ITHEA

International Journal INFORMATION CONTENT & PROCESSING

2014 Volume 1 Number 1

International Journal
INFORMATION CONTENT & PROCESSING
Volume 1 / 2014, Number 1

EDITORIAL BOARD

IJ ICP is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ ICP rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals as well as the **subscription fees** are given on www.ithea.org .
The papers should be submitted by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

International Journal "INFORMATION CONTENT AND PROCESSING" Volume 1, Number 1, 2014

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia.

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: Ina Markova

Printed in Bulgaria

ISSN 2367-5128 (printed) **ISSN 2367-5152 (online)**

PREFACE

ITHEA International Journal "***Information Content and Processing***" (**IJ ICP**) is aimed to present advances in theories of information content models and content processing technique by different systems/agents – natural, as well as artificial.

The concept "Information Content" has diverse interpretations. Chalker and Weiner point out in Oxford Dictionary of English Grammar (1994) that *"the notion of information content is related to statistical probability. If a unit is totally predictable then, according to information theory, it is informationally redundant and its information content is nil."* In linguistics and information theory, the Information Content/Corpus is the amount of information conveyed to a particular unit of language in a particular context, etc.

In this journal the concept "Information Content" is used in its broad sense - as the matter of information models created and used by natural and/or artificial intelligent systems. This is akin to terms used in current EU FP for ICT:

- Big Data, improving building innovative data products and services and solving fundamental and applied, market driven research problems related to the scalability and responsiveness of analytic capabilities such as KDD, gaming, semantics, DSP;
- Machine translation and other NLP, to overcome barriers to multilingual online communication which is still hampering a wider penetration of cross-border commerce, social communication and exchange of cultural content enabling the full deployment of the Single Digital market;
- Tools for creative, media, knowledge and learning industries, socio economical complex systems with visualization, cross boarder geo-bio-eco models and serious games;
- Multimodal, multimedia, hereditary and hybrid natural computer processing and interaction.

We kindly introduce the first issue of our new Journal by a content best correlated to the aims we already described.

Introductory papers raise the actual topics of Privacy Preserved Computation that is based on Big Data securing, searching, and compressing with special properties (distance preservation and homomorphic encryption). Broad machine learning technique is demonstrated in use of solving this perspective research direction.

This is continued demonstrating the new "vantage point" for next stage of developing the "Information Science" and the information content and its processing being the main object of studies for near future.

We kindly invite the broad research community to join to this important initiative for Data Content investigations convergences and for their business implementations for state, social and economic needs.

Sofia, March 2014.

Levon Aslanyan, Krassimir Markov

BRIEF ANALYZIS OF TECHNIQUE FOR PRIVACY PRESERVING COMPUTATION[1]

Levon Aslanyan, Vardan Topchyan, Haykaz Danoyan

***Abstract**: The privacy preserving computation research area is considered. The problem appear when one party have confidential data and need to do intense computations over that data, and computations will be done by the second party, which may be supposed being untrusted. So the content of the raw data should be kept private from the second party during the computations. Therefore these data are to be encrypted before sending them to the second party. Two possible solution scenarios are considered – one in physical and the second in theoretical levels. Physical level solution assumes some hardware integration and reorganizations. Theoretical level solution is based on cryptographic approach (homomorphic encryption).The main idea is to encrypt data in such a way that the owner, after getting the results of computations over the encrypted data, will be able to get the results on original data only by decrypting the received results. The paper brings description and analyzes of such known schemas. The final outcome is that practical cryptographic tools today are really not ready to be applied on privacy preserving computations, so that the way of solution is the use of heuristic data analyses models and algorithms that replace original data with synthesized data. Considering preparatory, this article is followed by the base research part where synthetic data generation is considered on base of CART algorithm and clustering type computational algorithms.*

***Keywords**: Privacy preserving computations, homomorphic encryption, synthetic data generation.*

***ACM Classification Keywords**: H.1 Information Systems – Models and principles, I.2.0 Artificial intelligence.*

1. Introduction

Privacy is one of the most important properties related to state or societal information and to information systems analyzing such data. When privacy restrictions applied by the user/owner, computation will guarantee that no leak of information happens during the computations [Ferrer, 96; Defays, 93].

Mission of statistical agencies and survey organizations is to disseminate summarized social or economical data. However, demand from an increasingly sophisticated and computationally capable research community for access to microdata - actual data records, as opposed to only summaries of the data - is high and growing. Dissemination of microdata vs. summaries greatly benefits society, as well as facilitates research and advances in economics, sociology, public health, and many other areas of knowledge. However, data disseminators cannot release microdata as collected, because doing so would reveal respondents' identities or values of sensitive attributes [Duncan, 91; Wallman, 04].

