) characterize this case of plurality in observations of binary response y.

And, at last, in minmax approach varying part of the conditions is considered to consist of two parts:

$$x_i = (x_i^{(1)}, x_i^{(2)}) \ , i = \overline{1, N},$$

while each of observations treats to be $(x_i^{(1)}, y_i), i = \overline{1, N}$, or more precisely $((x_i^{(1)}, f), y_i), i = \overline{1, N}$, with $x_i^{(2)} \in X_2$, where X_2 is known.

MoFS approach or modification of the ClasFsS within the plural model of uncertainties is interpreted just as the generalized logit- and probit- regression. Indeed, accordingly to basic statistical interpretation theorem from [Donchenko, 2005] membership function $\mu^{(P)}(e), e \in E$ has the uncertainty object P. So, MoFS observations may be treated as

$$(e_i, y_i), i = 1, N : e_i \in E, y_i = \begin{cases} 1, \text{when P is observed} \\ 0, \text{when P is not observed} \end{cases}.$$

Collections of the MoFS: situation

MoFS definition of the a fuzzy set as a pair $(E, \mu^{(P)}(\cdot)), \mu : E \to [0,1]$ with the membership function $\mu^{(P)}(\cdot)$ being the function of two arguments with one of them fixed ($P \in \wp$) let the problem of classification or clusterization to be considered: ascription each of elements $e \in E$ to the one of the K classes, described by predicates $P_k \in \wp, k = \overline{1, K}$ from a MoFS collection $(E, \mu_k^{(P_k)}(\cdot)), P_k \in \wp, k = \overline{1, K}$ with $\mu_k^{(P_k)}(\cdot) : \sum_{k=1}^{K} \mu_k^{(P_k)}(e) \underset{\geq}{\leq} 1,$

such collection may be complete:

$$\forall e \in E \sum_{k=1}^{K} \mu_k^{(P_k)}(e) = 1, \qquad (1)$$

as well as incomplete: with strong inequality.

Predicates collection $P_k \in \wp, k = \overline{1, K}$ from MoFS collection $(E, \mu_k^{(P_k)}(\cdot)), P_k \in \wp, k = \overline{1, K}$ may be interpreted as a collection of the alternatives which may take place for each of the elements of $e \in E$ with some probabilities, described by $H \mu_k^{(P_k)}(e)), e \in E, k = \overline{1, K}$.

In the ClasFsS the values of membership functions may be considered classically: as a confidence (certainty) functions. Actually, in this case the collection (list) of the objects of uncertainty (alternatives) are to be described, we will say, that MoFS collection $(E, \mu_k^{(P_k)}(\cdot)), P_k \in \wp, k = \overline{1, K}$, describe the situation for the elements $e \in E$ or the situation made concrete by $e \in E$. Ascription the $e \in E$ under consideration to the one of K classes, described by predicates $P_k \in \wp, k = \overline{1, K}$, we will call the classifying the situation.

Classifying the situation conception

When interpreting MoFS collection as a situation, each of the $e \in E$ make the situation concrete while membership functions $\mu_k^{(P_k)}(e), k = \overline{1, K}$ for each of the fix $e \in E$ evince the "degree of appearance" for the $P_k \in \wp, k = \overline{1, K}$. Such conception of the situation make it possible to estimate the situation for that or this $e \in E$ by the maximum (or minimum) of the confidence in each of $P_k \in \wp, k = \overline{1, K}$ for fixed $e \in E$ and

correspondingly to make ascription the element $e \in E$ to one of the classes, determined by $P_k \in \wp, k = \overline{1,K}$. Such approach to classifying the situation namely realizes the idea embodied in MLM (Maximum Likelihood Method). It is reasonable to remark that in MLM this idea is realized in a posteriori form: when having the observations.

Definition. Function $\hat{P}(e), e \in E, \hat{P} : E \to \{P_1,...,P_K\}$, determined on the common E of the MoFS collection $(E, \mu_k^{(P_k)}(\cdot)), P_k \in \wp, k = \overline{1,K}$ (situation) by the relation:

$$\hat{P}(e) = P_{k*}, k* = \arg\max_{k=\overline{1,K}} \mu_k^{(P_k)}(e), e \in E \tag{2}$$

said to be situation estimation for $e \in E$.

Remark 2. Generally speaking situation estimation may be plural: when maximum in (1) reached simultaneously for some k $k \in \{1,...,K\}$. In this case $\hat{P}(e), e \in E$ turned out to be plural: $\hat{P}(e) \subseteq \{P_1,...,P_K\}, e \in E$ and determined by the modification of the relation (2):

$$\hat{P}(e) = \{P_k : k \in K* = \operatorname{Arg}\max_{k=\overline{1,K}} \mu_k^{(P_k)}(e), e \in E\}. \tag{3}$$

Remark 3.Just as in (1) or (2) situation is estimated to be "best", it may be estimated to be worst. In this case "min" is substituted instead "max" in (1) or (2).

Remark 4. Term "situation estimation" by no means do not restrict "classification" character of the $\hat{P}(e), e \in E$, when $P_k, k = \overline{1,K}$ define classes and $\mu_k^{(P_k)}(\cdot)), k = \overline{1,K}$ define the membership probabilities for each $e \in E$.

In this case $\hat{P}(e), e \in E$ said to be classifying function. Note that classes are not necessarily mutually exclusive. Under exclusive alternatives the condition (1) is natural.

Remark 5. Situation estimation according (1) or (2) with "max" or "min" demonstrates that when operating with "fuzzy logic" operations it is reasonable to take into account the arguments on which the result of operation is reached but no the result of the operation by itself.

Situation Model example: clustering the distributions, probe sets (a priori data)

The classification problem for several distributions demonstrates the "classifying situation" approach proposed earlier. More precisely, let observed $e \in R^m$ may represent one of the K distributions $P^{(k)}(B)$,B –Borel set in $R^m, k = \overline{1,K}$ or, equally, may be the value of the one of the random (multivariate) variable (r.v.) $\varepsilon_k, k = \overline{1,K} : P^{(k)}(B) = P\{\varepsilon_k \in B\}, k = \overline{1,K}$,B – Borel set in R^m. One can consider the values of the distributions on "probe" sets $e + \pi$, $e \in E$, for appropriate fixed Borel π. As a π ta ball $S_\rho(0)$ for the fixed radius $\rho > 0$ can be considered. Obviously $\mu_{k,\rho}^{(P_k)}(e), e \in R^m, k = \overline{1,K}$, determined by the relations

$$\mu_{k,\rho}^{(P_k)}(e) = P^{(k)}(S_\rho(e)) = P^{(k)}(e + S_\rho(0)) = P\{\varepsilon_k \in e + S_\rho(0)\}, k = \overline{1,K}, e \in R^m, \tag{4}$$

are membership functions for MoFS with the "crisp"-predicates P_k : "to have distribution $P^{(k)}$". $k = \overline{1,K}$.

Also π one can be chosen to be symmetric, close convex set $V_\rho(0)$ with fixed radius $\rho > 0$: $e + \pi = e + V_\rho(0)$, $e \in E$. Last set denoted to be $V_\rho(e)$:

$$V_\rho(e) = e + V_\rho(0)\,. \tag{5}$$

So, the collection $(R^m, \mu_{k,\rho}^{(P_k)}(e)), k = 1, K$ is the MoFS collection with the "crisp" predicates P_k :" property to have distribution $P^{(k)}$ ", $k = \overline{1,K}$,- which describes the situation.

The balls $S_\rho(e) = e + S_\rho(0)$, $\rho > 0$ centered in $e \in R^m$ in a natural way probe the distributions. The results are represented by the MoFS membership functions $\mu_{k,\rho}^{(P_k)}(e)), k = 1, K$.

Remark 6. It is advisable to say that in the example under consideration memberships functions of the MoFS collection, which describe the situation are obviously statistically transparent. Ana in a addition standart statistical representation accordingly to theorem 2 above from for example [Donchenko, 2005] turned out to be of the next form:

$$\mu_{k,\rho}^{(P_k)}(e)) = P\{\varepsilon_k \in S_\rho(\xi) \mid \xi = e\}, k = \overline{1,K} \tag{6}$$

With R^m -valued vector fandom variable (r.v.) ξ determined on the probability space common with the collection $\varepsilon_k, k = \overline{1,K}$, independent of them and having the distribution to be nonsingular.

Indeed, we have:

$$P\{\varepsilon_k \in S_\rho(\xi) \mid \xi\} = M\{\chi_{S_\rho(\xi)}(\varepsilon_k) \mid \xi\} = M\{\chi_{S_\rho(0)}(\varepsilon_k - \xi) \mid \xi\} =$$

$$= M\{\chi_{S_\rho(0)}(\varepsilon_k - e)\Big|_{e=\xi} = P\{\varepsilon_k - e \in S_\rho(0)\}\Big|_{e=\xi} = P\{\varepsilon_k \in S_\rho(e)\}\Big|_{e=\xi}\,.$$

So, indeed, relation (6) gives the example of the standard (universal) statistical representation for collection of the ClasFsS or MoFS membership function accordingly to theorem 2 from the [Donchenko, 2005] has been already mentioned above.

Turning back to the MoFS membership functions (6) from the collection, defined the situation, we note that $\hat{P}(e), e \in R^m$ from (2) or (3) denoted as $\hat{P}_\rho(e), e \in R^m$ below are defined by one of the next two relations:

$$\hat{P}_\rho(e) = P_{k*} : k* = \arg\max_{k=\overline{1,K}} \mu_{k,\rho}^{(P_k)}(e), e \in R^m\}\,, \tag{7}$$

$$\hat{P}_\rho(e) = \{P_k : k \in K* = \operatorname{Arg}\max_{k=\overline{1,K}} \mu_{k,\rho}^{(P_k)}(e), e \in R^m\}\,. \tag{8}$$

So, classifying the situation accordingly to (7) or (8) obviously turned out to be the "maximum likelihood" estimation by P_k - probabilities for ρ - neighborhood of $e \in R^m$.

Remark 7. It is interesting to note that alternatives in (8) are not exclusive.

Situation Model example, continue: limit by the infinitive decreasing the measures of probe sets

Classifying the situation accordingly (7) or (8) shows transparent statistical (probabilistic) content. It is classification by maximum probability of the distributions collection under consideration on the "probe" set, neighboring $e \in E$. The items relating the classifying problem under "decreasing" the the "probe" set. It is naturally to normalize the membership functions in some way. It is turned out, that the classifying situation algorithm reduces the problem to the classification by the maximum of density functions for the distribution under consideration.

Indeed, let the distributions $P_k, k=1,K$ have the continues densities $h_k(z), z \in R^m, k=\overline{1,K}$, and the "size" ρ of the "probe" sets $S_\rho(e)=e+S_\rho(0)$ or $V_\rho(e)=e+V_\rho(0)$ is infinitely decreasing: $\rho \to 0$. As it has been mentioned about membership functions are to be normalized in some way. There are two variants for normalization: one for m=1 and another for the general case: for m>1. It is $\rho^{-1}, \rho>0$ in the first case (m=1) and inverse of Lebeague measure ϑ in R^m on the probe set in another (m>1). It is naturally to demand of "non singularity" for the density functions $h_k(z), k=1,K$ of distributions $P_k, k=1,K$."Non singularity" is understudied as non zero value of the distributions on the ""probe" sets for all $\rho>0$.

Then the next pairs of statements take place.

Theorem 3. Let the distribution functions $h_k(z), z \in R^m, k=\overline{1,K}$ of distributions $P_k, k=1,K$ are continuous and non singular.

Then

$$\lim_{\rho \to 0} \rho^{-1} \mu_{k,\rho}^{(P_k)}(e)) = h_k(e) \,||\, \mathrm{grad}_z h_k(e) \,||, e \in R^1, k=\overline{1,K}, \tag{9}$$

$$\lim_{\rho \to 0} \{\vartheta(S_\rho(0))\}^{-1} \mu_{k,\rho}^{(P_k)}(e)) = h_k(e), e \in R^m, m>1, \ k=\overline{1,K}. \tag{10}$$

Similar result take place for MoFS membership functions built by the "probe" sets $V_\rho(e)=e+\rho V$ with symmetric, convex, close set with radius equal to one, i.e. with the collection of the MoFS membership functions of the next type:

$$\mu_{k,\rho}^{P_k}(e) = P^{(k)}(V_\rho(e)) = P^{(k)}(e+\rho V) = P\{\varepsilon_k \in e+\rho V\}, e \in R^m, k=\overline{1,K}. \tag{11}$$

Theorem 4. Let conditions of theorem 3 take place and $P_k(V_\rho(e))>0, k=1,K$. Then there is exist $\varphi: 0<\varphi \le 1$ such, that:

$$\lim_{\rho \to 0} \rho^{-1} \mu_{k,\rho}^{(P_k)}(e)) = \varphi h_k(e) \,||\, \mathrm{grad}_z h_k(e) \,||, e \in R^1, k=\overline{1,K}, \tag{12}$$

$$\lim_{\rho \to 0} \{\vartheta(\rho V_1(0))\}^{-1} \mu_{k,\rho}^{(P_k)}(e)) = h_k(e), e \in R^m, m>1, \ k=\overline{1,K}. \tag{13}$$

Classifying the situation: "a priori maximum likelihood method"

Relations (9)-(10), (12)-(13) shows straight connection between situation classifying in the model example and classification by the density functions, namely, between classifying by the probabilities of distributions $P_k, k=1,K$ on the "probe" sets and classification by density functions of correspondent distributions.

So, let for the MoFS collection $(R^m, \mu_{k,\rho}^{(P_k)}(\cdot)), k=\overline{1,K}$ there exist limits $\lim_{\rho \to 0} \rho^{-1} \mu_{k,\rho}^{(P_k)}(e))$, denote it by $d_k(e)$, for $E=R^1$ or limit $\lim_{\rho \to 0} \{\vartheta(\rho V)\}^{-1} \mu_{k,\rho}^{(P_k)}(e))$ also denoted by $d_k(e)$, for some symmetric, convex close set V with radius equal 1.

Definition. Function $\hat{P}(e), e \in E, \hat{P}: E \to \{P_1,...,P_K\}$, defined by MoFS collection $(R^m, \mu_{k,\rho}^{(P_k)}(\cdot)), k=\overline{1,K}$, by the one of the relations:

$$\hat{P}_{\infty}(e) = P_{k*}, k* = \arg\max_{k=\overline{1,K}} d_k(e), e \in R^m, \quad (14)$$

$$\hat{P}_{\infty}(e) = P_{k*}, k* = \arg\max_{k=\overline{1,K}} d_k(e) \,||\, \mathrm{grad}_z h_k(e) \,||, e \in R^m, \quad (15)$$

said to be classifying the situation by the limit of normalized membership functions.

Model example shows the interpretation of the classifying algorithm represented by (14) or (15).

Theorems 3-4 show the relation between maximum probabilities classification and classifying the situation for the MoFS collection on the model example. It is reasonable to recollect here similar classification procedure; main-shift classification (see, for example, [Comaniciu, 2002]). In the mean-shift algorithm classification s implemented by the density functions generated for each of the classes by the observations of learning sample, i.e. that one may be called “a posteriori maximum likelihood method”.

Conclusion

In the paper modified variant of fuzziness from [Donchenko, 2005] (see also [Donchenko, 1998-3], [Donchenko, 1998-4] and [Donchenko, 2004]), is considered to develop the conception proposed by the ideas of “situation” and “classifying the situation”. The algorithm for “classifying situation” is proposed having natural probability content. This probability content is illustrated by the model example based on the idea of “probe” sets. Limit behavior of the objects in the model example establish the relation between “classifying the situation” and “a priori likelihood method”. “A priori likelihood method” is proposed to be extended on “classifying situation” construction. Similar classification algorithm: mean-shift method – is recollected having the same idea, but a posteriori character.

References

[Comaniciu, 2002] .D. Comaniciu. A robust approach towards Feature space analysis.//IEEE Transactions on Pattern Analysis and Machine Intelligence. – V.24,№5, May 2002.– p. 603-617.

[Donchenko, 1998-3] Donchenko V.S. Conditional distributions and Fuzzy sets. // Bulletin of Kiev University. Series: Physics and Mathematics, №3, 1998. (In Ukrainian)

[Donchenko, 1998-4] Donchenko V.S. Probability and Fuzzy sets. // Bulletin of Kiev University. Series: Physics and Mathematics, №4,1998. (In Ukrainian)

[Donchenko, 2004] Donchenko V.S. Statistical models of observations and Fuzzy sets. // Bulletin of Kiev University. Series: Physics and Mathematics, №1, 2004. (In Ukrainian)

[Donchenko, 2005]. Donchenko V. Fuzzy sets: Abstraction axiom, Statistical Interpretation. Observations of Fuzzy Sets. //International Journal “Information Theories and Applications”.– V.13, №3.– 2005.–p.233-239.

[Kaufmann, 1982]. Introduction to Fuzzy subset theory [Рус. Перевод Кофман А. Введение в теорию нечетких множеств.- Г.: Радио и связь. 1982.- 322 с.

[Kaufmann and Gupta, 1991] Arnold Kaufmann and Madan Gupta. Introduction to Fuzzy Arithmetic. - Thomson Computer Press, 1991.

[Zadeh, 1965] . Zadeh, Lotfi, Fuzzy Sets.// Information and Control, 8(3). June 1965. pp. 338-353.

Authors' Information

Donchenko, Volodymyr – Professor, Kyiv National Taras Shevchenko University, Cybernetics Faculty, Systems analysis and Decision Making Department; e-mail: vsdon@unicyb.kiev.ua, voldon@unicyb.kiev.ua.

THE INFLATION INDEX PROGNOSIS BASED ON THE METHOD OF DECISION-MARKING "TREE"

Alexei Voloshyn, Victoria Satyr

Abstract: *The description of the support system for marking decision in terms of prognosing the inflation level based on the multifactor dependence represented by the decision – marking "tree" is given in the paper. The interrelation of factors affecting the inflation level – economic, financial, political, socio-demographic ones, is considered. The perspectives for developing the method of decision – marking "tree", and pointing out the so-called "narrow" spaces and further analysis of possible scenarios for inflation level prognosing in particular, are defined.*

Keywords. *Method of decision - marking "tree", multifactor analysis, inflation index, expert information.*

Introduction

Economic growth is considered to be one of the most important social problems which is in the focus of economists' and politicians' attention. So the importance of inflation index prognosing means that the main tendencies of economic development are reflected in it and the charges in the dynamics of Gross Domestic Product (GDP) affect the changes of living standard of each citizen of Ukraine. There exist no universal and perfect approaches to finding solution to this problem nowadays. Some attempts for building possible scenarios for developing either phenomenon in future only have been made. Various method of qualitative characteristics based on using expert information are used for this purposes.

Economic prognosis should define and evaluate the main directions of economic development reflecting total combination of internal and external interrelations between the components on the macroeconomic level in the first place. Macroeconomic prognosis means that the investigation should be aimed at strategic level. All the main elements of economic and social sphere in progress should be considered in the context of their cause and effect relations and interdependence.

Irrespective of the aim any macroeconomic prognosis is based on definite theoretical grounds to correspond with scientific grounds. In terms of mechanisms of achieving adequacy and information unity the apparatus of economic mathematic modeling becomes indispensable because of using complex economic-mathematic methods based on apparatus of econometric modelling too often. The methods of quantitative prognosing (time lines, regressive analysis, imitation modelling, etc.) based on "continuation if the past" present poor results while prognosing "unstable" processes which are characterized by "breaking the monotony" and are based on sudden changes and cannot be referred to as distinctive features for describing the development of the process in the past [Popov, 1996]. The problem lies in representing the future as a usual thing related to continuation of the past as far as the future can acquire some principally new shares. This prognosing ("qualitative prognosing") is based on direct usage of human (expert) knowledge, inaccuracy of expert information being taken into account in the first place, which depends on the expert's professional and psychological characteristics (competency, independence, impartiality, real vision, risk taking, etc.) [Ivchenko, 1984].

The method of decision - marking "tree"

The method of decision - marking "tree" represented in this paper may be referred to as a basic ground for prognosing inflation level, expert information being used for building the tree itself and method of double comparison [Voloshin, 1999]. Expert information can be presented both as determined and inaccurate one. In processing expert information for finding out "collective" estimates the algebraic method based on using Hemming's metrics and measure of nonconformity of object ranks [Ushakov, 1979]. The group with $n(n \geq 1)$ of experts working together singles out the problems and sub-problems and builds the decision - marking tree, and also defines the importance (priority) of each task (each element of the tree). Processing expert information is

carried out taking into account “priorities” set by expert and degree of agreement on their estimates [Kozeletskyy, 1979]. The basic factors are presented at the top of the tree. Then these factors fall into smaller sub-problems, etc. As a result the decision - marking tree is built. The leaves of the tree stand for the factors which do no fall into further sub-problems. After heaving set the priorities in the decision - marking tree the importance of each factor is evaluated [Seber, 1980].

The tree is built by a group of experts (persons who make decisions). Each of the elements (key points) of the tree (expert of the leaves) has its sub-elements (i.e. each problem has a sub-problem). Then the priorities (chances) for coming from one top of the tree to another one are set.

The work of the expert group results in building the decision - marking tree aimed at prognosing the inflation index (Figure 1) which incorporates the following main problems:

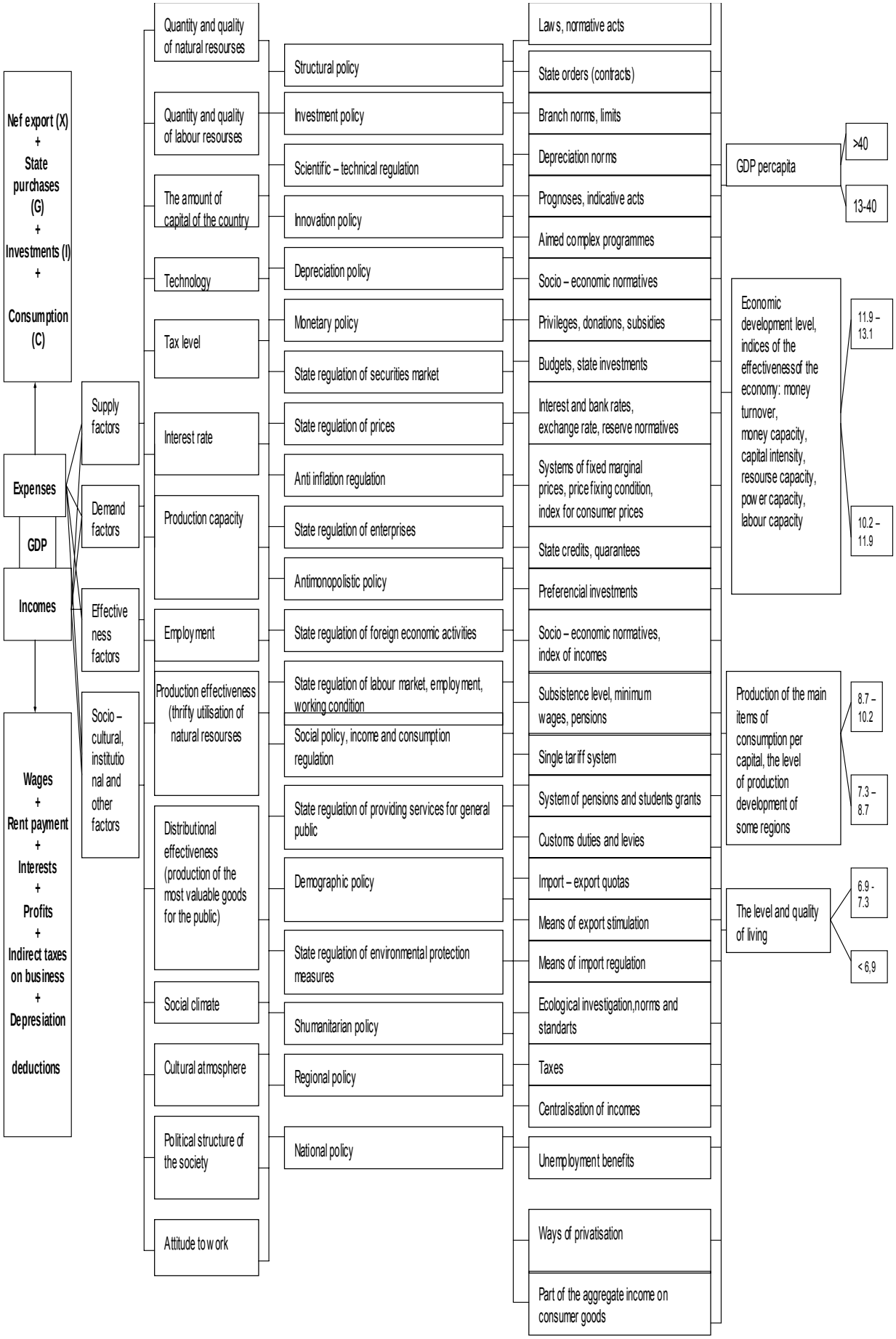
- economic - industry, agroindustrial complex, financial market, commerce, etc;
- political - irregular economy, investment, monetary and antimonopolistic policy, etc;
- social – demographic situation – unemployment, social-demographic load, tempo of population increment, etc;
- financial – financial-budget, monetary policy, sate regulation of securities market, state regulation of prices, etc.

To specify “narrow” spaces means to single out the affecting factors at each stage of the sub-problem, for instance, let’s consider the state of economic development: state orders (contracts), privileges, subsidies, state credits, warrants and taxes, internal and external investments, the allotment of the aggregate income expenses on consumer goods. Let’s consider the problem of unemployment, the number of jobs available, the system of allotment of pensions and students grants, state regulation of labour market. Employment and working conditions, etc. Having considered the importance the leaves of this tree we obtain the probable specific indices of inflation fir a certain period of time. Each of the experts produces estimates of three types: a_1^i - “optimistic”, a_2^i - “realistic”, and a_3^i - “pessimistic”. The resulting estimate is calculated in such a way. The average estimate of each expert $a_i = \frac{a_1^i \cdot \gamma_1 + a_2^i \cdot \gamma_2 + a_2^i \cdot \gamma_3}{\gamma_1 + \gamma_2 + \gamma_3}$, is considered in the first place, then with reference to the importance of all the experts, the resulting estimate is calculated. In this case: $a_1^j, j = \overline{1,3},\ i = \overline{1,n}$ is introduced indistinctly by means of function (vector of essential quality and looks like: $a_j^i = (a_{1j}^1,...,a_{kj}^i) j = \overline{1,3},\ i = \overline{1,n}$. Coefficients $\gamma_1, \gamma_2, \gamma_3$ are calculated empirically and present numerical characteristics [Tsurkov, 1981].

In accordance with one of the method $\gamma_1 = \gamma_3 = 1, \gamma_2 = 4$ (for expert-“realist”), according to other - $\gamma_1 = 3,\ \gamma_2 = 0,\ \gamma_3 = 2$ (for an “optimist”) and $\gamma_1 = 2,\ \gamma_2 = 0,\ \gamma_3 = 3$ (for a “pessimist”). To specify the psychological type of the expert (pessimist, optimist, realist) by means input in the system the psychological testing of the experts is carried out and then the coefficients of “realisticity” $k_1^i = (1/3 \le \lambda \le 2/3$ for realist, $0 \le \lambda \le 1/3$ for pessimist, $2/3 \le \lambda \le 1$ for optimist correspondingly are taken into account). The coefficients of competency k_2^i are calculated on the basis of the previous prognoses accuracy in conformity with the methods suggested. The initial coefficients k_2^i equal $\frac{1}{n}$ [Gladun, 1987].

The theory

To analyse the way the decision - marking tree can respond it is necessary to find the facts which dramatically affect the inflation level and could be taken into consideration in this model. Bearing this in mind we should single out the “narrow” spaces which depend greatly on being placed on a certain leaf of the tree – the estimate of the prognosed parameter.


Nef export (X)
+
State purchases (G)
+
Investments (I)
+
Consumption (C)
Expenses
GDP
Incomes
Wages
+
Rent payment
+
Interests
+
Profits
+
Indirect taxes on business
+
Depresiation
deductions
Supply factors
Demand factors
Effective ness factors
Socio – cultural, institutio nal and other factors
Quantity and quality of natural resourses
Quantity and quality of labour resourses
The amount of capital of the country
Technology
Tax level
Interest rate
Production capacity
Employment
Production effectiveness (thrifty utilisation of natural resourses
Distributional effectiveness (production of the most valuable goods for the public)
Social climate
Cultural atmosphere
Political structure of the society
Attitude to w ork
Structural policy
Investment policy
Scientific – technical regulation
Innovation policy
Depreciation policy
Monetary policy
State regulation of securities market
State regulation of prices
Anti inflation regulation
State regulation of enterprises
Antimonopolistic policy
State regulation of foreign economic activities
State regulation of labour market, employment, working condition
Social policy, income and consumption regulation
State regulation of providing services for general public
Demographic policy
State regulation of environmental protection measures
Shumanitarian policy
Regional policy
National policy
Laws, normative acts
State orders (contracts)
Branch norms, limits
Depreciation norms
Prognoses, indicative acts
Aimed complex programmes
Socio – economic normatives
Privileges, donations, subsidies
Budgets, state investments
Interest and bank rates, exchange rate, reserve normatives
Systems of fixed marginal prices, price fixing condition, index for consumer prices
State credits, quarantees
Preferencial investments
Socio – economic normatives, index of incomes
Subsistence level, minimum wages, pensions
Single tariff system
System of pensions and students grants
Customs duties and levies
Import – export quotas
Means of export stimulation
Means of import regulation
Ecological investigation,norms and standarts
Taxes
Centralisation of incomes
Unemployment benefits
Ways of privatisation
Part of the aggregate income on consumer goods
GDP percapita
>40
13-40
Economic development level, indices of the effectiveness of the economy: money turnover, money capacity, capital intensity, resourse capacity, pow er capacity, labour capacity
11.9 – 13.1
10.2 – 11.9
Production of the main items of consumption per capital, the level of production development of some regions
8.7 – 10.2
7.3 – 8.7
The level and quality of living
6.9 - 7.3
< 6,9

After this it is desirable to consider several scenarios for describing different switchovers in the tree. Under such conditions there is a possibility to obtain current statistic data for the model parameters to be estimated and using them to test statistic and economic effectiveness of the model and develop the more effective prognosis. To decide on what factors should be chosen for the model of this kind the theory of economic indicators might be of great importance. Commonly all indicators are interrelated and affect each other so inflation itself may be referred to as an economic indicator and there exist a number of indicators closed related to it. The wide range of indicators allows them to be grouped into certain system or models in conformity with the requirements specific for them.

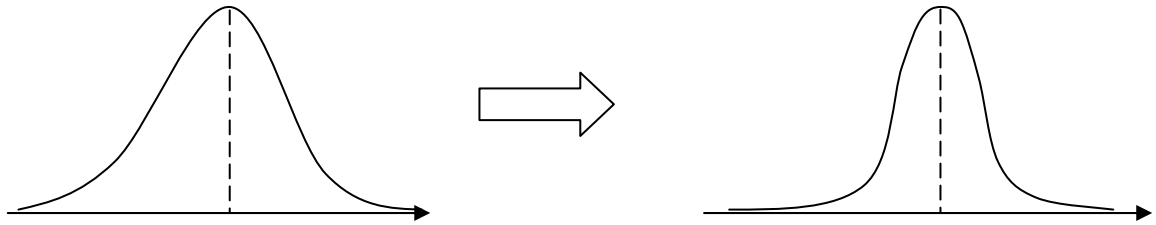
One of the methods of changing economic situation is combining economic indicators into a system in terms of their economic contents and the nature of statistic interrelation with inflation. There economic indicators can be grouped though certain elements in them can overlap. Among these indicators there may be distinguished such of them which characterise economy and economic growth, population and employment, state fiscal policy, consumption, investments industry and commerce, external cash flows, exchange rates, money and interest rates and salaries and wages. While considering indicators of each group as separate elements of the system and comparing their dynamics with the dynamics of GDD changes we may single out such of them which are directly related to inflation level and those which are not. The indicators which tend to be the components of inflation belong to the first group. They are: part of the aggregate income expenses on consumer goods, privileges, donations and subsidies, etc. In this context it is possible to define one more criterion of their relation to inflation whether they change together with economic situation or with the positive or negative lag.

There are the moments which form the basis of the statistic prognosing indicators system which is used to prognoses the inflation level. So, the indicator approach gives an opportunity to take into consideration those critical moments in the tree and the influence of different by their nature indices on the common result.

It is also by means of other methods to change the data obtained from the experts according to their characteristics. Let i- expert give the approximate estimate $a^j = (a_1^j, ..., a_k^j)$ of the problem that might become one of the tops of the decision – making tree. If for example, the expert is not risk taking enough it is advisable to process the obtained estimate in such a way [Zagoruiko, 1999].

The maximum $a_j^i \max$ and the minimum $a_j^i \min$ values are calculated then the average value a_j^i is calculated. Any a_j^i is changed in such a way: $a_j^i = a_j^i - \frac{\left(a_j^i - \min\right)\cdot\left(\max - a_j^i\right)}{\left(\max - \min\right)^2}$.

This changer will change the type of the function essential quality: (Figure 2.)



Judging from this figure we may arrive to the conclusion that this kind of transformation makes the function more vividly expressed. This may be understood as increasing the degree of expert's risk taking.

Besides, in accordance with the task it is also possible to perform certain transformations in terms of increasing expert's realisticity, independence, etc.

The main advantages of such approach are:

- It allows to solve effectively the tasks of technological forecast doing the research of unstable processes and phenomena with insufficient description as far as they are based on the experts knowledge and do not depend on the information abut the behaviour of these phenomena in the past;
- It allows the structuring of colleting expert information dividing the subject area into certain segments which makes it possible to select more highly specialized and hence more skilled experts/

Conclusion

This system was used for prognosing the inflation index on 1. 01. 2005, the system was used in July, 2005, the inflation index obtained was equal 12.8%, the prognosed estimates of the official institutes and external experts were fluctuating between 8% to 20% and plus. The official statistic data of the ministry of Economy of Ukraine come up to 10,5% and the Institute for Economics and Prognosing suggested 12,5% - 13%. Taking into consideration the level of irregular economy the second figure seems much more realistic.

Bibliography

[Popov, 1988] E. V. Popov Expert systems. – M.: Nauka, 1988. 183p

[Popov, 1988] E. V. Popov, I. B. Fominykn, E. B. Kisel. Statistic and dynamic expert systems: Educational material – M.: Finance and statistics, 1996. 320p.

[Ivchenko, 1984] G. I. Ivchenko, I. Medvedev. Mathematical statistics/ M.: Publishing House Vysshaya Shkola, 1984. 248p.

[Seder, 1980] J. Seber. Linear regressive analysis. – M.: Mir, 1980. 456p.

[Voloshin, 1999] A. F. Voloshyn, M. B. Panchenko, E. P. Pikhotnik. The expert system of supporting the prognosing hrivnia exchange rate Artificial intellect. – 1999. - №1. Pp. 354–359.

[Ushakov, 1979] I. A. Ushakov the tasks of optimal reserving. – M.: Znaniye, 1979. 343p.

[Tsurkov, 1981] U. I. Tsurkov Decomposition in great demensions sums – M.: Science, 1981. 352p.

[Koseletskyy, 1979] Yu. P. Koseleletskyy. Psychological theory of making decisions. – M.: Progress, 1979. 504p.

[Gladun, 1987] V. P. Gladun Decision Planning. – Kiev: Naukova dumka, 1987. 167p.

[Zagoruiko, 1999] N. G. Zagoruiko. The applied methods of analyzing data and Knowledge. – Novosibirsk: Publ. House NM SB RAN.

Authors' Information

Voloshyn Alexei Fiodorovich – Kiev National University "Taras Shevchenko", Faculty of Cybernetics, professor, Kiev, Ukraine; e-mail: voloshin @unicyb.Kiev.ua

Satyr Viktoria Valeriivna – Kiev National University "Taras Shevchenko"; Faculty Cybernetics, master of sciences, Kiev, Ukraine; e-mail: vicsatr@hotmail.com

SYNERGETIC METHODS OF COMPLEXATION IN DECISION MAKING PROBLEMS

Albert Voronin, Yury Mikheev

Abstract. *Synergetic methods of data complexation are proposed that make it possible to obtain a maximal amount of available information using a limited number of channels. Along with freedom degrees reducers, a mechanism of freedom degrees discriminators is proposed that enables all the channels to take part in the development of a cooperative decision in accordance with their informativeness in a current situation.*

Keywords: *Synergetics, data complexation, information channels, decision making*

Introduction

In advanced information systems, information on the same object (a process or an event) is usually transmitted over *several* channels. The problem lies in determining the channels over which more significant data are transmitted. Depending on this, it is required to combine (integrate) obtained data to develop a cooperative decision on the state of an object.

Taking into account a role of information accuracy in constructing the present-day decision-making systems the problem considered in this article should be regarded as topical.

Analysis of the problem state

The synergetic conception of data complexation is actively applied to the extraction of a maximum of useful information from an available collection of the various data that characterize a process or an object in various application domains.

As an example we shall cite a problem of the height of a plane estimation using indications on barometric, onboard and ground radar-tracking gadgets and, probably, on the visual channel. Each of the specified channels has its advantages and lacks in various flight conditions. It is required to combine (integrate) the obtained data for the most authentic height estimation in a current situation.

In the monograph [1], the problem of integration of the devices having different accuracy class indications is considered. Each of the devices brings its mite in the resulting indication according to its accuracy class. Also, the problem of integration of experts estimates here is put and solved in view of the different experts competence in a case in point.

In paper [2], an automatic classification method of the state of forests is described that is based on a satellite data map and the synergetic merging of data principle. The most informative (dominant) spectral channels of the sensor being used are detected, and the sought-for decision is taken upon their evidences.