[1] Partially supported by grants № SCS 13RF-088, and № 3-1B340 of State Committee of Science of Ministry of education and science of Republic of Armenia

So, for public microdata releases, a special technique - statistical disclosure limitation/control (SDL) is used to alter the data in a way that maintains the utility but limits disclosure risk. Methods for SDL can be divided in two types:

- Perturbative methods [Dalenius, 82; Ferrer, 01a; Ferrer, 01b; Kim, 86; Moore, 96; Tendik, 94], which distort original values, so that the distorted values are publishable. For example, ages or incomes can be recorded in aggregated categories; or data values can be swapped for selected records, e.g. switch the sexes of some men and women in the data, in hopes of discouraging users from matching, since matches may be based on incorrect data. Or, they add noise to numerical data values to reduce the likelihood of exact matching on key variables or to distort the values of sensitive variables;
- Synthetical methods [Feinberg, 06; Sanz, 99; Reiter, 02; Reiter, 05], which lead to release of synthetical data - random samples drawn from the distribution representing original data. It is necessary to release multiple versions of synthetic data in order to guarantee the validity of statistical inferences. Other variant of synthetical approach is release of partially synthetical data, when only some of the values, which are considered sensitive, are synthesized, while others are left unchanged.

The encryption is one of the techniques that provide privacy of information. As a limitation to this technique it is enough to mention that an information system which works with encrypted data can at most store or retrieve the data for the user; and any more complicated operations seem to be requiring the data decryption before being operated on. Effective search over the encrypted data requires that the encryption scheme preserves the distances and similarities [Miyoung, 13]. Some particular encryption functions which permit encrypted data to be operated on without preliminary decryption of the operands are known for several sets of interesting operations. These special encryption functions are called “privacy homomorphisms” which form an interesting subset of arbitrary encryption schemes called “privacy transformations”. The idea of the discussion below is to learn if such partial schemes can be combined to an integrated application system for privacy preservation.

For example, let us consider a statistical organization which owns confidential data (financial data, credit card data, health data, etc.). And let there is a computational environment (datacenter, cluster, grid, etc.) to be used for information analysis. In simple scenario we will suppose that data is stored on a data bank in encrypted form, Figure 1. For such organization to provide not archiving but the necessary computations over the data, the following 2 ways – one physical level solution and the second - a theoretical level solution can be considered:

I. Allowing the computational system to store the data in decrypted form at the time of computations. This approach needs some additional hardware modifications to provide the security and will be considered in section 2.

II. The data in computational system stays always in encrypted form. In this case the encryption function must satisfy some additional properties such that the results of computation on original data coincide with the decrypted results of computations on encrypted data. More precisely this approach will be considered in section 3.

Since I. and II. provide limited opportunities for privacy preservation after the section 3 we will discusse the privacy preservation computation heuristics.

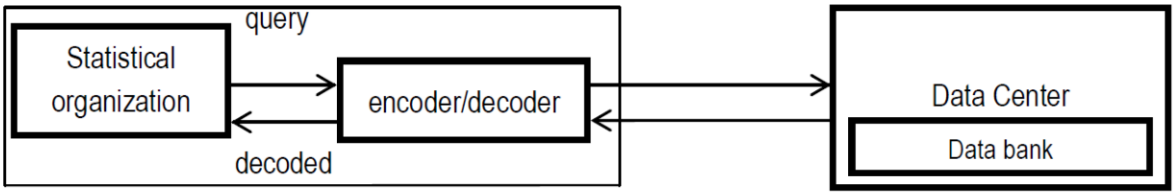


Figure 1. Encrypted data are stored in Data center

2. Physical Level Solution

Let us consider an example that shows how a computational system might be reorganized in a form to solve the problem of performing operations on decrypted data securely [Rivest, 78]. Those modifications are presented in Figure 2.

In this example, in addition to the standard register set and ALU, a secure register set and ALU is a added to the infrastructure. In this case, all communication of data between operation memory and the physically secure register set passes through an encoder-decoder E supplied with the user's key, so the unencrypted data can exist only within the physically secure register set. Moreover, it follows that all sensitive data in operating memory, data bank files, ordinary register set, and on the communications channel will be encrypted. During operation, a load/store instruction between operating memory and the secure register set will automatically cause the appropriate performed decryption/encryption operations.

An obvious problem is getting the encoder/decoder E which loaded with the user's key K without compromising the security of the user's key. In this case, one possible solution is to keep the user's key encrypted under control of a system key S. The encrypted form of key K, $E_S(K)$, can be transmitted over the insecure channel to the system, decrypted by the physically secure decoder F, and loaded into the encoder-decoder E.