In [3], the problem of integration of signals from navigating fields of different physical natures (radio-navigation fields such as GPS, geophysical fields, the field of stars and bodies of Solar system, etc.) is formulated for the most authentic estimation of the current coordinates of a space vehicle.

In [4], a method of complexation of signals for bi-static radar-location of small celestial bodies is described. To increase the accuracy of measurements in investigating parameters of motion of small celestial bodies, a bi-static configuration of radar-tracking systems is used. Data from each of the receiving antennas spaced on sizable distances are processed and compared among themselves so that the resulting signal is most reliable.

In the given examples, the concept of synergetics [5,6] – the science about cooperative processes is used. In the hierarchy of systems theories, synergetics occupies the upper level. As against the general systems theory, synergetics studies and organizes the processes running not under centralized actions but due to collective components interaction according to the result in view. The cooperation of components makes it possible to use reserve capabilities of a system and substantially increases the system effect degree.

By the definition of P.K. Anokhin, "we can call a system only such a complex of selectively involved components that their interaction and interrelation assume the character of *mutual assistance* of components that is oriented toward a fixed useful result" [7]. The stated fundamental property of mutual assistance is a synergetic process that is clearly expressed and everywhere manifests itself in biological systems [5].

During the synthesis of a synergetic functional system, redundant freedom degrees (Ashby law [8] about a requisite variety), should first be created that determine additional capabilities in the properties of the future system, and then they are reduced according to the dominant mechanism during the functioning of the system [5]. To achieve this end, “reducers of freedom degrees” are introduced into the system being synthesized with the help of a special control law.

Synergetic concept of complexation (confluence) of the data is actively applied to extraction of the maximal information from available set of the various data not only in biology, but also in other object domains to what the given examples testify.

Substantial analysis of the problem

In contrast to biological and similar synergetic control systems, the complexation systems do not contain, as a rule, redundant channels of data gathering. The number of freedom degrees is a priori limited, and the heart of the problem consists of obtaining a maximum amount of available information under these constraints. There exist two approaches to the problem. In the above examples, the action of "reducers of freedom degrees" has led

to the truncation of low-informative channels and to the selection of one or several most informative (dominant) in the current situation data acquisition channels on the basis of which the sought-for decision has been formed.

In applying this approach, some useful nuances contained in the truncate channels do not participate in the process of searching for the decision, i.e. some information items are lost. Figuratively speaking, one or several dominant “soloists" whose sounding does not contain the overtones that attach particular significance to a musical performance are artificially selected from the entire ensemble of data.

At the second approach, it is advisable to abandon the conception of a dominant and, instead of "reducers of freedom degrees" to include mechanisms that allow all channels of data acquisition to participate in the formation of the sought-for decision with weights corresponding to their informativeness degree in the current situation ("discriminators of freedom degrees”). As a result, all the available information items will be properly used and the "sounding" of the data ensemble will be harmonious and volumetric. Both approaches have their pluses and minuses and both are applied in practice.

Statement of the problem

It is given: quantity of data channels (number of freedom degrees in synergetic complexation system) $m \geq 3$. The array of initial data is represented in the form of the column matrix

$$A^T = \| \alpha_1 \; \alpha_2 \; ... \; \alpha_m \; \|, \tag{1}$$

where $\alpha_j, j \in [1,m]$ – the data on some numerical value a, received on j-th channels (components of the complexation system).

It is required to determine the most authentic estimation a^* of the value a.

Method of solution

If the number of channels is great enough and it is known, that their self-descriptiveness degrees are approximately identical, the problem is solved by simple averaging channels data as the maximum likelihood estimator:

$$\alpha^* = \frac{1}{m}\sum_{j=1}^{m}\alpha_j \; .$$

The need for increase of estimation reliability arises, when the number of data channels is small, and the relative degree of trust to them is different and not known beforehand. Hereinafter we shall apply, for example, the first approach to the complexation problem.

In this case, for the solution of a problem in view we shall take advantage of the “freedom degrees reducers” mechanism. The iterative synergetic process of adaptive mutual assistance of system components is organized.

Since the channel that deserves to be more believed is not yet known, we first assume that the degree of belief to all the channels data is the same and, in averaging them, their data are taken with one coefficient $k_j^I = 1, \; j \in [1,m]$.

As a result of averaging, the following mean estimate is obtained:

$$\alpha^I = \frac{1}{m}\sum_{j=1}^{m} k_j^I \alpha_j = \frac{1}{m}\sum_{j=1}^{m} 1 \cdot \alpha_j = \frac{1}{m}\sum_{j=1}^{m}\alpha_j .$$

We call it the estimate of the first iteration. The operation of averaging in matrix form is the multiplication of the matrix - column of the data from the left by a unit m-row matrix (summing vector)

$$E = \|1 \quad 1 \quad \ldots \quad 1\|$$

and division of the product obtained by the channels number :

$$\alpha^I = \frac{1}{m} EA.$$

We now have information on the mean estimate α^I and can compare it with estimates of each channel α_j from matrix (1). Of course, the difference between the mean estimate (the opinion of the majority) and the estimate proposed by a channel can form the basis for the change in the weight coefficient with which the "opinion" of the channel is taken into account. For those channels, whose estimates on the first iteration are closer to the mean estimate, it is expedient to increase the coefficient k_j, and, on the contrary, this coefficient should be decreased for the channels whose estimates are far different from the mean estimate. In our procedure, the relatively rare cases when the opinion of the minority is "true" are omitted.

Let us introduce the following measure (the "reducers of freedom degrees")

$$\delta_j^{II} = \left| \alpha^I - \alpha_j \right|, j \in [1,m],$$

that is the quantitative representation of the trust degree of the *j*-th channel at the second iteration. It make sense to select coefficients k_j^{II} such that they would be functions inversely proportional to δ_j^{II} :

$$k_j^{II} = c / \delta_j^{II}, \; c = \text{const}, \tag{2}$$

under the condition

$$\sum_{j=1}^{m} k_j^{II} = m. \tag{3}$$

Solving the system of equations (2) and (3), we eliminate the unknown coefficient of proportionality c and obtain

$$k_j^{II} = \left(\frac{m}{\delta_j^{II}}\right) / \sum_{t=1}^{m} \left(\frac{1}{\delta_t^{II}}\right).$$

Then we perform the averaging operation at the second iteration but, in this case, take into account the different trust to channels according to the results of the first iteration,

$$\alpha^{II} = \frac{1}{m} \sum_{j=1}^{m} k_j^{II} \alpha_j. \tag{4}$$

Introducing a row matrix

$$K^{II} = \left\| k_1^{II} \quad k_2^{II} \quad \ldots \quad k_j^{II} \quad \ldots \quad k_m^{II} \right\|,$$

we represent expression (4) in matrix form

$$a^{II} = \frac{1}{m} K^{II} A.$$

At the third iteration, the measure

$$\delta^{III} = \left|\alpha^{II} - \alpha_j\right|, j \in [1, m],$$

is first established and so on.

The iterative procedure

$$a^{(g)} = \frac{1}{m} K^{(g)} A, g \in [1, h], K^1 = E$$

continues until the condition of termination

$$\left|\alpha^{(h)} - \alpha^{(h-1)}\right| \le \varphi$$

is false, where φ is a given small quantity. The result of the iterative procedure described is the obtaining of a refined estimate $a^* = a^{(h)}$ determined with due regard for the heterogeneity of channels. In practical cases, the iterative process converges after 3-4 iterations and the most informative channel is determined.

Synergetic aspects of mathematical statistics

The synergetic principle of data complexation has much in common with ideas of mathematical statistics [9]. Really, when the synergetic conception of data merging is applied to the most authentic estimation of characteristics of processes (objects) from an available data set, then the mathematical statistics studies methods for the most authentic estimation of moments of distribution of random quantities from an available set of sample units. The commonality of problems of both theories testifies to the topicality of the problem of investigating synergetic aspects of mathematical statistics both for synergetics, and also for the development of statistical methods. In order to illustrate the second approach, here we shall provide the mechanism of "discriminators of freedom degrees" for the problem solving.

Let us consider a problem of the information processing at the limited number of data channels as calculation of the best (in a sense) estimate θ^* of the unknown distribution parameter θ of the random quantity X with probability density $f(x|\theta)$ on the basis of a limited statistical material $x = x^{(n)} = (x_1, x_{2,}, ..., x_n)$ – analogue of freedom degrees of a data integration synergetic system.

An efficient instrument of increasing the efficiency of statistical estimation is the Bayesian approach [9]. The aprioristic information that the unbiased estimate of parameter θ, assumed as the random quantity, is distributed under the same law as X is used. Minimization of the risk function for the square-law loss function gives the expression for an optimum estimate as the posteriori mathematical expectation of parameter θ, calculated on to the given vector of observations:

$$\theta^* = \int_{-\infty}^{+\infty} \theta f(\theta|x) d\theta \Big|_{x=x^{(n)}}. \quad (5)$$

Let us make use of the posteriori density definition under Bayes theorem [9]:

$$f(\theta|x) = \frac{f(x|\theta) f_a(\theta)}{f(x)},$$

where the normalizing marginal distribution is expressed by the formula

$$f(x) = \int_{-\infty}^{+\infty} f(x|\theta) f_a(\theta) d\theta.$$

Then expression (5) will be transformed to a kind

$$\theta^* = \left. \frac{\int_{-\infty}^{+\infty} \theta f(\theta|x) f_a(\theta|\theta') d\theta}{\int_{-\infty}^{+\infty} f(\theta|x) f_a(\theta|\theta') d\theta} \right|_{\bar{x}=\bar{x}(n)}, \quad (6)$$

where θ' – an unknown constant. Since the sought-for estimate must be computed from a given vector of observations, we should pass in expression (6) from integrals to summation over the elements of this sample and replace unknown constants by their estimates:

$$\theta^* = \frac{\sum_{i=1}^{n} x_i f(x_i|\theta^*) f_a(x_i|\theta^*)}{\sum_{i=1}^{n} f(x_i|\theta^*) f_a(x_i|\theta^*)}. \quad (7)$$

Formula (7) expresses dependence of the quantity θ^* on itself,

$$\theta^* = \varphi(x_1, x_2, ..., x_n; \theta^*),$$

As is well known [10], the equation in such a form can be solved by an iterative method. The iterative procedure is organized according to the recurrent formula

$$\theta^*[l] = \varphi(x_1, x_2, ..., x_n; \theta^*[l-1]), l \in [1, L],$$

and iterative process terminates when the condition

$$\theta^*[L] - \theta^*[L-1] \le \lambda_\theta$$

becomes true, where l - number of the current iteration and L is the number of iterations; λ_θ – a preassigned accuracy of computation of the sought-for estimate. If the questions of convergence should be analyzed, then we can use the well-known theorem [10] according to which, to provide convergence of the iterative process, it is sufficient that the following inequality be true in the considered interval of refinement of the estimate θ^*:

$$\left| d\varphi(x_1, x_2, ..., x_n; \theta^*) / d\theta^* \right| < 1.$$

The general expression for refined estimate (7) fully complies with the following idea of Gauss [11]. Most probable is such a value of parameter being estimated that minimizes the sum of squares of differences between the actually observable and computed values multiplied by the weight coefficient k_i that reflects the relative confidence in observations:

$$\theta^* = \arg\min_{\theta^*} \sum_{i=1}^{n} k_i (x_i - \theta^*)^2. \qquad (8)$$

In [12,13], it is shown, that expression (7) really is obtained from (8) if the posteriori probability density ("discriminators of freedom degrees") is introduced in the capacity of the measure of relative confidence in observations.

Thus, the methodology proposed provides the individual approach to each realization of a random quantity (weighing in accordance with the posteriori probability of its occurrence), which makes it possible [14] to avoid the information loss in computing the sought-for estimates from a small sample.

It is important to note that an estimate is elaborated by means of the organization of an iterative process in which sample units adaptively interact among themselves during each iteration. Similarly, synergetics provides a process characterized by self-control and self-organization according to the objective formulated. Adaptable processes are developed owing to the collective interaction of components. The cooperation of components activates reserve capabilities of a system and considerably increases the extant of system effect.

Bibliography

1. A.N. Voronin, Ju.K. Ziatdinov, A.V. Khartchenko, Complex Engineering and Ergatic Systems: Methods of Investigation [in Russian], Fact, Kharkov (1997).
2. V.I. Ljal'ko, A.D. Fedorovskij, M.A. Priests, et al., "Using satellite data for studying problems of nature resources", in: Space Exploration in Ukraine (2002-2004), NSAU, Kiev (2004), p.p. 7-14.
3. I.D. Varlamov, D.V. P'yaskovskij and S.V. Vodop'yan, "Adaptive correlation-extremal algorithm of a space vehicle navigation over geophysical fields on the basis of differential Taylor transformations", Space Science and Technology, No. 4, 141-146 (2001).
4. A.N. Voronin, "A method of signals complexation for bistatic radiolocation of small celestial bodies", in: 9-th International Conference "Systems analysis and control", Izd. MAI, Moscow (2004), p.p. 113-114.
5. A.A. Kolesnikov, Synergetic Control Theory [in Russian], Energoatomizdat, Moscow (1994).
6. H. Haken, Sinergetics [Russian translation], Mir, Moscow (1980).
7. P.K. Anokhin, Essays on Physiology of Functional Systems [in Russian], Medicine, Moscow (1975).
8. W.R. Ashby, An Introduction to Cybernetics [Russian translation], Izd Inostr. Lit., Moscow (1959).
9. D.R. Cox , D.V. Hinckley, Theoretical statistics [Russian translation], Mir, Moscow (1978).
10. R.S. Guter and P.T. Resnikovskij, Programming and Computational Mathematics [in Russian], Nauka, Moscow (1971).
11. A. Seige and J. Mells, Estimation Theory and its Application to Communication and Control [Russian translation], Svyaz', Moscow (1976).
12. A.N. Voronin, "Increasing the efficiency of statistical estimates of ergatic systems parameters", Cybernetics and computer technology, No. 50, 29-31 (1980).
13. A.N. Voronin, "On the rise of efficiency of statistical estimates for parameters of ergatic systems", Zentralblatt fur Mathematik und ihre Grenzgebiete, Mathematics Abstracts, 484, Berlin. Heidelberg. New York, (1983), p. 375.
14. D.V. Gaskarov, V.I. Shapovalov, Small samples [in Russian], Statistics, Moscow, (1978).

Authors' Information

Albert N. Voronin – National aviation university, faculty of computer information technologies, Dr.Sci.Eng., professor; 03058, Kiev - 58, Kosmonavt Komarov avenue, 1, Ukraine

Jury I. Mikheev – Zhitomir military institute of radioelectronics, adjunct; 10004, Zhitomir, Mir avenue, 22, Ukraine

OPERATING MODEL OF KNOWLEDGE QUANTUM ENGINEERING FOR DECISION-MAKING IN CONDITIONS OF INDETERMINACY

Liudmyla Molodykh, Igor Sirodzha

***Abstract**: The operating model of knowledge quantum engineering for identification and prognostic decision-making in conditions of α-indeterminacy is suggested in the article. The synthesized operating model solves three basic tasks: A_t-task to formalize tk-knowledge; B_t-task to recognize (identify) objects according to observed results; C_t-task to extrapolate (prognosticate) the observed results. Operating derivation of identification and prognostic decisions using authentic different-level algorithmic knowledge quantum (using tRAKZ-method) assumes synthesis of authentic knowledge quantum database (BtkZ) using induction operator as a system of implicative laws, and then using deduction operator according to the observed tk-knowledge and BtkZ a derivation of identification or prognostic decisions in a form of new tk-knowledge.*

***Keywords**: operating model, decision-making object, knowledge quantum database, target feature, method of different-level algorithmic knowledge quantum, implicative law.*

***ACM Classification Keywords**: I.2.3 Deduction and Theorem Proving; I.2.4 Knowledge Representation Formalisms and Methods; I.2.5 Programming Languages and Software*

Introduction

Knowledge-oriented modelling of human being's intellectual skills to make decisions in conditions of indeterminacy to recognize patterns and prognostic situations for artificial intelligence systems (AIS) is being developed in the article. Operating model for knowledge quantum engineering for decisions derivation in conditions of indeterminacy, which is based on using the **method of authentic different-level algorithmic knowledge quantum or portions (tRAKZ-method)** is suggested. The existing artificial neural networks (ANN) and knowledge engineering methods, based on frame, production and other knowledge models, are not effective enough because of the imperfection of representation ways and computer knowledge manipulation. Unlike these approaches the suggested model has a form of strictly formalized knowledge quantum, different in the level of complexity (tk-knowledge). Such tk-knowledge as substantial algorithmic structures of authentic data allow computer manipulation of knowledge using an finite predicates algebra and vector-matrix operators, and also inductive synthesis of **knowledge quantum database (BtkZ)** while teaching computer using selective plot examples of situations from the concrete data domain.

1. Target setting

The model-based process of **human's classification** and **prognostic** decision-making in conditions of *indeterminacy* is always aimed (motivated by a target criterion) at the **decision-making object** (**DMO**), which can be described with a set of characteristics (features), measured in different scales and allowing logical representation. **Target** features are also contained in this set. Their values determine the **class** and **pattern** of the considered **DMO**. To **identify** the class (pattern) of DMO, i.e. to **make a classification decision**, means to define a value of the **target feature** according to the observed initial characteristics, relying on the **knowledge quantum database** (BtkZ), represented by **classification law** systems. Analogically to make a **prognostic decision** it is necessary to have a **prognostic BtkZ**, allowing to define the value of the **target prognostic feature** on the segment **t+Δt**, according to the situation on the time segment **t**.

The discussed **α-indeterminacy** is characterized by such limitations:

- data about DMO are of different type (i.e. measured in quantitative as well as in qualitative scales) and can be reached in incomplete volumes and from different sources (experts, technical documentation, reference books, instruments measurements etc.);

- the target criteria are given implicitly, it is unknown which ones, in what quantity and how to select informative features of DMO according to targets of decision-making;
- the rules of making classification and prognostic decisions are unknown, and also the inductive principles of their building by teaching on selective experimental data are unknown too;
- the sought rules of decision-making are impossible to be defined by regular calculus of approximations directly, but it is possible to create knowledge engineering tools to model and imitate intellectual human's skills to find solutions, relying on intuition and knowledge database.

In α-indeterminacy the **authentic** k-knowledge (**tk-knowledge**) are used.

The main task of this article is to create a method of synthesis for operating model in knowledge quantum engineering to derive classification and prognostic decisions in conditions of α-indeterminacy. In general this task is deducted to solving three basic tasks [Sirodzha, 2002]:

1. **A_t-task** for formalization of tk-knowledge;
2. **B_t-task** for object recognition (identification) according to observation results;
3. **C_t-task** for extrapolation (prognostic) of observation results.

In the **A_t-task** it is required to define the terms "**tk-knowledge**" and "**tRAKZ-models**" formally in conditions of **α-**indeterminacy, to describe their algorithmic design using quantum structuring of different-type data about DMO considering its semantics in a concrete data domain.