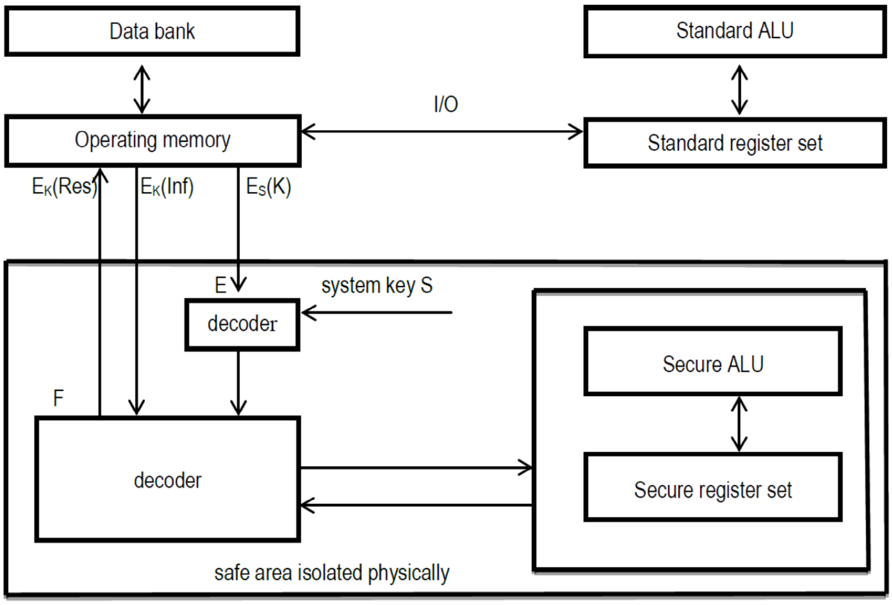


Figure 2. Computation in a secure environment

Remark: In addition to key management problems, in this scheme there appear speed degradation type questions, caused by invoking the encryption/decryption in each load and store.

3. Theoretical Level Solution

3.1 Basic Concepts of Encryption

Encryption (enciphering) is one of the main methods to protect data privacy. A process of enciphering is the transformation of one message (initial data), called a plaintext (irrespective to the data type), to another message, called ciphertext, using some specific transformation. As a rule in a cryptographic scheme this transformation is open, publicly known, and a secret key is incorporated to provide the security. The process of transformation of ciphertext into plaintext is called deciphering.

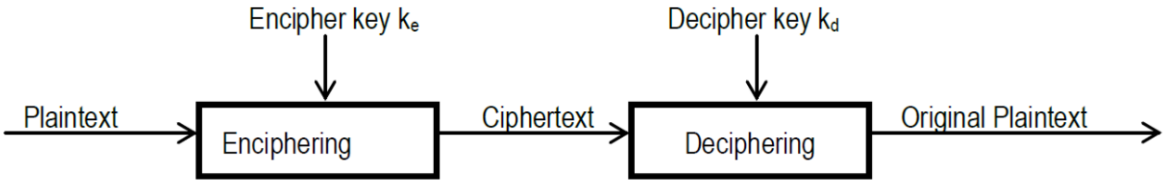


Figure 3. Enciphering/deciphering cycle

Let us denote a set of all possible keys by K and let each $\mathrm{k} \in \mathrm{K}$ can be represented as a tuple $\mathrm{k} = (\mathrm{k_e}, \mathrm{k_d})$, where $\mathrm{k_e}$ is an encipher key and the $\mathrm{k_d}$ is a decipher key. Let P be the set of all possible plaintexts and C be the set of all possible ciphertexts. The enciphering function is denoted by $\mathrm{E_k}: \mathrm{P} \rightarrow \mathrm{C}$ and the deciphering function is denoted by $\mathrm{D_k}: \mathrm{C} \rightarrow \mathrm{P}$.

So a cryptosystem is a five-tuple $(\mathrm{P}, \mathrm{C}, \mathrm{K}, \mathrm{E}, \mathrm{D})$, where the following condition is satisfied:

for each $\mathrm{k} \in \mathrm{K}$, there is an enciphering function $\mathrm{E_k} \in \mathrm{E}$ and corresponding deciphering function $\mathrm{D_k} \in \mathrm{D}$, each $\mathrm{E_k} : \mathrm{P} \rightarrow \mathrm{C}$ and $\mathrm{D_k} : \mathrm{C} \rightarrow \mathrm{P}$ are functions such that $\mathrm{D_k}(\mathrm{E_k}(\mathrm{p})) = \mathrm{p}$ for every plaintext $\mathrm{p} \in \mathrm{P}$.

Later, as mentioned, we will consider that the structure of $\mathrm{E_k}$ is known, or in other words - the safety of data do not depend on the secret structure of the encryption algorithm.

There are two main types of cryptosystems based on the keys: symmetric and asymmetric (public-key).

Symmetric cryptosystems are those where the encryption key can be calculated from the decryption key and vice versa. In most symmetric cryptosystems encryption and decryption keys are the same. These cryptosystems also called secret-key cryptosystems or cryptosystems with a single key require that the sender and recipient have agreed to use the key before secure messaging.

Asymmetric cryptosystems, also called public-key cryptosystems, are those when the encryption key is known but it is practically impossible to calculate the decryption key, even having some additional information (known plaintext attack, known ciphertext attack, etc.) So, in asymmetric cryptosystems $\mathrm{E_k}$ can be safely made public without allowing an adversary to decrypt messages.

3.2. Privacy Homomorphism

The idea of performing simple computations on encrypted data was first put forward by Rivest, Adleman, and Dertouzous [Rivest, 78] who referred to such computations as privacy homomorphism. The original motivation for privacy homomorphism was to allow for encrypted database to be stored by the untrusted second party, while still allowing the owner to perform simple updates and queries such that noting about the database content is revealed to the third party.

Let us consider two algebraic systems to represent plaintext and ciphertext systems. First system is the plaintext system $\mathcal{P}$, which consists from plaintext set P, and some operations $\mathrm{f}_1, \dots, \mathrm{f}_\mathrm{n}$. And the second system is the ciphertext system $\mathcal{C}$, which consists of ciphertext set C, and some operations $\mathrm{g}_1, \dots \mathrm{g}_\mathrm{n}$. For example, the system consisting of integers under the usual set of operations might be denote $< Z, +,\times >$; where Z is set of integers. Formally, the plaintext/ciphertext formalism looks as follows: $\mathcal{P} =< P, \mathrm{f}_1, \dots, \mathrm{f}_\mathrm{n} >$, and $\mathcal{C} =< C, \mathrm{g}_1, \dots, \mathrm{g}_\mathrm{n} >$. We must have also a set of encryption functions $\mathrm{E} = \{\mathrm{E}_\mathrm{k} : \mathrm{P} \to \mathrm{C} \ / \ \mathrm{k} \in \mathrm{K}\}$ and the set of decryption functions $\mathrm{D} = \{\mathrm{D}_\mathrm{k} : \mathrm{C} \to \mathrm{P} \ / \ \mathrm{k} \in \mathrm{K}\}$.

The encryption schema will be called a privacy homomorphism if the following takes place [Brickell, 87; Fontaine, 07]:

$$\forall \mathrm{i} \ \mathrm{D}_\mathrm{k}(\mathrm{g}_\mathrm{i}(\mathrm{E}_\mathrm{k}(\mathrm{a}), \dots, \mathrm{E}_\mathrm{k}(\mathrm{b})) = \mathrm{f}(\mathrm{a}, \dots, \mathrm{b}), \forall \mathrm{a}, \mathrm{b} \in \mathrm{P}.$$

In [Rivest, 78] is brought more general definition which requires that the decryption function will be a homomorphism.

For example, suppose we want to compute $\mathrm{f}_1(\mathrm{a}, \mathrm{b})$. We need only to ask the system to compute $\mathrm{g}_1(\ \mathrm{E}(\mathrm{a}), \mathrm{E}(\mathrm{b}))$. Since the schema is a privacy homomorphism $\mathrm{f}_1(\mathrm{a}, \mathrm{b}) = \mathrm{D}(\ \mathrm{g}_1(\ \mathrm{E}(\mathrm{a}), \mathrm{E}(\mathrm{b})\ )$, so we arrive at the encrypted form of the answer without having to decrypt the intermediate results.

Below the requirements on the choice of the algebraic system $\mathcal{C}$ and the functions E_k, D_k are provided:

1. Encryption and decryption functions, respectively E_k and D_k , should be easy to compute.
2. The operations g_i in ciphertexts system $\mathcal{C}$ should be efficiently computable.
3. An encrypted version of a plaintext d_i , $\mathrm{E}_\mathrm{k}(\mathrm{d}_\mathrm{i})$, should not require much more space to represent than a representation of d_i.
4. Knowledge of $\mathrm{E}_\mathrm{k}(\mathrm{d}_\mathrm{i})$ for many plaintexts d_i should not be sufficient to reveal E_k. (Ciphertext only attack).
5. Knowledge of d_i and $\mathrm{E}_\mathrm{k}(\mathrm{d}_\mathrm{i})$ for several values of d_i should not reveal E_k. (Chosen plaintext attack).
6. The operations and predicates in ciphertext system $\mathcal{C}$ should not be sufficient to provide efficient computation of decryption function D. (This applies primarily to use comparisons).

4. Examples of Privacy Homomorphisms

This section presents some examples of privacy homomorphisms to support the hypothesis that useful privacy homomorphisms may exist for many applications. Moreover, we present cryptanalysis of these examples [Rivest, 78; Brickell, 87]. Mention, that simply, criptosystems RSA and ElGamal are multiplicatively homomorphic [Fontaine, 07].