A_t-task is described formally using the multiple four:

$$A_t = \langle S, K_t, \Pi_t, Q_t \rangle \qquad (1)$$

and consists in building the class M_t of substantial algorithmic structures and operating tools for manipulating them on a character language S from a set of letters, numbers, special symbols and algorithmic operations of algorithm theory on the basis of using rules for constructing t-quantum Π_t to terminal t-quantum from K_t with a help of finite set Q_t of semantic codes. Under semantic code $tk_s \in Q_t$ (s=0,1,2,…) we assume symbols, coding t-quantum, which corresponds the form and content of authentic knowledge of level s.

The **B_t-task** is to synthesize **recognizing** tRAKZ-models and algorithms to manipulate tk-knowledge to define values of **target characteristic** for the recognized DMO, i.e. its **identification** with the given reliability according to the external observations, relying on the preliminary cumulated BtkZ.

The **C_t-task** is to synthesize **prognostic** tRAKZ-models and algorithms for manipulation tk-knowledge to **predict** with the given reliability of **DMO permanent characteristics** values according to the measured values of the observed characteristics, relying on the preliminary built BtkZ.

To solve B_t- and C_t-tasks it is required:

1) **to synthesize the induction operator** $INDS\left(tk_2\Sigma_0; AZ; tk_2\overline{\Sigma_{BM}}\right)$ for inductive derivation of the sought BtkZ from a set of selected teaching tk-knowledge, where in brackets the parameters of INDS operator are shown: $tk_2\Sigma_0$ - teaching selective tk-knowledge of the 2nd level; AZ – operating algorithm of inductive derivation for BtkZ as new knowledge; $tk_2\overline{\Sigma_{BM}}$ - minimized BtkZ in the form of a matrix t-quantum of the 2nd level as a system of implicative laws.
2) **to synthesize the deduction operator** $DED(tk_2\Sigma_0; tk_1Y_\omega; AL; tk_sR)$ for deductive derivation of the sought decision as a new tk-knowledge of the level s (s=1,2) tk_sR in observations tk_1Y_ω for DMO ω, relaying on BtkZ= $tk_2\overline{\Sigma_{BM}}$, where AL is a deduction *algorithm*.

2. Algorithmic Formalization and Vector-Matrix Representation of tk-knowledge (A_t-task)

The general structure of **t-quantum of knowledge (tk-knowledge)** has two compounds: ***semantic*** and ***informational*** to represent a ***knowledge portion*** about DMO conditions in **semantic**, **informational** and **algorithmic** aspects at the same time. It is supposed that a portion (quantum) of knowledge about the DMO

condition describes some authentic quantum event (QE) in a production form "***message - consequence***" according to the scheme (2)

IF (logical combination of messages e_i), THEN (consequence C_j),
i =1, k; j = 1, h. (2)

Semantic compound of t-quantum in a form of **special structure of data** represents **meaning information** about this **QE**, showing the *scales for measuring* the DMO features, *semantic code* and quantum purpose as *knowledge model* about facts or laws. **Semantic code** from the set $\mathbf{Q_t}$ has a symbolic form $\mathbf{tk_sY_\omega}$, **k** is a quantum symbol; $\mathbf{s} \in \{0,1,2...\}$ is a level**, Y** is a name and $\omega \in$ **{p, tr, b, t...}** – quantum status (**p**recondition, **t**a**r**get, **b**asic, **t**erminal).

Information compound describes different-type features (characteristics) of DMO in a sectioned (domain) vector-matrix form, suitable to ***manipulate tk-knowledge*** and ***logical derivation*** using **computer algebra**. In a substantial and formal representation the domains $\mathbf{d_j}$ meet ***non-target*** (precondition) and ***target*** features of DMO, they are called **active** and are separated by a symbol **" : "**. Binary components of active domains $\alpha^i_j \in \mathbf{d_j}$ correspond to the features values. All the active domains define the QE logics, as far as a postulate is taken about the fact that active domains are connected with a **conjunction** (**":"** is a strap **"∧"**), the compounds in domains – with a **disjunction** (**","** is a strap **"∨"**), and precondition domains to a target – with an **implication** (**«⇒»**) in a form of (2). The logics of QE can be described in *sentential formulas* of ***propositional logic*** or in *finite predicates*, where the arguments are components of α_i^j domains.

The main idea of **strictly formalization** is in axiomatic building of tRAKZ-model on the basis of postulating the three **terminal** quanta $\mathbf{tk_1y_\tau}$, $\mathbf{tk_0a_\tau}$, $\mathbf{tk_1b_\tau}$ and using operators of **superposition (∏-operator)** known in the theory of algorithm, a **string concatenation (CON⟨•⟩-operator)** and a **column concatenation (CON[•]-operator).**

The *generalized* **terminal** quantum $\mathbf{tk_1y_\tau}$ represents a **vector of domains**, corresponding to different-type features $\mathbf{x_1, \dots, x_n}$ DME with values (in the domain components) from the finite sets $\mathbf{X^j}$, (**j**=1,2,..., **n**):: $X^1 = \{\alpha_1^1,...,\alpha_{r_1}^1\},...,X^n = \{\alpha_1^n,...,\alpha_{r_n}^n\}$, .The *generalized* quantum $\mathbf{tk_1y_\tau}$ has a form:

$$tk_1y_T = [\overset{x_1}{d_1} : \overset{x_2}{d_2} : \dots : \overset{x_n}{d_n}] = [\alpha_1^1,\dots,\alpha_{r_1}^1 : \alpha_1^2,\dots,\alpha_{r_2}^2 : \dots : \alpha_1^n,\dots,\alpha_{r_n}^n], \quad (3)$$

where $tk_1 \in Q_T$; name $y_T \in S_v$.

The *generalized* **terminal selecting** quantum $\mathbf{tk_0a_\tau}$ is described with a **selection function** $\mathbf{V_k^{(t)}}$ of the argument α_k from t-consequence of numbers or symbols:

$$\mathbf{tk_0a_\tau} = [V_k^{(\ell)}(\alpha_1,\dots,\alpha_k,\dots,\alpha_\ell)=\alpha_k],$$
$$\text{where } tk_0 \in Q; \quad \text{name } a_T, V_k^{(\ell)} \in S; \quad (4)$$

The *generalized* **terminal characteristic** quantum $\mathbf{tk_1b_\tau}$ is described with a characteristic function χ_{Yj} of a set $\mathbf{Y_j}$ for admissible values α_i^j of the j feature x_j:

$$\mathbf{tk_1b_\tau} = [\chi_{Y_j}(\alpha_k^j)] = \begin{cases} 1, \text{if } \alpha_k^j \in Y_j, \\ 0, \text{if } \alpha_k^j \notin Y_j, \end{cases} \qquad k = (1,2,...,r_j). \quad (5)$$

Definition 1. The different-level algorithmic structures, being received from terminal quantum $\mathbf{tk_1y_\tau}$(3), $\mathbf{tk_0a_\tau}$(4) and $\mathbf{tk_1b_\tau}$(5) with a help of finite number of applying **∏-operator**, **CON⟨•⟩-operator** and **CON[•]-operator**, are called **different-level algorithmic tk-knowledge** or **tRAKZ-models** of knowledge in conditions of **α-indeterminacy**, which form a class of authentic tRAKZ-models M_t.

In Fig.1 a quantum area $B_t^{(3)}$ of tRAKZ-model of DMO is shown, being described by three features: $\mathbf{x_1}$ with $\mathbf{r_1}$ = 2 values from $\mathbf{X^1} = \{\alpha_1^1, \alpha_2^1\}$; $\mathbf{x_2}$ with $\mathbf{r_2}$ = 4 values from $\mathbf{X^2} = \{\alpha_1^2, \alpha_2^2, \alpha_3^2, \alpha_4^2\}$ and $\mathbf{x_3}$ with $\mathbf{r_3}$ = 3 values from the set $\mathbf{X^3} = \{\alpha_1^3, \alpha_2^3, \alpha_3^3\}$.

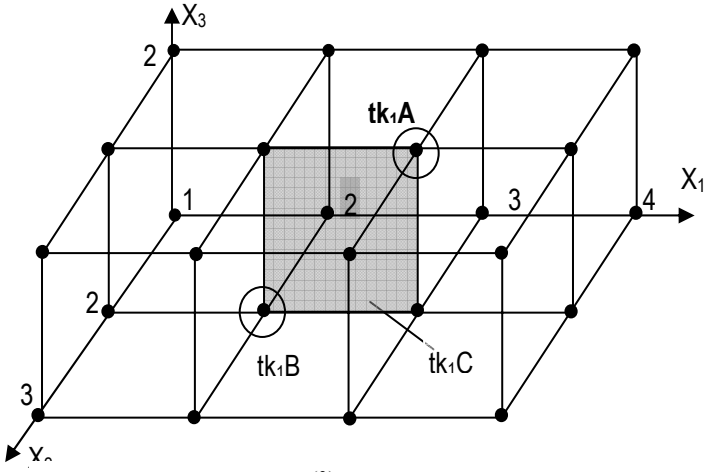


Fig.1. Area $B_t^{(3)}$ of **tRAKZ**-model

Vector domains are separated with a semicolon «:» and meet the different-type features of DMO, and components of domains – for the features values so that **i** component of **j** domain should contain «1», if we observe **i** value of **j** feature, otherwise **i** component equals to «0». If every domain of a **quantum of the 1st level** contains strictly only one «1», it is called an **element** one, otherwise it is called – an **interval** vector quantum. The points A and B of the area $B_t^{(3)}$ are responsible for element vector tk-knowledge $\mathbf{tk_1A}$ and $\mathbf{tk_2B}$:

$$tk_1A = \left[\overbrace{01}^{x_1} : \overbrace{0010}^{x_2} : \overbrace{010}^{x_3}\right], \qquad tk_1B = [10 : 0100 : 010], \tag{6}$$

The **interval C** $\subset B_t^{(3)}$ corresponds with an authentic **interval vector quantum of the 1st level**

$$tk_1C = [\overbrace{11}^{x_1} : \overbrace{0110}^{x_2} : \overbrace{010}^{x_3}], \tag{7}$$

which can be represented by a **matrix t-quantum of the 2nd level $\mathbf{tk_1C}$**, containing the joint 4 element vector t-quantum of the 1st level:

$$tk_2C = \begin{bmatrix} \overbrace{01}^{x_1} : \overbrace{0010}^{x_2} : \overbrace{010}^{x_3} \\ 10 : 0010 : 010 \\ 01 : 0100 : 010 \\ 10 : 0100 : 010 \end{bmatrix} \tag{8}$$

Besides, is t-quantum tk_1C (7) represents a conjunct, an elementary conjunction corresponds to it:

$$\left(x_1 \in \left\{\alpha_1^{(1)}, \alpha_2^{(1)}\right\}\right) \wedge \left(x_2 \in \left\{\alpha_2^{(2)}, \alpha_3^{(2)}\right\}\right) \wedge \left(x_3 \in \left\{\alpha_2^{(3)}\right\}\right) \tag{9}$$

The elementary conjunction (9) can be represented as a predicate equation:

$$\left(\left(x_1 = \alpha_1^{(1)}\right) \vee \left(x_1 = \alpha_2^{(1)}\right)\right) \wedge \left(\left(x_2 = \alpha_2^{(2)}\right) \vee \left(x_2 = \alpha_3^{(2)}\right)\right) \wedge \left(x_3 = \alpha_2^{(3)}\right) = 1 \tag{10}$$

So, the class $\mathbf{M_t}$ of **tRAKZ**-models represents a set of *uniform quantum tools* for describing **implicative laws**, and also different **facts** to represent them in the three equivalent forms: ***multiple*** (points, intervals of area $\mathbf{B_t^{(n)}}$); ***vector-matrix*** (domain structures); ***analytic*** (finite predicates).

3. Inductive search and deductive derivation of solutions as tk-knowledge

Under the **facts** we understand the *measured* DMO features of different type and their logical combinations, and also any *observed* events and situations, having relation to DMO and being represented by ***knowledge quantum*** of different levels, i.e. by **tRAKZ-models**. The tables **of empirical data** (TED) $\mathbf{T_o(m,n)}$ are typical examples of real facts.

Under the **laws** (DMO are subordinated to them) we consider **implicative** (***forbidden***) **logical connections** between *features* of **DMO**, they are rather **stable** to be defined while analyzing a limited TED $\mathbf{T_o(m,n)}$.

Definition 2. A **stable connection** between **r** characteristics of DMO from the general number of **n**, (**r≤n**), expressing ***inadmissibility*** of at least one combination of their values on a set of **tk-knowledge**, is called an **implicative law** or a **prohibition** of **r rank**.

In **tRAKZ**-method of decision-making the ***inductive derivation of tk-knowledge*** is used to build a general "**world model**" in a form of **BtkZ** as a range of ***implicative*** laws being found by ***learning tk-knowledge***, represented in a form of TED.

The ***deductive derivation*** of **tk-knowledge** is necessary to receive partials ***conclusions*** for the *observed* facts, basing on the BtkZ.

3.1. Inductive derivation operator of implicative BtkZ (INDS-operator)

The existence of implicative law as some forbidden knowledge quantum of s-level $tk_s\overline{Y}$ from $\mathbf{T_r}$, according to TED $\mathbf{T_0(m,N)}$, (**s**=1,2), is defined by the ***evaluation*** of its ***certainty***, satisfying the inequality

$$\mathbf{M_S\{m,N,r\}}=\frac{N!\cdot 2^{r(1-m)}\cdot(2^r-1)^m}{r!\cdot(N-r)!}\le M_S^* \qquad (11)$$

where the given **possible** *limit* **value** (*threshold*) of $\mathbf{M^*_S}$ [Sirodzha,1992] evaluation**.**

In a practical diapason of values **m** and **N** rank $\mathbf{r_{max}}$ turns out to be **small.** This allows defining all the **implicative** laws using a check for intervals «**forbiddances**» of a rank that is *not more than* $\mathbf{r_{max}}$. The disjunctive union of all the found forbidden intervals as conjunctions of combinations of informative features of DMO forms an analytic (predicate) description of the ***forbidden area***, corresponding BtkZ.

Definition 3. The algorithmic procedure

$$\mathrm{INDS}\left(tk_2\Sigma_0;AZ;tk_2\overline{\Sigma_{BM}}\right)=tk_2\Sigma_0\xrightarrow{\frac{INDS}{AZ}}tk_2\overline{\Sigma_{BM}}, \qquad (12)$$

implementing **inductive derivation** of non-odd **BtkZ=** $tk_2\overline{\Sigma_{BM}}$ in a form of a set of **simple** *prohibitions* from the learning knowledge quantum $\mathbf{tk_2\Sigma_0}$ using the algorithm **AZ**, is called an **operator of inductive derivation of implicative tk-knowledge (INDS-operator)** [Sirodzha, 1992].

Algorithm AZ

Input: TED in a form of quantum $tk_2\Sigma_0$ of size m×n, threshold $M^*_S=10^{-2}$, maximal rank $r_{max}=3$.

Output: minimized BtkZ = $tk_2\overline{\Sigma_{BM}}$ as a system of simple forbidden quanta, i.e. that do not result one from another.

Steps:

1. according to r_{max} patterns of features prohibitions combinations are formed. For $r_{max}=3$ there are 8 patterns: <000>, <001>, <010>, <011>, <100>, <101>, <110>, <111>. Forbidden combinations are searched between domains components, but not inside a domain.

2. In the cycle in $tk_2\Sigma_0$ all the combinations of features values are taken as doubles, and then as triples, etc. till r_{max}. The non-found in $tk_2\Sigma_0$ pattern combinations are added to $tk_2\overline{\Sigma_B}$.

3. The formed quantum of prohibitions $tk_2\overline{\Sigma_B}$ is ***minimized*** in BtkZ = $tk_2\overline{\Sigma_{BM}}$ using operators of gluing, merging and compression.

Let's assume that in the result of step 2 in the algorithm AZ we got a quantum $tk_2\overline{\Sigma_B}$. DMO is characterized by three features x_1, x_2, x_3.

$$tk_2\overline{\Sigma_B} = \begin{bmatrix} \underbrace{x_1}\ \underbrace{x_2}\ \underbrace{x_3} \\ 01-:-1:-1-- \\ 01-:0-:-1-- \\ -10:1-:---0 \\ ---:1-:---0 \\ 1--:-0:--1- \\ 0--:-0:--1- \end{bmatrix},$$ where «-» defines «it is indifferent if it is 0 or 1».

3.1. Gluing $(xy \vee x\bar{y} = x)$

$$tk_2\overline{\Sigma_B} = \begin{bmatrix} 01- : -1 : -1-- \\ 01- : 0- : -1-- \\ -10 : 1- : ---0 \\ --- : 1- : ---0 \\ \mathbf{1-- : -0 : --1-} \\ \mathbf{0-- : -0 : --1-} \end{bmatrix} \Rightarrow tk_2\overline{\Sigma_{B1}} = \begin{bmatrix} 01- : -1 : -1-- \\ 01- : 0- : -1-- \\ -10 : 1- : ---0 \\ --- : 1- : ---0 \\ \mathbf{--- : -0 : --1-} \end{bmatrix}$$

3.2. Merging $(xy \vee x = x)$

$$tk_2\overline{\Sigma_{B1}} = \begin{bmatrix} 01- : -1 : --0- \\ 01- : 0- : -1-- \\ \mathbf{-10 : 1- : ---0} \\ \mathbf{--- : 1- : ---0} \\ --- : -0 : --1- \end{bmatrix} \Rightarrow tk_2\overline{\Sigma_{B2}} = \begin{bmatrix} 01- : -1 : -1-- \\ 01- : 0- : -1-- \\ \mathbf{--- : 1- : ---0} \\ --- : -0 : --1- \end{bmatrix}$$

3.3. Compression *(union of quanta different with one domain only)*

$$tk_2\overline{\Sigma_{B2}} = \begin{bmatrix} \mathbf{01- : -1 : -1--} \\ \mathbf{01- : 0- : -1--} \\ --- : 1- : ---0 \\ --- : -0 : --1- \end{bmatrix} \Rightarrow tk_2\overline{\Sigma_{BM}} = \begin{bmatrix} \mathbf{01- : 01 : -1--} \\ --- : 1- : ---0 \\ --- : -0 : --1- \end{bmatrix}$$

After steps 3.1-3.3 under the whole forbidden quantum database we get the searched minimized implicative BtkZ= $tk_2\overline{\Sigma_{BM}}$.