A. In this example the system of plaintext data is $\mathrm{P} =< \mathrm{Z_{p-1}}, +_{\mathrm{p-1}}>$, system of integers modulo $(\mathrm{p}-1)$ with operation of addition by modulo $(\mathrm{p}-1)$, where p is a prime number such that [Brickell, 87]

$$p-1=\prod_{i=1}^{k} p_i^{d_i}, \text{ and for all } i, p_i \leq B, \text{ for some small } B. \qquad (1)$$

And the corresponding system of ciphertext data is $\mathrm{C} =< \mathrm{Z_n}, \times_\mathrm{n}>$, system of integers modulo n with operation of multiplication modulo n, with the product of p and large prime q. Encryption process is defined as follows: plaintext P is encrypted by computing $\mathrm{E_k(M)} \equiv \mathrm{g^M mod\, n}$, where g is a generator of multiplicative group Z_p^* under multiplication by modulo p (i.e. $\forall \mathrm{a} \in \mathrm{Z_p^*}$ can be represented as $\mathrm{g^i}$ for some integer i) and k is an encryption key, $\mathrm{k} = (\mathrm{p}, \mathrm{q})$. And invers process, decryption, is $\mathrm{D_k(C)} \equiv \log_\mathrm{g} \mathrm{C}\ (\mathrm{mod\ p})$. The structure of p enables computation of the discrete logarithm by the method of Pohlig and Hellman [Pohling, 78] in time $\mathrm{O(B^{1/2})}$.

To show that this schema is a privacy homomorphism let us prove that $\forall \mathrm{M_1}, \mathrm{M_2} \in \mathrm{P}$ $\log_\mathrm{g} \mathrm{g^{M_1}} \times_\mathrm{n} \mathrm{g^{M_2}}\ \mathrm{modp} = \mathrm{M_1} +_{\mathrm{p-1}} \mathrm{M_2}$. We have that $\log_\mathrm{g}(\mathrm{g^{M_1}} \times_\mathrm{n} \mathrm{g^{M_2}}) = \log_\mathrm{g} \mathrm{g^{M_1+M_2}}\ \mathrm{modp}$, where the sum $\mathrm{M_1} + \mathrm{M_2}$ is taken by its usual means (not by modulo p-1). Let us denote $\mathrm{x} = \log_\mathrm{g} \mathrm{g^{M_1+M_2}}\ \mathrm{modp}$. By definition we have that $\mathrm{g^x} \equiv \mathrm{g^{M_1+M_2}modp}$. By definition of g we have that $\mathrm{g^{M_1+M_2}} = \mathrm{g^{M_1+_{p-1}M_2}modp}$. Therefore $\mathrm{x} = \mathrm{M_1} +_{\mathrm{p-1}} \mathrm{M_2}$, which completes the proof.

We will say that the number n is B-powersmooth if each prime power dividing n is less than or equal to B [Pollard, 74].

This system is insecure because (1) allow us to factor n (with high probability) by the Pollard $(\mathrm{p}-1)$ method [Brickell, 87; Cohen, 93]. The basic idea of this factoring algorithm is the following. We have number n and let p be a prime divisor of n. Let $\mathrm{a} > 1$ be an integer such that $\mathrm{GCD(a, n)} = 1$, otherwise consider the factor of n to be found. According to Fermat's little theorem, $\mathrm{a^{p-1}} \equiv 1(\mathrm{modp})$. Let $\mathrm{p}-1$ to be B-powersmooth for a small *B*. Then by definition $\mathrm{p} - 1$ divides the least common multiple of the numbers from 1 to *B,* which we will denote by $\mathrm{lcm}\{1, .., \mathrm{B}\}$. Hence, $\mathrm{a^{lcm\{1,..,B\}}} \equiv 1(\mathrm{modp})$, which implies that $\gcd\left(\mathrm{a^{lcm\{1,..,B\}}} - 1, \mathrm{n}\right) > 1$.

Note that in this case the adversary does not know the plaintext data set, but in public-key cryptosystems there are considered that the adversary knows everything expect the private key.

B. Suppose that the system of plaintext data is the integers modulo p with operation multiplication and test for equality $< \mathrm{Z_p}, \times_\mathrm{p}>$, where p is prime number [Rivest, 78]. A corresponding system of ciphertext data is the integers modulo n with operation multiplication and test for equality $< \mathrm{Z_n}, \times_\mathrm{n}>$, as in the previos example, letting

$\mathrm{n} = \mathrm{p} \cdot \mathrm{q}$, where q is large prime and supposing that n is difficult to factor. The encoding and decoding functions are the same as that used by Rivest, Shamir, and Adleman in their method of implementing public-key cryptosystems. Specifically, the encryption and decryption functions $\mathrm{E_k}$ and $\mathrm{D_k}$ are $\mathrm{E_k(M)} \equiv \mathrm{M^e (mod\, n)}$, for a message M and $\mathrm{D}_K(\mathrm{C}) \equiv \mathrm{C^d (mod\, p)}$, for a ciphertext C, where e and d are integers, such that $\mathrm{GCD}(\mathrm{e}, \varphi(\mathrm{pq})) = 1$ and $\mathrm{ed} \equiv 1 \mathrm{mod} \varphi(\mathrm{pq})$. Recall that by $\varphi(\mathrm{n})$ is denoted the Euler function. The encryption key is thus the pair of positive integer e and n, (e, n). Similarly, the decryption key is the pair of positive integer d and n, (d, n). Each user makes his encryption key public, and keeps the corresponding decryption key private. Since $(\mathrm{x^e})(\mathrm{y^e}) = (\mathrm{xy})^{\mathrm{e}}$, this is a homomorphism. The security of this system should be very good, even if the computer system is given the both e and n.

Note that in this case the adversary does not know the plaintext data set too because the number p is unknown.

C. In this example, the system of plaintext data is $< \mathrm{Z_n}, +_n, \times_n >$, integers modulo n with operations of addition and multiplication by modulo n, where n is the product of two large primes p and q , $\mathrm{n} = \mathrm{p} \cdot \mathrm{q}$. In turn, the system of ciphertext data we take a set $\mathrm{Z_p} \times \mathrm{Z_q}$ and operations are componentwise version of operations on plaintext data (i.e. operations on the first component is performed by modulo p, and over the second - by modulo q). It remains to describe the encryption and decryption functions (process). Encryption function defined as $\mathrm{E_k(a)} = (\mathrm{a \bmod p}, \mathrm{a \bmod q})$. The encryption key k, defined as a pair of numbers p and q, $\mathrm{k} = (\mathrm{p}, \mathrm{q})$. And decryption is: given key $\mathrm{k} = (\mathrm{p}, \mathrm{q})$, decryption function $\mathrm{D_k}((\mathrm{b}, \mathrm{c}))$ is computed using the Chinese remainder theorem.

Now we will show that this method of encryption is Privacy Homomorphism. For simplicity, we prove the case with the addition operation. For multiplication, the proof is similar.

Suppose we have the following plaintexts P1, P2, P3 and their corresponding ciphertexts C1, C2 and C3, which have the form $\mathrm{E_k(P_i)} = \mathrm{C_i} = (\mathrm{c_i^p}, \mathrm{c_i^q})$; for all i, $1 \le \mathrm{i} \le 3$, where

$$c_i^p = P_i\, modp \;\; ; \;\; 1 \le i \le 3 \qquad (2)$$

$$c_i^q = P_i\, modq \;\; ; \;\; 1 \le i \le 3 \qquad (3)$$

More, let $\mathrm{C_3}$ is a sum of $\mathrm{C_1}$ and $\mathrm{C_2}$, i.e. $\mathrm{c_3^p} = (\mathrm{c_1^p} + \mathrm{c_2^p})\mathrm{modp}$ and $\mathrm{c_3^q} = (\mathrm{c_1^q} + \mathrm{c_2^q})\mathrm{modq}$. In this case we need to show that $\mathrm{P_3} = \mathrm{P_1} + \mathrm{P_2 modn}$.

The expressions (2) and (3), they can be written differently:

$$\mathrm{P_i} = \mathrm{c_i^p}\ \mathrm{modp} \;\; ; \;\; 1 \le \mathrm{i} \le 3$$

$$\mathrm{P_i} = \mathrm{c_i^q}\ \mathrm{modq} \;\; ; \;\; 1 \le \mathrm{i} \le 3$$

Based on the properties of modular arithmetic, we can find that

$$P_1 + P_2 = (c_1^p + c_2^p) \text{ modp}$$
$$P_1 + P_2 = (c_1^q + c_2^q) \text{ modq}$$

Finally, since n is a multiple of p and q, then we have the following expressions:

$$(P_1 + P_2)\text{mod n} = \left((c_1^p + c_2^p)\text{modp}\right) \text{modn} = \left(c_1^p + c_2^p\right)\text{modp}$$
$$(P_1 + P_2)\text{mod n} = \left((c_1^q + c_2^q)\text{modq}\right) \text{modn} = \left(c_1^q + c_2^q\right)\text{modq}$$

So we got that $P_3 = (P_1 + P_2)\text{modn}$, that exactly what we wanted to prove.