3.2. Deductive derivation operator of decisions from implicative tk-knowledge.

It is necessary to solve the task of building the algorithm AL, implementing ***deductive*** *operating process* to search the *needed* decisions being correspondent with the ***logical consequence*** $tk_2\|Y\|, tk_1Y, tk_0\beta_{ik}^{(j)}$**:**

$$tk_2\overline{\Sigma_{BM}} \underset{AL1}{\overset{DED}{\Rightarrow}} tk_2\|Y\|, \quad tk_2\overline{\Sigma_{BM}} \underset{AL3}{\overset{DED}{\Rightarrow}} tk_1Y, \quad tk_2\overline{\Sigma_{BM}} \underset{AL2}{\overset{DED}{\Rightarrow}} tk_0\beta_{ik}^{(j)}, \qquad (13)$$

where $tk_2\overline{\Sigma_{BM}}$ is a known database of ***implicative*** **tk-knowledge**.

The searched sequences $tk_2\|Y\|$, tk_1Y $tk_0\beta_{ik}^{(j)}$ (13) represent the different-level tk-knowledge, characterizing the decisions being made in basic tasks $\mathbf{B_t}$ and $\mathbf{C_t}$ according to the observed results.

Let a base of implicative tk-knowledge $tk_2\overline{\Sigma_{BM}}$ is given and a quantum tk_1Y_ω of knowledge about the *observed* DMO $\omega \in \Omega$ of a data domain being investigated. The **algorithm AL** to evaluate the ***possible condition of DMO*** ω according to *quanta of observations* tk_1Y_ω, based on a ***known*** **BtkZ,** is a implementation of ***deductive derivation*** for the searched decision according to the scheme (13). Let's note that under the possible condition of DMO ω we understand a *class* or *pattern* and the DMO ω is concerned to it while solving the $\mathbf{B_t}$**-**task or a *category* (*value*) of prognosis connected with DMO ω if we solve the $\mathbf{C_t}$**-task**.

Algorithm AL

Input**:** tk-knowledge **BtkZ=** $tk_2\overline{\Sigma_{BM}}$ and observations tk_1Y_ω for DMO ω.

Output: deductively derived **tk**-knowledge $\mathrm{tk}_2\left\|\overline{Y^*_\omega}\right\|$ from **BtkZ** about the possible condition of DMO ω, according to the observations $\mathrm{tk}_1 Y_\omega$.

Steps:

1. To make a substitution of quantum values $\mathrm{tk}_1 Y_\omega$ in **BtkZ**= $\mathrm{tk}_2\overline{\Sigma_{BM}}$ in this way: to delete columns in a matrix quantum $\mathrm{tk}_2\overline{\Sigma_{BM}}$, meeting the features of the observed quantum $\mathrm{tk}_1 Y_\omega$.

2. To delete the rows, which are orthogonal to the observation $\mathrm{tk}_1 Y_\omega$ row, from the formed minor (respectively to the known features; 'orthogonal' means those having opposite in the meaning). In such a way we get $\mathrm{tk}_2\left\|Y^*_\omega\right\|$.

3. To invert the received quantum and consider it to be the result $\mathrm{tk}_2\overline{\Sigma}^*_\omega = \mathrm{tk}_2\left\|\overline{Y^*_\omega}\right\|$.

The algorithm is analogical for deriving the logical sequences $\mathrm{tk}_1 Y$, $\mathrm{tk}_0\beta^{(j)}_{ik}$ [Sirodzha, 2002].

Let's assume the DMO is characterized with 4 features (x_1,x_2,x_3,x_4), and the BtkZ has been inductively received in a form of:

$$\mathrm{tk}_2\overline{\Sigma_{BM}} = \begin{bmatrix} \overbrace{01-}^{x1} : \overbrace{-1}^{x2} : \overbrace{-1---}^{x3} : \overbrace{01}^{x4} \\ 01- : 0- : -1--- : 1- \\ -10 : 1- : ---0 : 1- \\ --- : 1- : ---0 : 0- \\ 1-- : -0 : --1- : -0 \\ 0-- : -0 : --1- : -1 \end{bmatrix}$$

There is also a quantum to observe the DMO $\mathrm{tk}_1 Y_\omega$=[001:10:0100:--]. It is required to define the possible value of the non-measured feature $\mathbf{x_4}$. According to the algorithm steps we get:

$\mathrm{tk}_1 Y_\omega$ = [001 : 10 : 0100 : --]

$$\mathrm{tk}_2\overline{\Sigma_{BM}} = \begin{bmatrix} 01- : -1 : -1-- : 01 \\ 01- : 0- : -1-- : 1- \\ -10 : 1- : ---0 : 1- \\ --- : 1- : ---0 : 0- \\ 1-- : -0 : --1- : -0 \\ 0-- : -0 : --1- : -1 \end{bmatrix} \Rightarrow \begin{bmatrix} 01- : -1 : -1-- : 01 \\ 01- : 0- : -1-- : 1- \\ -10 : 1- : ---0 : 1- \\ --- : 1- : ---0 : 0- \\ 1-- : -0 : --1- : -0 \\ 0-- : -0 : --1- : -1 \end{bmatrix}$$

After applying algorithm AL steps 1,2 a quantum [--- : 1- : ---0 : 0-] is left. After the inversion (step 3 of the algorithm AL) $\mathrm{tk}_0\beta^{(j)}_{ik}$=[1], i.e. the 4th feature (the 4th domain corresponds to it) takes the first value. Analogically the tasks to prognosis the several features values are being solved. In such a way the B_t-, C_t-tasks have been solved with a help of the algorithm AL.

Conclusion

Operating derivation of identification and prognostic decisions using tRAKZ-method suppose such a sequence of ***operating*** transformations of *different-level tk-knowledge*: using **induction operator** according to the given **table of empirical data** (TED) as learning tk-knowledge the **database of *authentic* knowledge quanta** (BtkZ) is synthesized. Then using **deduction operator** according to the observed (*input*) tk-knowledge of DMO, the searched ***identification*** or ***prognostic*** decisions are derived on the basis of BtkZ in a form of *resulting* tk-khowledge.

Operating method of decision derivation is based on the computer manipulation of vector-matrix structures (unlike the existing methods), that allows to abbreviate the time for BtkZ synthesis as a conclusive rule and to increase the efficiency of computer decision-making.

Bibliography

[Sirodzha, 2002] Sirodzha, I.B. Quantovye modeli I metody iskusstvennogo intellekta dlya prinyatiya reshenij I upravleniya. (Quantum models and methods of artificial intelligence for decision-making and management). Naukova dumka. – Kyiv: 2002. – 420 pp.

[Sirodzha,1992] Sirodzha, I.B. Matematicheskoe I programmnoe obespechenie intellektualnykh compiuternykh sistem. (Mathematical provision and programming software of intellectual computer systems.) – Kharkiv: KhAI,1992.

Authors' Information

Liudmyla Molodykh – a post-graduate student of Computer System Software Department, National Aerospace University named after N.I. Zhuckovsky "Kharkov Aviation Institute"; room 518, Impulse Building, Chkalova st., 17, Kharkiv, Ukraine, 61070; e-mail: molodykh@onet.com.ua; flamelia@mail.ru

Igor B. Sirodzha – Professor, Doctor of Technical Sciences, Head of Computer System Software Department, National Aerospace University named after N.I. Zhuckovsky "Kharkov Aviation Institute"; room 414, Impulse Building, Chkalova st., 17, Kharkiv, Ukraine, 61070.

CONSTRUCTING OF A CONSENSUS OF SEVERAL EXPERTS STATEMENTS*

Gennadiy Lbov, Maxim Gerasimov

***Abstract**: Let Γ be a population of elements or objects concerned by the problem of recognition. By assumption, some experts give probabilistic predictions of unknown belonging classes γ of objects $a \in \Gamma$, being already aware of their description $X(a)$. In this paper, we present a method of aggregating sets of individual statements into a collective one using distances / similarities between multidimensional sets in heterogeneous feature space.*

***Keywords**: pattern recognition, distance between experts statements, consensus.*

***ACM Classification Keywords**: I.2.6. Artificial Intelligence - knowledge acquisition.*

Introduction

We assume that $X(a) = (X_1(a),...,X_j(a),...,X_n(a))$, where the set X may simultaneously contain qualitative and quantitative features X_j, $j = \overline{1,n}$. Let D_j be the domain of the feature X_j, $j = \overline{1,n}$. The feature space is given by the product set $D = \prod_{j=1}^{n} D_j$. In this paper, we consider statements S^i, $i = \overline{1,M}$; represented as sentences of type "if $X(a) \in E^i$, then the object a belongs to the γ-th pattern with probability p^i", where $\gamma \in \{1,...,k\}$, $E^i = \prod_{j=1}^{n} E_j^i$, $E_j^i \subseteq D_j$, $E_j^i = [\alpha_j^i, \beta_j^i]$ if X_j is a quantitative feature, E_j^i is a finite subset of feature values if X_j is a nominal feature. By assumption, each statement S^i has its own weight w^i. Such a value is like a measure of "assurance".

Without loss of generality, we can limit our discussion to the case of two classes, $k = 2$.

* The work was supported by the RFBR under Grant N04-01-00858.

Distances between Multidimensional Sets

In the works [1, 2] we proposed a method to measure the distances between sets (e.g., E^1 and E^2) in heterogeneous feature space. Consider some modification of this method. By definition, put

$\rho(E^1,E^2)=\sum_{j=1}^{n}k_j\rho_j(E_j^1,E_j^2)$ or $\rho(E^1,E^2)=\sqrt{\sum_{j=1}^{n}k_j(\rho_j(E_j^1,E_j^2))^2}$,

where $0\le k_j\le 1$, $\sum_{j=1}^{n}k_j=1$.

Values $\rho_j(E_j^1,E_j^2)$ are given by: $\rho_j(E_j^1,E_j^2)=\frac{|E_j^1\Delta E_j^2|}{|D_j|}$ if X_j is a nominal feature, $\rho_j(E_j^1,E_j^2)=\frac{r_j^{12}+\theta|E_j^1\Delta E_j^2|}{|D_j|}$ if X_j is a quantitative feature, where $r_j^{12}=\left|\frac{\alpha_j^1+\beta_j^1}{2}-\frac{\alpha_j^2+\beta_j^2}{2}\right|$.

It can be proved that the triangle inequality is fulfilled if and only if $0\le\theta\le 1/2$.

The proposed measure ρ satisfies the requirements of distance there may be.

Consider the set $\Omega_{(1)}=\{S_{(1)}^1,...,S_{(1)}^{m_1}\}$, where $S_{(1)}^u$ is a statement concerned to the first pattern class, $u=\overline{1,m_1}$. Let E^u be the relative sets to statements $S_{(1)}^u$, $E^u\subseteq D$, $u=\overline{1,m_1}$. By analogy, determine $\Omega_{(2)}=\{S_{(2)}^1,...,S_{(2)}^{m_2}\}$, $S_{(2)}^v$, $\widetilde{E}^v$ as before, but for the second class.

By definition, put $k_j=\frac{\tau_j}{\sum_{i=1}^{n}\tau_i}$, where $\tau_j=\sum_{u=1}^{m_1}\sum_{v=1}^{m_2}\rho_j(E_j^u,\widetilde{E}_j^v)$, $j=\overline{1,n}$.

Consensus

We first treat single expert's statements concerned to a certain pattern class: let Ω be a set of such statements, $\Omega=\{S^1,...,S^m\}$, E^i be the relative set to a statement S^i, $i=\overline{1,m}$.

Denote by $E^{i_1i_2}:=E^{i_1}\oplus E^{i_2}=\prod_{j=1}^{n}(E_j^{i_1}\oplus E_j^{i_2})$, where $E_j^{i_1}\oplus E_j^{i_2}$ is the Cartesian join of feature values $E_j^{i_1}$ and $E_j^{i_2}$ for feature X_j and is defined as follows.

When X_j is a nominal feature, $E_j^{i_1}\oplus E_j^{i_2}$ is the union: $E_j^{i_1}\oplus E_j^{i_2}=E_j^{i_1}\bigcup E_j^{i_2}$.

When X_j is a quantitative feature, $E_j^{i_1}\oplus E_j^{i_2}$ is a minimal closed interval such that $E_j^{i_1}\bigcup E_j^{i_2}\subseteq E_j^{i_1}\oplus E_j^{i_2}$.

Denote by $r^{i_1i_2}:=d(E^{i_1i_2},E^{i_1}\bigcup E^{i_2})$.

The value $d(E,F)$ is defined as follows: $d(E,F)=\max_{E'\subseteq E\setminus F}\min_{j\|E'_j|\ne|E_j|}\frac{k_j|E'_j|}{diam(E)}$, where E' is any subset such that its projection on subspace of quantitative features is a convex set.

By definition, put $I_1=\{\{1\},...,\{m\}\}$, ..., $I_q=\{\{i_1,...,i_q\}\,|\,r^{i_ui_v}<\varepsilon\ \ \forall u,v=\overline{1,q}\}$, where ε is a threshold decided by the user, $q=\overline{2,Q}$; $Q\le m$.

Take any set $J_q=\{i_1,...,i_q\}$ of indices such that $J_q\in I_q$ and $J_q\not\subset J_{q+1}\ \ \forall J_{q+1}\in I_{q+1}$.

Now, we can aggregate the statements S^{i_1}, ..., S^{i_q} into the statement S^{J_q}:

S^{J_q} ="if $X(a) \in E^{J_q}$, then the object a belongs to the γ-th pattern with probability p^{J_q}", where

$E^{J_q} = E^{i_1} \oplus ... \oplus E^{i_q}$, $p^{J_q} = \dfrac{\sum_{i \in J_q} c^{iJ_q} w^i p^i}{\sum_{i \in J_q} c^{iJ_q} w^i}$, $c^{iJ_q} = 1 - \rho(E^i, E^{J_q})$.

By definition, put to the statement S^{J_q} the weight $w^{J_q} = \left(1 - d(E^{J_q}, \bigcup_{i \in J_q} E^i)\right) \dfrac{\sum_{i \in J_q} c^{iJ_q} w^i}{\sum_{i \in J_q} c^{iJ_q}}$.

The procedure of forming a consensus of single expert's statements consists in aggregating into statements S^{J_q} for all J_q under previous conditions, $q = \overline{1,Q}$.

After coordinating each expert's statements separately, we can construct an agreement of several independent experts for each pattern class. The procedure is as above, except the weights: $w^{J_q} = \sum_{i \in J_q} c^{iJ_q} w^i$.

Solution of Disagreements

After constructing of a consensus for each pattern, we must make decision rule in the case of contradictory statements. Take any sets $E^u_{(1)}$ and $E^v_{(2)}$ such that $E^u_{(1)} \bigcap E^v_{(2)} = E^{uv} \neq \varnothing$, where the set $E^u_{(\gamma)}$ corresponds to a statement $S^u_{(\gamma)}$ from the experts agreement concerned to the γ-th pattern class, $\gamma = 1,2$.

Consider the sets $I^{uv}_{(\gamma\}} = \left\{ i \mid (S^i \in \Omega_{(\gamma)}) \text{ and } (\rho(E^i, E^{uv}) < \varepsilon^*) \right\}$, where ε^* is a threshold, $0 < \varepsilon^* < 1$.

By definition, put $p^{uv}_{(\gamma)} = \dfrac{\sum_{i \in I^{uv}_{(\gamma)}} (1 - \rho(E^i_{(\gamma)}, E^{uv})) w^i p^i}{\sum_{i \in I^{uv}_{\{\gamma)}} (1 - \rho(E^i_{\{\gamma)}, E^{uv})) w^i}$. Denote by $\gamma^* := \arg\max_{\gamma}(p^{uv}_{(\gamma)})$.

Thus, we can make decision statement:

S^{uv} =" if $X(a) \in E^{uv}$, then the object a belongs to the γ^*-th pattern with probability $p^{uv}_{(\gamma^*)}$"

with the weight $w^{uv} = \left| \dfrac{\sum_{i \in I^{uv}_{(1)}} (1 - \rho(E^i_{(1)}, E^{uv})) w^i - \sum_{i \in I^{uv}_{(2)}} (1 - \rho(E^i_{(2)}, E^{uv})) w^i}{\sum_{i \in I^{uv}_{(\gamma^*)}} (1 - \rho(E^i_{(\gamma^*)}, E^{uv}))} \right|$.

Bibliography

[1] G.S.Lbov, M.K.Gerasimov. Determining of distance between logical statements in forecasting problems. In: Artificial Intelligence, 2'2004 [in Russian]. Institute of Artificial Intelligence, Ukraine.

[2] G.S.Lbov, V.B.Berikov. Decision functions stability in pattern recognition and heterogeneous data analysis [in Russian]. Institute of Mathematics, Novosibirsk, 2005.

Authors' Information

Gennadiy Lbov – Institute of Mathematics, SB RAS, Koptyug St., bl.4, Novosibirsk, Russia; e-mail: lbov@math.nsc.ru

Maxim Gerasimov – Institute of Mathematics, SB RAS, Koptyug St., bl.4, Novosibirsk, Russia; e-mail: max_post@bk.ru

APPLICATION OF THE HETEROGENEOUS SYSTEM PREDICTION METHOD TO PATTERN RECOGNITION PROBLEM[1]

Tatyana Stupina

Abstract*: An application of the heterogeneous system prediction method to solving the problem pattern recognition with respect to the sample size is considered in this paper. The criterion of multivariate heterogeneous variable recognition is used in this approach. The relation of this criterion with probability of error is shown. For the fixed complexities of probability distribution and logical decision function class the examples of pattern recognition problem are presented.*

Keywords*: the prediction of heterogeneous variables system, the pattern recognition, the complexity of distribution, logical decision function.*

ACM Classification Keywords*: G.3.*

Introduction

The reducing relation problem with respect to sample is one of important problem in data mining. The method quality of constructing sample decision function depends on the size of the sample, the complexity of the distributions, and the complexity of the class of functions used by the algorithm for constructing sample decision functions. When the distribution is known the quality of decision function, for example, is risk function for one prediction variable. For the pattern recognition problem, for example, it is well-known probability of error. We can define a quality of the method so as average of the quality of sample decision function on samples of fixed size. When the distribution is unknown the problem estimating of this function quality with respect to the complexity of the distributions, of the functions class and sample size is appeared. At present time there are approaches solving this problem [Vapnik V.N., Chervonenkis A.Ya, 1970]. In addition the complexity of the functions class is assigned differently [Lbov G.S., Starceva N.G, 1999]. The method quality we can estimate if the class of distribution is known or by mathematical modeling. But that approaches consider the case of one prediction variable and one variable type.

However there are many important applied when we what to predict or recognize several (system) heterogeneous variables. In work [Lbov G.S., Stupina T.A., 2002] was presented this problem statement. It is necessary to construct the sample decision function on the small sample in the multivariate heterogeneous space, so the most proper class is a class of logical decision functions [Lbov G.S., Starceva N.G, 1999]. In this paper for the fixed probability distribution the relation of the criterion with probability of error is shown for pattern recognition problem. The quality of constructing sample decision function method is shown for that problem.