Unfortunately, this Privacy Homomorphism can be broken, i.e. p and q can be discovered - using a known plaintext attack. Namely, assume that cryptanalyst has plaintext-ciphertext pairs: $P_i, (C_1^p, C_2^q), 1 \leq i \leq r$, where C_i^p and C_i^q computed according to (2) and (3), i.e.

$$C_i^p \equiv P_i \text{ mod p} \quad \text{and} \quad C_i^q \equiv P_i \text{ mod q} \ ; \ 1 \leq i \leq r$$

According (3) follow that p divides $(C_i^p - P_i)$, for all $i = 1,2,3, \dots, r$. Suppose p' is the gcd of numbers $\{(C_i^p - P_i) \mid 1 \leq i \leq r\}$. Since p is a common divisor for the numbers $\{(C_i^p - P_i) \mid 1 \leq i \leq r\}$, then it will be a divisor and for p', i.e. p'/p. Similarly, if q' is GCD of numbers $\{(C_i^q - P_i) \mid 1 \leq i \leq r\}$, then we see that q'/q. And finally, If p'=p and q'=q, then cryptanalyst can decrypt any ciphertext. Even for small r, there is high probability that p'=p, q'=q. Even if it's not as, then if for any new ciphertext cryptanalyst is given the plaintext, he can improve his knowledge of p or q. In particular, given ciphertext (C^p, C^q), the cryptanalyst can find P' such that $P' \equiv C^p \text{ mod } p'$ and $P' \equiv C^q \text{ mod } q'$. If P' ≠ P, then follow that $P \neq C^p \text{ mod } p'$ or $\neq C^q \text{ mod } q'$. So if cryptanalyst given P, it can improve the value of p' and q' by replacing p' by $\text{GCD } (P - C^p, p')$ and q' by $\text{GCD } (P - C^q, q')$.

D. For this example as a system of plaintext data we take the set of integers under the usual operations of addition and multiplication, $P = < Z, +, \times >$. Encryption process is performed as follows: the user chooses an integer n and represents all of his data in radix-n notation, i.e. if $y = d_m n^m + d_{m-1} n^{m-1} + \cdots + d_1$ then the n-radix of y will be the vector $(d_m, \dots, d_1)$.. Corresponding operations on ciphertext data defined the same as in the case of algebraic polynomials by allowing individual coordinate positions to exceed n. For example, if n=15, we have

$E_k(42)$ = (2 , 12), $E_k(23)$ = (1 , 8), $E_k(65)$ = (3 , 20), $E_k(966) = (2,28,96)$

where k is encryption key, which defined by the integer n, k = (n). Easy to see that the computer system can operate on ciphertext data without knowledge of n, this means that the encryption is Privacy Homomorphism. Without loss of generality, we will show that in the case of the addition operation. To do this we need to show that, if we have two plaintexts P1 and P2, for which:

$$E_k(P_1) = (a_r, a_{r-1}, \dots, a_1) \tag{4}$$

$$E_k(P_2) = (b_s, b_{s-1}, \dots, b_1) \tag{5}$$

$$E_k\,(P_1) + E_k(P_2) = (c_m, c_{m-1}, \dots, c_1)$$

Then we need to show that the following equality holds: $(\mathrm{P}_1 + \mathrm{P}_2) = D_k\,((\mathrm{c}_\mathrm{m}, \mathrm{c}_{\mathrm{m}-1}, \dots, \mathrm{c}_1))$, where m = max(r, s) and c_i are the result of adding the values of the corresponding coordinates. For that let calculate the value of$(\mathrm{P}_1 + \mathrm{P}_2)$. From (4), (5) follow that P_1 and P_2 can be represented as follows:

$$P_1 = a_r \cdot n^{r-1} + a_{r-1} \cdot n^{r-2} + \dots + a_2 \cdot n + a_1 \tag{6}$$

$$P_2 = b_s \cdot n^{s-1} + b_{s-1} \cdot n^{s-2} + \dots + b_2 \cdot n + b_1 \tag{7}$$

According to (6) and (7) follows that $(\mathrm{P}_1 + \mathrm{P}_2)$ is equal to $\mathrm{P}_1 + \mathrm{P}_2 = \mathrm{c'}_{\mathrm{m'}} \cdot \mathrm{n}^{\mathrm{m}-1} + \mathrm{c'}_{\mathrm{m'}-1} \cdot \mathrm{n}^{\mathrm{m}-2} + \dots + \mathrm{c'}_2 \cdot \mathrm{n} + \mathrm{c'}_1$. From the last equality and the definitions of coordinates c_i and m it follows that $\mathrm{m'}$ coincide with m and coefficients $\mathrm{c'}_\mathrm{i} = \mathrm{c}_\mathrm{i}$, for all $\mathrm{i} = 1,2, \dots$. And this in turn proves that $(\mathrm{P}_1 + \mathrm{P}_2) = D_k((\mathrm{c}_\mathrm{m}, \mathrm{c}_{\mathrm{m}-1}, \dots, \mathrm{c}_1))$.