Problem Statement

In the probabilistic statement of the problem, the value *(x,y)* is a realization of a multidimensional random variable (X,Y) on a probability space $<\Omega, \mathrm{B}, \mathrm{P}>$, where $\Omega = D_X \times D_Y$ is μ-measurable set (by Lebeg), B is the borel σ-algebra of subsets of Ω, P is the probability measure (we will define such as c, the strategy of nature) on B, D_X is heterogeneous domain of under review variable, $\dim D_x = n$, D_y is heterogeneous domain of objective variable, $\dim D_y = m$. The given variables can be of arbitrary types (quantitative, ordinal, nominal). For the pattern recognition problem the variable *Y* is nominal. Let us put $\Phi_\circ$ is a given class of decision functions. Class $\Phi_\circ$ is μ-measurable functions that puts some subset of the objective variable $E_y \subseteq D_Y$ to each value of

[1] This work was financially supported by **RFBR-04-01-00858**

the under review variable $x \in D_X$, i.e. $\Phi_\circ = \{f : D_x \to 2^{D_y}\}$. For example the domain E_y can contains the several patterns $\{\omega_1,...,\omega_k\}$ for pattern recognition problem.

The quality $F(c,f)$ of a decision function $f \in \Phi_\circ$ under a fixed strategy of nature c is determined as $F(c,f) = \int_{D_x} (P(E_y(x)/x) - \mu(E_y(x)))dP(x)$, where $E_y(x) = f(x)$ is a value of decision functions in x, $P(y \in E_y(x)/x)$ is a conditional probability of event $\{y \in E_y\}$ under a fixed x, $\mu(E_y(x))$ is measurable of subset E_y. Note that if $\mu(E_y(x))$ is probability measure, than criterion $F(c,f)$ is distance between distributions. If the specified probability coincides with equal distribution than such prediction does not give no information on predicted variable (entropy is maximum). On the nominal-real space $\Omega = D_н \times D_в$ a measure μ is defined so as any $E \in \mathrm{B}$, $E = \bigcup_{j=1}^{|E_н|} E_в^j \times \{z^j\}$, $\mu(E) = \sum_{j=1}^{|E_н|} \frac{\mu(E_в^i)}{|D_н|\mu(D_в)}$, were $E_н$ is projection of set E on nominal space $D_н$, z^j - item of $E_н$, $E_в^j$ - set in $D_в$ corresponding to z^j, $\mu(E_в^j)$ - lebeg measure of set $E_в^j$. For any subset of domains D_X or D_Y the measure μ is assigned similarly. Clearly, the prediction quality is higher for those E_y whose measure is smaller (accuracy is higher) and the conditional probability $P(y \in E_y(x)/x)$ (certainty) is larger. For a fixed strategy of nature c, we define an optimal decision function $f_\circ(x)$ as such as $F(c,f_\circ) = \sup_{f \in \Phi_\circ} F(c,f)$, where $\Phi_\circ$ is represented above class of decision functions.

When we solve this problem in practice the size of sample is very smaller and type of variables different. In this case is used class of logical decision function Φ_M complexity *M* [Lbov G.S., Starceva N.G, 1999]. For the prediction problem of the heterogeneous system variables class Φ_M is defined as $\Phi_M = \{f \in \Phi_\circ \,|\, f \sim <\alpha, r(a)>, \alpha \in \Psi_M, r(\alpha) \in R_M\}$ (the mark '~' denotes the correspondence of pair $<\alpha, r(a)>$ to symbol *f*), were Ψ_M is set of all possible partitioning $\alpha = \{E_X^1,...,E_X^M \,|\, E_X^t = \prod_{j=1}^n E_{X_j}^t,\ E_{X_j}^t \subseteq D_{X_j}, t = \overline{1,M}, \bigcup E_X^t = D_X\}$ of domain D_X on *M* noncrossing subsets, R_M is set all possible decisions $r(\alpha) = \{E_y^1,...,E_y^M \,|\, E_y^t \in \Im_{D_Y}, t = \overline{1,M}\}$, $\Im_{D_Y}$ - set of all possible *m*-measuring intervals. For that class the measure $\mu(E_y(x)) = \frac{\mu(E_y)}{\mu(D_Y)} = \prod_{j=1}^m \frac{\mu(E_{y_j})}{\mu(D_{y_j})}$ is the normalized measure of subset E_y and it is introduced with taking into account the type of the variable. The measure $\mu(E_y(x))$ is measure of interval, if we have a variable with ordered set of values and it is quantum of set, if we have a nominal variable (it is variable with finite non-ordering set of values and we have the pattern recognition problem). A complexity of Φ_M class is assigned as *M* if we have univariant prediction (decision is presented by form: if $x \in E_X^t$, than $y \in E_y^t$), $M_\Phi = M$, and it is assembly $(k_1,...,k_M)$ if we have multivariant, i.e. $E_y^t = \bigcup_{i=1}^{k_t} E_y^i$, $t = 1,...,M$ and $E_y^i \cap E_y^j = \varnothing$ for $i \neq j$ (decision is presented by form: if $x \in E_X^t$, than $y \in E_y^1 \vee E_y^2 \vee ... \vee E_y^{k_t}$). In this work for pattern recognition problem we consider the case $M_\Phi = M$.

Properties of the Criterion

For the fixed strategy of nature *c* the relation of the criterion $F(c,f)$ with probability of error P_f is shown.

Statement 1. For any strategy of nature *c* the quality criterion $F(c,f)$ is represented by risk function such that $1 - R(c,f) = \int_{D_X} \int_{D_Y} (1 - L(y,f(x)))p(x,y)dxdy$, where the loss function $L(y,f)$ such as $L(y,f) = \begin{cases} \mu(E_y), & y \in E_y \\ 1 + \mu(E_y), & y \notin E_y \end{cases}$.

Remark that risk function $R(c,f)$ is probability of error P_f if the loss function $L(y,f)$ is indicator function.

Statement 2. For recognition *k* patterns by decision function *f* the quality criterion is $F(c,f)=\frac{k-1}{k}-P_f$.

Consequence 1. For recognition two patterns we have equation $F(c,f)=\frac{1}{2}-P_f$.

Consequence 2. For pattern recognition problem the optimal decision function coincides with bayes function such as $\sup_{x\in(-\infty,+\infty)} F(c,f)=\inf_{x\in(-\infty,+\infty)} P_f$.

Definition 1. Define a nature strategy c_M (generated by logical decision function $f\in\Phi_M$, $f\sim<\alpha,r(\alpha)>$) such as $c_M=\{p^t(x,y)=p_x^t p_{y/x}^t=P(x\in E_X^t)P(y\in E_Y^t/x\in E_X^t),t=1,...,M\}$, where 1) $\sum_{t=1}^M p_x^t=1$; 2) $P(E_Y^t/E_X^t)=p_{y/x}^t$, 3) $P(\bar{E}_Y^t/E_X^t)=1-p_{y/x}^t$, where $E_X^t\in\alpha$, $E_Y^t\in r(\alpha)$, $<\alpha,r(\alpha)>\in\Phi_M$, 4) $\forall\ A_X\subseteq E_X^t\ P(A_X)=p_x^t\frac{\mu(A_X)}{\mu(E_X^t)}$, $\forall\ A_Y\subseteq E_Y^t\ P(A_Y/E_X^t)=p_{y/x}^t\frac{\mu(A_X)}{\mu(E_Y^t)}$.

In the paper [Lbov G.S., Stupina T.A., 2002] is proved that $F(c,f)=\sum_{t=1}^M p_x^t(p_{y/x}^t-\mu(E_y^t))$ for this nature strategy. Let for *k* pattern recognition the domain D_Y is the set $\{\omega_1,...,\omega_k\}$.

Statement 3. Let the nature strategy c_k for *k* pattern recognition is generated by logical decision function f^* such as $f^*(x)=E_y^i$ for $x\in E_X^i$, then the probability of error P_f for decision function *f* such as $f(x)=\omega_i$, $\omega_i\in E_y^i$, for $x\in E_X^i$ is $P_f=1-\sum_{i=1}^k\frac{1}{k\mu(E_y^i)}p_x^i p_{y/x}^i$.

Consequence 3. From the statement 3 it follows equation $P_f+F(c_k,f^*)=1+\sum_{i=1}^k p_x^i\left[p_{y/x}^i\left(1-\frac{1}{k\mu(E_y^i)}\right)-\mu(E_y^i)\right]$.

Let us illustrate these statements. Let there is *n=1, m=1*, *X*- continuous variable, *Y*-nominal variable, *M*=2. The nature strategy c_2 is generated by f^*, that $\alpha^*=\{E_X^1,E_X^2\}$, $E_X^1=(0.364,\ 1.0]$, $E_X^2=[0.0,\ 0.364]$, $r(\alpha^*)=\{E_y^1,E_y^2\}, E_y^1=\{1\}, E_y^2=\{1,0\}$, $\omega_1=$'1', $\omega_2=$'0', $p_x^1=\frac{19}{50}$, $p_x^2=\frac{31}{50}$, $p_{y/x}^1=0{,}95$, $p_{y/x}^2=1$. So we have $F(c_2,f^*)=0{,}171$. Obviously that c_2 is such that conditional distribution $p(x/\{1\})$ and $p(x/\{0\})$ is intersected for every pattern, if we have f' ($f'(x)=\{1\}$, if $x\in E_X^1$, and $f'(x)=\{0\}$, if $x\in E_X^2$) or f'' ($f''(x)=\{1\}$, if $x\in E_X^1$, and $f''(x)=\{1\}$, if $x\in E_X^2$). Let calculate $P_{f'}$ using definition and compare with criterion $F(c_2,f^*)$: $P_{f'}=P(\{1\})P(E_X^2/\{1\})+P(\{0\})P(E_X^1/\{0\})$ = $P(E_X^2)P(\{1\}/E_X^2)+P(E_X^1)P(\{0\}/E_X^1)$ = $\frac{31}{50}\cdot\frac{1}{2}\cdot 1+\frac{19}{50}\cdot(1-0{,}95)=0{,}329$. Similarly we can provide the probability of error $P_{f''}$. For this case we have $F(c_2,f^*)=\frac{1}{2}-P_{f'}=0{,}5-0{,}329=0{,}171$ (statement 2 and 3).

The method of Constructing Sample Decision Function

If the strategy of nature is unknown the sampling criterion $F(\bar{f})$ is used by presented method $Q(v_N)$ of constructing sample decision function $\bar{f}$, $\bar{F}(\bar{f})=\sum_{t=1}^{M'}\bar{p}_x^t(\bar{p}_{y/x}^t-\bar{\mu}_y^t)$, were $\bar{p}_x^t=\frac{N(\tilde{E}_X^t)}{N(D_X)}=\frac{N^t}{N}$, $\bar{p}_{y/x}=\frac{N(\hat{E}_y^t)}{N(\tilde{E}_X^t)}=\frac{\hat{N}^t}{N^t}$, $\bar{\mu}_y=\mu(\hat{E}_y)$, $N('*')$ is number of sample points, generating the set "*", $\bar{f}\sim<\alpha,r(\alpha)>$, $\alpha=\{\tilde{E}_X^1,...,\tilde{E}_X^{M'}\}\in\Psi_{M'}$, $r(\alpha)=\{\hat{E}_y^1,...,\hat{E}_y^{M'}\}\in R_{M'}$. The optimal sample decision function is $\bar{f}^*=\arg\max_{\alpha\in\Psi_{M'}}\max_{r(\alpha)\in R_{M'}}\bar{F}(\bar{f})$. In order to solver this extreme problem we apply the algorithm *MLRP* of step-by-step increase attachments of decision trees. It do the branching of top point on that value criterion $\bar{F}(\bar{f})$ is maximum and the top point is divisible or $\bar{F}(\bar{f})\geq F^*$. The top point is indivisible if 1) number of final top point is $M'=M^*$

or 2) $\hat{N}^t \le N^*$. That criterion and parameters F^*, M^*, N^* assign method of constructing sample decision function.

In order to estimate the presented method quality we do statistical modeling. The average of the criterion of sample decision function on samples of fixed size $m_F(c) = \mathrm{E}_{V_N} F(c, \bar{f})$ is estimated for fixed nature strategy.

Conclusion

An approach to solving the problem of heterogeneous multivariate variable recognition with respect to the sample size was considered in this paper. The solution of this problem was assigned by means of presented criterion. The universality of the logical decision function class with respect to presented criterion makes the possible to introduce a measure of distribution complexity and solve this problem for small sample size. For the nature strategy and the class of logical decision function the criterions properties are presented by means of statements and consequences for pattern recognition problem. The relationship of the $\overline{m}_F(c) = \mathrm{E}_{V_N} F(c, \bar{f})$ estimate with respect to decision function class complexity for fixed nature strategy complexity demonstrates the method quality.

Bibliography

[Lbov G.S., Starceva N.G, 1999] Lbov G.S., Starceva N.G. Logical Decision Functions and Questions of Statistical Stability. Inst. Of Mathematics, Novosibirsk.

[Lbov G.S., Stupina T.A., 2002] Lbov G.S., Stupina T.A. Performance criterion of prediction multivariate decision function. Proc. of international conference "Artificial Intelligence", Alushta, pp.172-179.

[Vapnik V.N., Chervonenkis A.Ya, 1970] Vapnik V.N., Chervonenkis A.Ya .Theory of Pattern Recognition, Moscow: Nauka.

Author's Information

Tatyana A. Stupina – Institute of Mathematics SBRAS, Koptuga 4 St, Novosibirsk, 630090, Russia; e-mail: stupina@math.nsc.ru

ON THE QUALITY OF DECISION FUNCTIONS IN PATTERN RECOGNITION

Vladimir Berikov

***Abstract**: The problem of decision functions quality in pattern recognition is considered. An overview of the approaches to the solution of this problem is given. Within the Bayesian framework, we suggest an approach based on the Bayesian interval estimates of quality on a finite set of events.*

***Keywords**: Bayesian learning theory, decision function quality.*

ACM Classification Keywords: I.5.2 Pattern recognition: classifier design and evaluation

Introduction

In the problem of decision functions quality analysis, one needs to find a decision function, not too distinguishing from the optimal decision function in the given family, provided that the probability distribution is unknown, and learning sample has limited size. Under optimal decision function we shall understand such function for which the risk (the expected losses of wrong forecasting for a new object) is minimal. In particular, the following questions should be solved at the analysis of the problem.

a) With what conditions the problem has a decision?

b) How the quality of decision function can be evaluated most exactly on learning sample?

c) What principles should be followed at the choice of decision function (in other words, what properties must possess a class of decision functions and learning algorithm) under the given sample size, dimensionality of variable space and other information on the task?

The principle possibility to decide the delivered problem is motivated by the following considerations. Firstly, for the majority of real tasks of forecasting on statistical information it is possible to expect a priori the existence of certain more or less stable mechanism being the basis of under study phenomena. Secondly, it is possible to expect that available empirical information in one or another degrees reflects the functioning of this unknown mechanism. Thirdly, it is required for the successful solution that the class of decision functions and learning algorithm should possess certain characteristics, on which will be said below.

Basic Approaches

A number of different approaches to the solution of the problem can be formulated. In experimental approaches [1] (one-hold-out, bootstrap, cross-validation) the data set is divided on learning sample (for decision function finding) and test sample (for evaluation of quality of the decision function) repeatedly. The performance of the given method for decision function construction is then evaluated as the average quality on test samples. The shortcoming of this approach is its computational expensiveness.

Probabilistic approach is based on the preliminary estimation of the distribution law. A learning problem can be solved if this law can be reconstructed from the empirical data. The given approach can be used only when sufficiently large a priori information on the distribution is available. For instance, it is possible to show that the problem of distribution density reconstruction from the empirical data is in general an ill-posed problem [2].

Two directions within the framework of this approach are known. The first one deals with the asymptotic properties of learning algorithms. In [3], the asymptotic evaluations of quality for such methods as a K-nearest neighbor rule, minimum Euclidean distance classifier, Fisher's linear discriminant etc are given.

The second direction takes into account the fact that the size of learning sample is limited. This approach uses the principles of multivariate statistical analysis [4] and restricts the set of probabilistic models for each class, for instance, assumes the Gauss distribution low. The determined types of decision functions (for instance, linear or quadratic; perceptron) are considered.

Vapnik and Chervonenkis [2] suggested an alternative approach to the solution of the given problem ("statistical learning theory"). This approach is distribution-free and can be applied to arbitrary types of decision functions. The main question is "when minimization of empirical risk leads to minimization of true unknown risk for arbitrary distribution?". The authors associate this question and the question of existence of the uniform convergence of frequencies to probabilities on the set of events related to the class of decision functions. The fundamental notions of growing function, entropy, VC-dimension that characterize the difficulty of decision functions class are suggested. It is proved that the frequency converges to probability uniformly if and only if the amount of entropy per element of sample converges to zero at the increase of sample length. As far as these evaluations are received for the worst case, on the one hand they are distribution-independent, but on the other hand give too pessimistic results. In [5] the notion of efficient VC-dimension is offered, which is dependent from distribution. With this notion, the authors managed to perfect greatly the accuracy of evaluations.

Within the framework of statistical learning theory the structured risk minimization method was suggested. The idea of the method consists in consequent consideration of classes of decision functions, ranked on growth of their complexity. The function minimizing empirical risk in the corresponding class and simultaneously giving the best value for guaranteed risk is chosen. Support vectors machine [6] uses an implicit transformation of data to the space of high dimensionality by means of the given kernel function. In this space, a hyperplane maximizing the margin between support vectors of different classes is found. It is proved that the main factor influences the risk is not the dimensionality of the space but margin width.

In "PAC-learning" approach [7,8] ("Probably Approximately Correct"; developed within the framework of computational learning theory) the complexity of learning algorithm is taken into consideration. It is required that

learning algorithm, with probability not smaller than η finding the decision function for which the probability of mistake does not exceed ε, has time of work polynomially depended on sample size, complexity of class of decision functions, and on values $1/\eta$, $1/\varepsilon$. The existence of such algorithms for some classes of recognition decision functions is proved, for instance, for conjunctions of Boolean predicates, linear decision functions, some types of neural networks, decision trees.

Statistical and computational learning theories suggest the worst-case analysis. From standpoints of statistical decision theory, their evaluations are of minimax type. However it is possible to use the average-case analysis (in Bayesian learning theory) for which it is necessary to define certain priory distribution (either on the set of distribution parameters or on the set of decision functions) and to find the evaluations at the average [9,10]. The main task is to find the decision function for which a posterior probability of error is minimal. As a rule, the finding of such function is computationally expensive, so the following rule of thumb can be used. Instead of optimum decision function search, less expensive function which is close (under determined conditions) to optimum is found. An example is minimum description length principle (minimizing the sum of the code length describing the function and the code length describing the data misclassified by this function). Another example is maximum a posterior probability function. From the other hand, the estimations can be done by statistical modeling (Markov Chain Monte Carlo method).

The main problem at the motivation of the Bayesian approach is a problem of choice of a priori distribution. In the absence of a priori information, it is possible to follow Laplace principle of uncertainty, according to which uniform a priori distribution is assumed. If the uncertainty in the determining of a priory distribution presents, the robust Bayesian methods can be applied.

Bayesian learning theory was used for discrete recognition problem [10], for decision trees learning algorithms [11] etc. Within the Bayesian framework, the case of one discrete variable is mostly suitable for analytical calculations. Hughes [10] received the expression for the expected probability of recognition error depending on sample size and the number of values of the variable. It was shown that for the given sample size, an optimum number of values exists for which the expected probability of error takes minimum value. Lbov and Startseva [12] received the expressions for the expected misclassification probability for the case of available additional knowledge about the probability of mistake for the optimum Bayes decision function. In [13-16] this approach was generalized for arbitrary class of decision functions defined on a finite set of events. The expert knowledge about the recognition task is taken into account. Below we give the summary of the obtained results:

a) The functional dependencies are obtained between the quality of an arbitrary method of decision functions construction and learning sample size, number of events [14,16].
b) The theoretical investigation of empirical risk minimization method is done [14,15].
c) The posterior estimates of recognition quality for the given decision function are found (with respect to number of events, empirical risk, sample size) [13].
d) New quality criteria for logical decision functions are suggested on the basis of above mentioned results. An efficient method for classification tree construction is proposed [15].
e) New methods are suggested for the following data mining tasks: regression analysis, cluster analysis, multidimensional heterogeneous time series analysis and rare events forecasting [15].

Main Definitions

Let us consider a pattern recognition problem with $K \geq 2$ classes, input features $X_1, X_{2,\ldots}, X_n$ and output feature Y with domain $D_Y=\{1,\ldots,K\}$. Denote D_i as a set of values of feature X_i, $i=1,\ldots,n$. Suppose that the examples from general sample are extracted by chance, therefore the features Y, X_i are random. A function $f : \prod_{i=1}^{n} D_i \to D_Y$ is called the *decision function*. A special kind of the decision function is a *decision tree T*. The decision function is built by the random sample of observations of X and Y (learning sample). Let learning sample be divided into two parts. The first part is used to design decision tree T, and the second part to prune it. Let T_{pr} be a *pruned decision tree*. During the pruning process, one or more nodes of T can be pruned. By numbering the leaves of a tree, we can reduce the problem to one feature X. The values of this feature are coded by numbers $1,\ldots,j,\ldots, M$,

where M is number of leaves ("events", "cells"). Let p_j^i be the probability of joint event *"X=j,Y=i"*. Denote a priori probability of the i-th class as p^i. It is evident that $\Sigma_i p^i=1$, $\Sigma_j p^i_j=p^i$. Let N be sample size, n_j^i be a frequency of falling the observations of i-th class into the j-th cell. Denote $s=(n_1^1,n_1^2,...,n_1^K,n_2^1,...,n_M^K)$. $j=1...M$, $i=1...K$. Let $\widetilde{N}$ be a number of errors on learning sample for the given decision function.

Let us consider the family of models of multinomial distributions with a set of parameters $\mathbf{\Theta}=\{\theta\}$, where $\theta=(p_1^1,p_1^2,...,p_1^K,p_2^1,...,p_M^K)$, $p_j^i\geq 0, \sum_{i,j} p_j^i=1,$ In applied problems of recognition, vector θ (defining the distribution law of a recognition task) is usually unknown. We use the Bayesian approach: suppose that random vector $\Theta=(P_1^1,...,P_1^K,P_2^1,...,P_M^K)$ with known priory distribution $p(\theta)$ is defined on the set of parameters. We shall suppose that Θ is subject to the Dirichlet distribution (conjugate with the multinomial distribution): $p(\theta)=\frac{1}{Z}\prod_{l,j}(p_j^l)^{d_j^l-1}$, where $d_j^l > 0$ are some given real numbers expressing a priori knowledge about distribution of Θ, l=1,... ,K, j=1,..., M, Z is normalizing constant. For instance, under $d_j^l\equiv 1$ we shall have uniform a priori distribution ($p(\theta)\equiv const$) that can be used in case of a priori uncertainty in the specification of a class of recognition tasks. The value M will be called the **complexity** of decision function class.

Let K=2, $d_j^l\equiv d$ for all l,j, where d>0 is a parameter. Thus we assume that there is no a priori information on the preferences between events, however a priori distribution is not uniform ($d\neq 1$). For the fixed vector of parameters θ, the probability of error for the Bayesian classifier f_B is: $P_{f_B}(\theta)=\sum_j \min\{p_j^1,p_j^2\}$. In [16], the expected probability of error $EP_{f_B}(\Theta)$ was found, where the averaging is done over all random vectors Θ with distribution density $p(\theta)$:

Theorem 1 [16]. $EP_{f_B}(\Theta)=I_{0,5}(d+1,d)$, where $I_x(p,q)$ is Beta distribution function.

The value d allows to express expert's knowledge about the expected degree of the "intersection" between patterns. For example, if d is small, then it is assumed that the recognition tasks with small probability of error are more probable to appear.

The choice of optimal complexity of the decision function class

Let μ* denotes the minimum empirical error minimization method: f= μ*(s), where f is classifier from the given class Φ, s is learning sample. Consider the expected probability of error for this method: $EP_{\mu^*}=E_{\Theta,S}P_{\mu^*(S)}(\Theta)$, where the averaging is done over all random vectors Θ and samples S.

Proposition 1. $EP_{\mu^*}=\frac{N!M}{(2Md)_{(N+1)}}\sum_{s_1}\frac{(2Md-2d)_{(\bar{n}_1)}d_{(n_1^1)}d_{(n_1^2)}}{\bar{n}_1!n_1^1!n_1^2!}(d+\min\{n_1^1,n_1^2\})$,

where $x_{(n)}$ denotes multiplication $x(x+1)...(x+n-1)$, operator $\sum_{s_1}$ denotes the summation over all vectors $s_1=(n_1^1,n_1^2)$ such that $n_1^1+n_1^2\leq N$.

The **proof** directly follows from the results given in [16].

Let us form the sequence of classes with the increasing complexities M=1,2,...,M_{max}. When increasing the complexity of the class, it is naturally to expect that the averaged probability of error $EP_{f_B}(\Theta)$ also changes.

For M=1 this value is maximal since the decision is made on a priori probabilities of classes only. When M increases, the value $EP_{f_B}(\Theta)$ usually monotonously decreases (converges to zero when the class is formed by partition of real variables space). Herewith under small values of M the complication of class, as a rule, causes the meaningful reduction of $EP_{f_B}(\Theta)$, but under the large values of M the effect of the complication is less noticeable. Let us denote the expected probability of error through $EP_B(M)$, the corresponding value of Dirichlet parameter through d_M, the expected probability of error through $P_{\mu^*}(N, M, d_M)$.

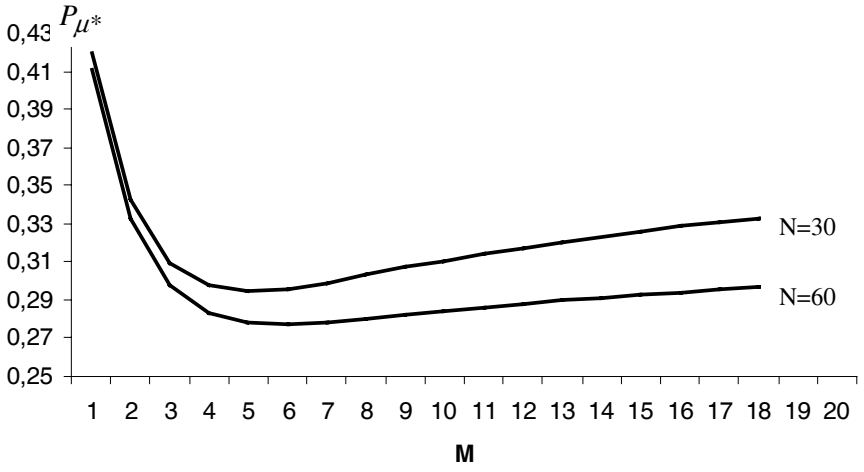


Figure 1.

The choice of specific values $d_1, d_2, \ldots, d_{Mmax}$ (or corresponding values $EP_B(1)$, $EP_B(2)$, …,$EP_B(M_{max})$ should be done with use of expert knowledge. It is possible to offer, for instance, the following way. Consider a model of the dependency of $EP_B(M)$ from M (power or exponential), as well as edge values $EP_B(1)$, $EP_B(M_{max})$. Then in accordance to Proposition 1, the vales $d_1, d_2, \ldots, d_{\text{Mmax}}$ are calculated. Hereinafter the set of expected probabilities of error is calculated for different values M.

Fig. 1 shows the example of the dependencies between the expected misclassification probability and M for the model $EP_B(M) = (EP_B(1) - EP_B(M_{\max}))\exp(-0{,}75(M-1)) + EP_B(M_{\max})$, M=2,3,…,M_{max}–1, M_{max}=20, $EP_B(1)$=0,4, $EP_B(M_{max})$=0,25. One can see that between the edge values of M there exists the best value, depending on sample size, for which the expected probability of error is minimal.

Acknowledgements

This work is supported by the Russian Foundation of Basic Research, grants 07-01-00331-a, 07-01-00393-a.

Bibliography

[1] Breiman L. Bagging predictors // Mach. Learn. 1996. V. 24. P. 123-140.

[2] Vapnik V. Estimation of dependencies based on empirical data. Springer- Verlag. 1982.

[3] Fukunaga K. Introduction to statistical pattern recognition. Academic Press, NY and London. 1972.

[4] Raudys S. Statistical and Neural Classifiers: An integrated approach to design. London: Springer-Verl., 2001.

[5] Vapnik V., Levin E. and Le Cun Y. Measuring the VC-Dimension of a Learning Machine // Neural Computation, Vol. 6, N 5, 1994. pp. 851--876.

[6] Vapnik V.N. An Overview of Statistical Learning Theory // IEEE Transactions on Neural Networks. 1999. V.10, N 5. P. 988-999.

[7] Valiant L.G. A Theory of the Learnable, CACM, 17(11):1134-1142, 1984.

[8] Haussler D. Probably approximately correct learning // Proc. 0f the 8th National Conference on Artificial Intelligence. Morgan Kaufmann, 1990. pp. 1101-1108.

[9] D.Haussler, M.Kearns, and R.Schapire. Bounds on sample complexity of Bayesian learning using information theory and the VC dimension // Machine Learning, N 14, 1994. pp. 84-114.

[10] Hughes G.F. On the mean accuracy of statistical pattern recognizers // IEEE Trans. Inform. Theory. 1968. V. IT-14, N 1. P. 55-63.

[11] W. Buntine. Learning classification trees // Statistics and Computing. 1992. V. 2. P. 63--73.

[12] Lbov, G.S., Startseva, N.G., *About statistical robustness of decision functions in pattern recognition problems.* Pattern Recognition and Image Analysis, 1994. Vol 4. No.3. pp.97-106.

[13] Berikov V.B., Litvinenko A.G. *The influence of prior knowledge on the expected performance of a classifier.* Pattern Recognition Letters, Vol. 24/15, 2003, pp. 2537-2548.

[14] Berikov, V.B. A Priori Estimates of Recognition Accuracy for a Small Training Sample Size // Computational Mathematics and Mathematical Physics, Vol. 43, No. 9, 2003. pp. 1377- 1386

[15] Lbov G.S., Berikov V.B. Stability of decision functions in problems of pattern recognition and heterogeneous information analysis. Inst. of mathematics Press, Novosibirsk. 2005. (in Russian)

[16] Berikov V.B. Bayes estimates for recognition quality on a finite set of events // Pattern Recognition and Image Analysis. 2006. V. 16, N 3. P. 329-343.

Author's Information

Vladimir Berikov – Sobolev Institute of Mathematics SD RAS, Koptyug pr.4, Novosibirsk, Russia, 630090; e-mail: berikov@math.nsc.ru

MEASURE REFUTATIONS AND METRICS ON STATEMENTS OF EXPERTS (LOGICAL FORMULAS) IN THE MODELS FOR SOME THEORY[1]

Alexander Vikent'ev

Abstract. The paper discusses a logical expert statements represented as the formulas with probabilities of the first order language consistent with some theory T. Theoretical-models methods for setting metrics on such statements are offered. Properties of metrics are investigated. The research allows solve problems of the best reconciliation of expert statements, constructions of decision functions in pattern recognition, creations the bases of knowledge and development of expert systems.

Keywords*: pattern recognition, distance between experts' statements.*

ACM Classification Keywords*: I.2.6. Artificial Intelligence - Knowledge Acquisition.*

Introduction

As the increasing interest to the analysis of the expert information given as probabilities logic statements of several experts is now shown, questions on knowledge of the experts submitted by formulas of the first order language with probabilities are interesting also. With the help of suitable procedure the statement of experts it is possible to write down as formulas of Sentence Logic or formulas of the first order language. Clearly, the various statements of experts (and the formulas appropriate to them) carry in themselves different quantity of the information. To estimate and analyses this information it is necessary to define the degree of affinity of statements that allows to estimate a measure of refutation statements of experts (a measure of refutation above at formula of the first order language with smaller number of elements satisfying it) and to specify their probabilities (an average share probabilities realizations for formulas with variables). It allows solve problems of the best reconciliation of expert statements, constructions of decision functions in pattern recognition, creation of bases of knowledge and expert systems [1].

[1] This work was financially supported by the Russian Foundation for Basic Research, project no. **04-01-00858a**.

A number of natural metrics on probabilities knowledge of experts is offered with use of suitable class of models (with metrics) some theory and modifications symmetric difference, by analogue, par example [4] for logical Lbov's the predicat for unique model. Properties of these metrics, connected to them measures of refutation of formulas (distance from the formula up to class of equivalence of identically true formula) and probability are established. From the point of view of importance of the information presented by an expert, it is natural to assume that a measure of refutation of the formula (nonempty predicate) the above, than it is les measure of elements satisfying it (i.e. a measure determined on subsets, set by predicate formulas).

We introduce the measure of refutation similarly to a case of formulas without probabilities. We call

$$R\left(P\left(\overline{x}\right)\right)=\rho_{вер}\left(P\left(\overline{x}\right),\ 1\right)$$

the measure of refutation of formula $P\left(\overline{x}\right)$, where 1 is an identical true predicate, that is, $\overline{x}=\overline{x}$. All stated for predicates (and Lbov's predicate aussi without probability) fairly and for formulas of the first order language with probabilities.

The distance between the formulas of Sentence Logic is entered in [1], properties of the entered distance are given and proved in the same place. Ways of introduction of distance between the formulas of the first order language are offered in [2]. Measures of refutation and probabilities of formulas are entered and their properties are formulated in [3]. The distance between the formulas of Sentence Logic with probabilities is entered in [3,5]. In the given work the way of introduction of distances between probabilities statements of experts represented as the formulas of the first order language theory T with probabilities is offered.

Distance between statements of experts represented as the formulas of the first order language with probabilities in theory

Let experts speak about probabilities of predicates on the product $\prod_{j=1}^{p} D_{x_j}$.

Then the given by expert probability is interpreted as follows: "the knowledge" $B_l^i =< P_l^i(x_1,...,x_p), p_l^i >$ means, that the predicate $P_l^i(x_1,...,x_p)$ is true on $n_{P_l^i}=\left\lfloor n\cdot p_l^i \right\rfloor$ trains of length p in model M_i, where $n=\prod_{j=1}^{k}\left|D_{x_j}\right|$ - measure of model.

Let's find distance between predicates P_l and P_j. For this purpose all over again we shall calculate distance $\rho^i(B_l^i,B_j^i)$ between probabilities interpretations $B_l^i =< P_l^i(\overline{x}), p_l^i >$ and $B_j^i =< P_j^i(\overline{x}), p_j^i >$ of predicates P_l and P_j in each model M_i. Distances are calculated between predicates of identical district and from the same variables plus measure protjaga (разброса в пространстве модели) p.e. [4], and without stable theory.

Then interpreting the probabilities given by experts the described above way we receive that the predicate $P_l^i(\overline{x})$ is true on $n_{P_l^i}$ trains in model M_i and the predicate $P_j^i(\overline{x})$ is true on $n_{P_j^i}$ trains in model M_i *theory* T. We shall note that is not known on what trains each predicate true and number (or mera) of trains on which these predicates are simultaneously true.

We shall consider the following task. Let the predicate $P_l^i(\overline{x})$ is true on $n_{P_l^i}$ trains in model M_i and the predicate $P_j^i(\overline{x})$ is true on $n_{P_j^i}$ trains in model M_i and k^i - number of trains on which these predicates are simultaneously true. It is required to calculate distance between $B_l^i =< P_l^i(\overline{x}), p_l^i >$ and $B_j^i =< P_j^i(\overline{x}), p_j^i >$. Distances arising in further we shall designate through $\rho_{k^i}(B_l^i,B_j^i)$, where, $k^i=t,t+1,\ldots,\min(n_{P_l^i},n_{P_j^i})$, $t=\max(0,n_{P_l^i}+n_{P_j^i}-n)$ hereinafter.

Distance $\rho_{k^i}(B_l^i, B_j^i)$ we shall define as a modifies (as ask above) symmetric difference, i.e.

$$\rho_{k^i}(B_l^i, B_j^i) = \frac{1}{n}(n_{P_l^i} + n_{P_j^i} - 2k^i), \quad (1)$$

for every one $k^i = t, t+1, \ldots, \min(n_{P_l^i}, n_{P_j^i})$. All properties of distances formulated in [1] are fair for $\rho_{k^i}(B_l^i, B_j^i)$. Let's offer some ways of calculation distance $\rho^i(B_l^i, B_j^i)$ between probabilities interpretations $B_l^i =< P_l^i(\overline{x}), p_l^i >$ and $B_j^i =< P_j^i(\overline{x}), p_j^i >$ of predicates P_l and P_j in each model M_i *theory* T. If the number k^i is not known (the number of trains on which these predicates are simultaneously true in model M_i) and if there are no preferences for value k^i (preference can be stated by experts) it is possible to act as follows. We shall assume, that all values for number k^i are equally probability. Then distance between probabilities interpretations $B_l^i =< P_l^i(\overline{x}), p_l^i >$ and $B_j^i =< P_j^i(\overline{x}), p_j^i >$ of predicates P_l and P_j in model M_i we shall define as average of distances $\rho_{k^i}(B_l^i, B_j^i)$ on all values k^i, i.e.

$$\rho(B_l^i, B_j^i) = \frac{\sum_{k^i=t}^{\min(n_{P_l^i}, n_{P_j^i})} \rho_{k^i}(B_l^i, B_j^i)}{\min(n_{P_l^i}, n_{P_j^i}) + 1 - t}. \quad (2)$$

For this distance all properties of distances formulated in [1] also are executed.

If by experts it is stated what value for k^i is more preferable in quality $\rho^i(B_l^i, B_j^i)$ it undertakes $\rho_{k^i}(B_l^i, B_j^i)$, i.e.

$$\rho(B_l^i, B_j^i) = \rho_{k^i}(B_l^i, B_j^i). \quad (3)$$

In the offered formulas (1) – (3) of distances the kind of formulas between which the distance is calculated is not taken into account. Therefore it is natural to offer distance by which takes into account a kind of formulas. Applying the model approach [1-3] to elements of set $\{M_i\}_{i=1}^s$ we shall find probabilities $P_{M_i}(P_j^i)$, $P_{M_i}(P_j^i)$ ([3]) and distance $\rho_{M_i}(P_l^i, P_j^i)$ in model M_i ([2]), then we shall calculate probability $P_{M_i}(P_l^i \wedge P_j^i) = \frac{1}{2}(P_{M_i}(P_l^i) + P_{M_i}(P_j^i) - \rho(P_l^i, P_j^i))$ ([3]) and we shall find $k_0^i = [P_{M_i}(P_l^i \wedge P_j^i) \cdot n]$ - the number of trains on which predicates are simultaneously true. Having k_0^i (calculated on models), it is possible to reduce number of possible values for k^i. Three cases here are possible: 1) if $t < k_0^i < \min(n_{P_l^i}, n_{P_j^i})$, then $k^i = k_0^i - 1, k_0^i, k_0^i + 1$; 2) if $k_0^i = t$ or $k_0^i = \min(n_{P_l^i}, n_{P_j^i})$, then, for example, $k^i = k_0^i, k_0^i + 1$; or $k^i = k_0^i - 1, k_0^i$; 3) if $k_0^i < t$ or $k_0^i > \min(n_{P_l^i}, n_{P_j^i})$, then $k^i = t$ or $k^i = \min(n_{P_l^i}, n_{P_j^i})$.

And already to these values for k^i apples offered above formulas (1) - (3) of distances. As required probably some expansion of number of values for k^i.

The offered ways it is possible to calculate distance between the following statements: $B_l^i =< P_l^i(\overline{x}), p_l^i >$ - the information received from one expert, and $B_l^j =< P_l^j(\overline{x}), p_l^j >$ - the information received from other expert. Thus we have calculated distance $\rho^i(B_l^i, B_j^i)$ between probabilities interpretations $B_l^i =< P_l^i(\overline{x}), p_l^i >$ and $B_j^i =< P_j^i(\overline{x}), p_j^i >$ of predicates P_l and P_j and degree elongate in each model M_i *theory* T.

Then as distance $\rho_{вер}(P_l, P_j)$ between predicates P_l and P_j we shall take size

$$\rho_{вер}(P_l, P_j) = \frac{1}{s}\sum_{i=1}^{s} \rho^i(B_l^i, B_j^i).$$

For all properties of distance formulated in ([5]) are carried out for $\rho_{вер}(P_l, P_j)$.

Acknowledgements

This work was financially supported by the Russian Foundation for Basic Research, project no. 04-01-00858a.

Bibliography

[1] Lbov G.S., Startseva N.G. Decision Logical Functions and Statistical Robustness. Novosibirsk: Izd. Inst. Math., 1999.

[2] Vikent'ev A.A., Koreneva L.N. "Setting the metric and measures of informativity in predicate formulas corresponding to the statements of experts about hierarchical objects", *Pattern Recognition and Image Analysis*, V. 10, N. 3, (2000), 303--308.

[3] Vikent'ev A.A., Koreneva L.N. "Model approach to probabilities expert statements", "*Mathematical Methods for Pattern Recognition – 10", Moscow,* (2001), 25-28.

[4] G.S.Lbov, M.K.Gerasimov. Determining the distance between logical statements in forecasting problems. In: Artificial Intelligence, 2'2004 [in Russian]. Institute of Artificial Intelligence, Ukraine.

[5] Викентьев А.А., Лбов Г.С., Коренева Л.Н. "Расстояние между вероятностными высказываниями экспертов", *Искусственный интеллект, 2'2002, НАН Украины,* 58-64.

[6] Keisler G., Chang C. Model theory. M.: Mir, 1977.

Author's Information

Alexander Vikent'ev – Institute of Mathematics, SB RAS, Acad. Koptyuga St., bl.4, Novosibirsk, Russia; e-mail: vikent@math.nsc.ru

ANALYSIS AND COORDINATION OF EXPERT STATEMENTS IN THE PROBLEMS OF INTELLECTUAL INFORMATION SEARCH[1]

Gennadiy Lbov, Nikolai Dolozov, Pavel Maslov

Abstract*: The paper is devoted to the matter of information presented in a natural language search. The method using the statements agreement process is added to the known existing system. It allows the formation of an ordered list of answers to the inquiry in the form of quotations from the documents.*

Keywords*: Search engine, natural language, coordination of statements, semantic graph*

ACM Classification Keywords*: I.2.7 Computing Methodologies – Text analysis*

Introduction

Efficiency of the search engine is determined by the use of various methods of relevant documents revealing and insignificant ones eliminating, as well as methods peculiar to the specific search engine or their certain kind (for example, specialized search engines). Existing search engines are based on the oversight of index databases of

[1] This work was financially supported by **RFBF-04-01-00858**

the processed documents. The purpose is revealing the objects satisfying some criteria. However, such systems do not analyze the sentences of the document for revealing their structure and interrelations.

In the paper an approach to the search engine construction based on the analysis of semantic structure of sentences and their interrelations in the document is offered. Such method allows to do the search considering the logic of sentences thus taking into account the sense of a document. Generally it provides a stricter criterion of significant documents selection, based on accordance to a certain logic structure reflecting the sense of inquiry.

The main issue solved by the offered algorithm consists in doing the logic analysis of sentences for the subsequent search, i.e. in formation of the ranged list of answers to the inquiry in the form of quotations from documents instead of the list of these documents. Intellectuality of this method lies in its simplification of sentences perception and analysis by a person.

This system was developed as a superstructure over an existing search engine ISS2 (Internal Search System) [1]. However, independent functioning of the offered system, for example, for doing the analysis in some interesting documents is also possible. The purpose is in providing search service on local and public network catalogues being storehouses of the information. For the effective search within several storehouses there is an option for aggregation of several search servers to a distributed system. The software contains the means of carrying out a safe remote management as well as all components status analysis done by a search engine.

Selection of Search System

To derive sentences structure the system uses working results a natural text translation system [2]. It describes the methods of translated documents processing for "natural" translation considering specific features of languages. In [5] various systems of parse such as «Dialing»: L. Gershenzon, T. Kobzareva, D. Pankratov, A. Sokirko, I. Nozhov (www.aot.ru); the program of scientific group FtiPL (Institute of linguistics) RGGU (T.Yu. Kobzareva, D.G. Lakhuti, I. Nozhov); LinkParser (www.link.cs.cmu.edu/link). The selection of basis for the developed method was stipulated among other things by a good description and demonstration of system abilities [2]. In this system the analysis is done through several steps, which simplified sequence is as follows: primary, morphological, parse and semantic. Each step uses the results achieved on the previous one. The purpose of primary analysis is in the analysis of the initial document which identifies its sentences, paragraphs, notes, stable statements, electronic addresses etc. As a result the table consisting of some fragments of the initial text and their descriptors is formed. At the following step words morphoanalysis and lemmatization is done, that is each word becomes respectfully attributed with its normal form, morphological part of speech and the set of grammems, defining its grammatical gender, number, case etc. In parse syntactic groups characterized by certain parameters (type of a group, position, parental group) are defined. On the step of the semantic analysis semantic relations describing certain binary links between dependent and operating members are formed. These binary relations are just used in the offered algorithm. Resulting semantic graph characterizes interrelated binary links in the initial text sentences which reflect their logic.

For the solution of the search issue the agreement of statements described in [3] is required on a certain step. So far the resulting sets of relations in the initial text are determined by multiple expert statements whereas in the inquiry text the are defined by a set of certainly true and agreed statements. The algorithm is offered for cases with one or several experts. At first the algorithm agrees the statements of one expert which leads to a number of formulas, and then a process of overall agreement of already agreed opinions of each expert is accomplished. The specific feature if this algorithm is that he identifies absolutely all regularities. Therefore the paper [4] describes an approach to reduction of dimension statements set given on sample with the purpose of the maximal reduction of its dimension at the minimal loss of information.

Co-ordination

Basing on the intermediate results of the system work [2], which are the semantic graphs of the sentences, the logic form is constructed for each sentence. This form is a model in the language of predicates calculus of two variables united in conjunctions. Each of such predicates is an elementary statement. The following problem is to accomplish the procedure of statements agreement in the models on the base of these received models of

sentences in the text and inquiry. To do it the predicates of one type are isolated and their set (for each type of predicates) corresponding to the sentences the text is a set of agreed statements whereas their set corresponding to the inquiry is an agreed in advance statement. Considering that each predicate is a part of a sentence model, the crossing of the sets corresponding to agreed predicates of different types is taken. This crossing can be considered as the result of search in the document.

Hypotheses

For the further description of algorithm it is necessary to introduce the following assumptions:

1. Sentences having different predicate structures and different variables in them are considered as the facts of different types supplementing each other.
2. A sentence with the same predicates and with the same (i.e. synonymous) variables are considered supplementing each other, therefore one-type variables are designated by the same letter with an identical index.
3. In case of crossing variables from different predicates we obtain more complicated variant of sense addition.

Each semantic link in the graph defines some type of a two variables predicate. Let's designate with letters Xi, Yi, Zi etc. each predicate variable. As the predicate defines the relation between its variables, the sets of the one-type variables standing in a certain position in the predicate are designated by the same letter with different indexes. Variables in predicates crossing are respectively designated by the same letter with an identical index. Predicates are designated by the name of semantic links. Synonymous words standing in identical positions and in identical predicates are designated by the same variables.

Analysis and Co-ordination

For the sentences and inquiry agreement, inquiry predicates are considered separately. The predicates are picked out one by one from the inquiry and in the same time the predicates of respective types are picked out from the text sentences. Expert statements are agreed with the elementary inquiry predicate which is considered to be agreed in advance.

Decision of the Formulated Task Requires Some Modification of the Algorithm Offered in [3]

Let some statement with known characteristics requires to define its belonging to the certain image. The predicate sets corresponding one or another image are considered separately. The general formal writing of a sentence is done in the form of two-place predicates conjunction. The area of predicate is defined by nominal variables satisfying the list of admissible values. We shall designate T_{ji}^{k} the truthful areas of function and argument variables in the initial sentences inquiry, where i, j, k are the numbers of predicates, statements and the links between argument and function variables, respectively.

As variables are nominal the area of true statements is defined by variables satisfying the list of admissible values. Such list has to be based on a synonyms dictionary. Besides the lists of synonyms it is also necessary for such a dictionary to contain also factors of words affinity. For example, each word from a synonymic group corresponds to the list of synonyms with decreasing weights. To simplify the finding of truthful area it is possible to define the truthfulness of statement on the base of variables satisfying the list consisting of one admissible value. But this list can be expanded with synonyms. Aprioristic probabilities of statements are equal to 1/n (S), where n is a number of statements S.

In the offered system it is enough to accomplish the agreement at a level of one expert as for simplification the analysis is done only in one document, not between many documents. Since predicates are two-placed and variables in them are from different truthful areas, then for the agreement of one expert statement it is necessary to consider separately variables in predicates. Assuming that the statement obtained from the inquiry is true and agreed we define truthful areas from each predicates included in it. The further procedure is done for each

separate predicate. T_{pi}^1 is a truthful area of the first variable in the predicate i the inquiry p. T_{pi}^2 is the same for the second variable. The order of choice of the first and the second (the function and the argument) variable can be interchanged for altering the character of agreement, but the choice of the second variable in a predicate as the main one is more logical. Lets designate T_{ji}^1, T_{ji}^2 truthful areas of variables in predicates of the initial text. Respectively, the statement satisfying:

1. $m(T_{ji}^2 \wedge T_{pi}^2) \geq \beta_{r1}$ and $m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is true,
2. $m(T_{ji}^2 \wedge T_{pi}^2) \geq \beta_{r1}$ and $\neg m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is not likely
3. $\neg m(T_{ji}^2 \wedge T_{pi}^2) \geq \beta_{r1}$ and $\neg m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is denying
4. $\neg m(T_{ji}^2 \wedge T_{pi}^2) \geq \beta_{r1}$ and $m(T_{ji}^1 \wedge T_{pi}^1) \geq \beta_{r2}$ is denying at a choice of the second variable as the main, and not likely in other case. β_{r2} is a parameter.

Thus we receive sets of statements: ω_1 - not likely, ω_2 - true, Ω - denying.

The following steps of the one expert statements agreement are similar to described in [3].

Ranging

Let's designate Nsi the number of all predicates in a sentence, Nsoi the number of agreed predicates of a sentence, Nr the number of predicates in an inquiry. Then for determination of the sentences relevance we have to calculate the ratio:

$$k = \frac{(N_{so_i})^2}{N_{s_i} \cdot N_r}$$

As a result we receive a set of agreed statements for the first type of predicates. The procedure of agreement is repeated separately for all other predicates and we obtain the sets of agreed statements of different type, each of which defines the sentence. Finding the crossing of all these sets we receive the set of sentences satisfying to the inquiry. The outcoming set forms the result in a usual language considering text paragraphs and document headings. Thus the trial algorithm of significant sentences allocation in the text is obtained; it reflects the first and the second assumption about the usual language.

Example (in Russian)

The simple text: **Рыбак собрался ловить рыбу. Рыбак взял удочку и ведро. Рыбак забросил крючок в реку и стал ждать. По реке проплывала лодка.**

And simple inquiries: **1. Рыбак ловит рыбу. 2. Рыбак взял наживку. 3. Мокрый рыбак.**

The sentence graphs constructed by the system [1] look as follows:

Sentences in the text:

The formula of the 1st sentence: $SUB(z_1, x_1) \cup OBJ(z_1, y_1)$

The formula of the 2nd sentence: $SUB(z_2, x_1) \cup OBJ(z_2, y_2) \cup OBJ(z_2, y_3)$

The formula of the 3rd sentence: $SUB(z_3, x_1) \cup OBJ(z_3, y_4) \cup TP(z_3, t_1) \cup SUB(z_4, x_1)$

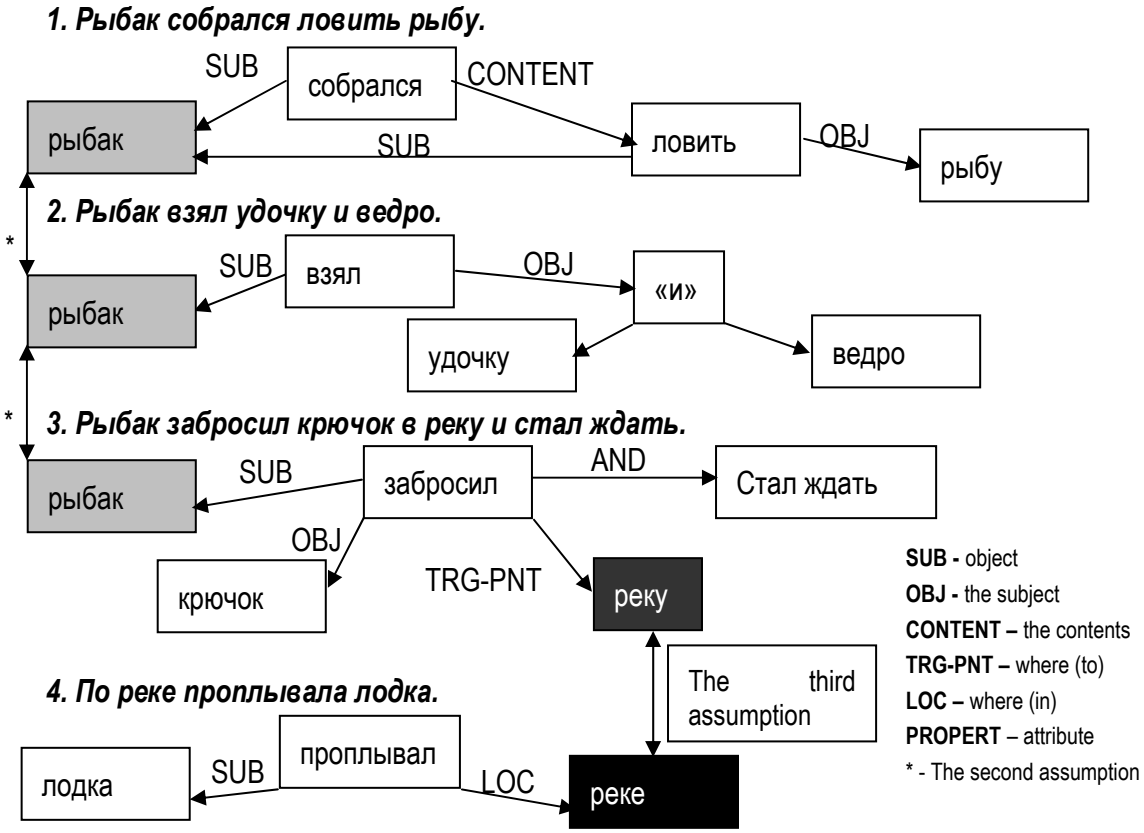
The formula of the 4th sentence: $SUB(z_5, x_2) \cup LOC(z_5, l_1)$

Sentences of the inquiry:

The formula of the 1st sentence: $SUB(z_1, x_1) \cup OBJ(z_1, y_1)$

The formula of the 2nd sentence: $SUB(z_2, x_1) \cup OBJ(z_2, y_5)$

The formula of the 3rd sentence: $PRT(x_1, p_1)$



Conclusion

For the inquiry 1 the structure of inquiry and predicate variables are similar to one of the text sentences, therefore at least one sentence is in complete agreement with such inquiry. In the second inquiry there the structure is concurrent, variables in a predicate are distinct - the full agreement is not present, therefore the ranging will show only 25%, whereas a simple phrase «рыбак взял» will show 100%. The third inquiry contains the single predicate PRT designating the property of an object. Such predicate is not present in the text, therefore the algorithm agrees nothing. In other words, the sense of inquiry is not crossed with the sense of the text.

Bibliography

[1] P.P. Maslov. Designing Materials of the All-Russian scientific conference of young scientists in 7 parts. Novosibirsk: NGTU, 2006. Part 1. - 291 p. // pp. 250-251

[2] Automated text processing “DIALING” // www.aot.ru

[3] G.S. Lbov T.I. Luchsheva. The analysis and the coordination of expert’s knowledge in problems of recognition // 2’2004, NAS of Ukraine, pp. 109-112.

[5] Nozhov I. The Parse // http://www.computerra.ru/offline/2002/446/18250/

Authors' Information

Gennadiy Lbov - SBRAS, The head of laboratory; P.O.Box: 630090, Novosibirsk, 4 Acad. Koptyug avenue, Russia; e-mail: lbov@math.nsc.ru

Nikolai Dolozov - NSTU, The senior lecturer, Cand.Tech.Sci.; P.O.Box: 6300092, Novosibirsk, 20 Marks avenue, Russia; e-mail: dnl@interface.nsk.su

Pavel Maslov - NSTU, The post-graduate student of of FAMI; P.O.Box: 6300092, Novosibirsk, 20 Marks avenue, Russia; e-mail: altermann@ngs.ru

TABLE OF CONTENTS OF VOLUME 14, NUMBER 1