As in the previous example, this system also insecure and can be broken by a known plaintext attack. Assume that the cryptanalyst has plaintext-ciphertext pairs P_i , $(\mathrm{c}^\mathrm{i}_{\mathrm{k}_\mathrm{i}}, \mathrm{c}^\mathrm{i}_{\mathrm{k}_{\mathrm{i}-1}}, \dots, \mathrm{c}^\mathrm{i}_1)$ for all $\mathrm{i} = 1,2, \dots, \mathrm{s}$. Based on (6) and (7) follows that $(\mathrm{P}_\mathrm{i} - \mathrm{c}^\mathrm{i}_1)/\mathrm{n}$. Let n' be the GCD of numbers $\{\mathrm{P}_\mathrm{i} - \mathrm{c}^\mathrm{i}_1 \mid 1 \le \mathrm{i} \le \mathrm{s}\}$. Since n is common divisor for numbers $\{\mathrm{P}_\mathrm{i} - \mathrm{c}^\mathrm{i}_1 \mid 1 \le \mathrm{i} \le \mathrm{s}\}$, then this means that n is divisor for n', $\mathrm{n'}/\mathrm{n}$ .If $\mathrm{n} = \mathrm{n'}$, the cryptanalyst can decrypt any ciphertext. Even for small s, there is a high probability that $\mathrm{n} = \mathrm{n'}$. Even if it's not the case, then for any new ciphertext cryptanalyst is given the plaintext, he can improve his knowledge of n. Specifically, given ciphertext $(\mathrm{c}_{\mathrm{k}_1},\ \mathrm{c}_{\mathrm{k}_1-1}, \dots, \mathrm{c}_1)$, the cryptanalyst can find $\mathrm{P'}$ such that $\mathrm{P'} \equiv \mathrm{c}_1 \bmod \mathrm{n'}$. If $\mathrm{P'} \ne \mathrm{P}$, then follow that $\mathrm{P} \ne \mathrm{c}_1 \bmod \mathrm{n'}$. So if cryptanalyst given P, it can improve the value of n' by replacing it with $\mathrm{GCD}(\mathrm{P} - \mathrm{m}_1, \mathrm{n'})$.

E. As in the previous example, as a system of plaintext data we take the set of integers under the usual operations of addition, and multiplication $\mathrm{P} = < Z, +, \times >$. More, let $\mathrm{a}_0, \mathrm{a}_1, \dots, \mathrm{a}_{\mathrm{n}-1}$ by randomly chosen positive integers and A be the matrix:

$$A = \begin{pmatrix} 1 & \cdots & a_0^{n-1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & a_{n-1}^{n-1} \end{pmatrix}, \tag{8}$$

where n is chosen so that all intermediate results used in any calculation are less than 2^{n}. Encryption function defines as follows: given plaintext P, whose binary representation is $\overline{\mathrm{P}} = (\mathrm{p}_0, \mathrm{p}_1, \dots, \mathrm{p}_{\mathrm{n}-1})_2$, the encryption of P is the column vector $\overline{\mathrm{C}}$:

$$\overline{\mathrm{C}} = \mathrm{E_k(P)} = \mathrm{A} \cdot \overline{\mathrm{P}} = \begin{pmatrix} \mathrm{f_p(a_0)} \\ \vdots \\ \mathrm{f_p(a_{k-1})} \end{pmatrix}, \qquad (9)$$

where k is the encryption key, which constructed from the integers $\mathrm{a_0}, \mathrm{a_1}, \dots, \mathrm{a_{n-1}}$, $\mathrm{k} = (\mathrm{a_0}, \mathrm{a_1}, \dots, \mathrm{a_{n-1}})$. And function $\mathrm{f_p(x)}$ defined as follows: $\mathrm{f_P(x)} = \sum_{\mathrm{i=0}}^{\mathrm{k-1}} \mathrm{p_i} \cdot \mathrm{x^i}$.

Operations on encrypted data are component wise version of addition and subtraction over the integers. Decryption is performed by multiplying ciphertext (column vector) $\overline{\mathrm{C}}$ by matrix A^{-1}: $\mathrm{D_k}(\overline{\mathrm{C}}) = \mathrm{A}^{-1} \cdot \overline{\mathrm{C}}$. This encryption is Privacy Homomorphism. It is true, let we have two plaintexts $\mathrm{P_1}, \mathrm{P_2}$ and their corresponding ciphertexts are $\overline{\mathrm{C_1}} = \mathrm{E_k(P_1)} = \mathrm{A} \cdot \overline{\mathrm{P_1}}$ and $\overline{\mathrm{C_2}} = \mathrm{E_k(P_2)} = \mathrm{A} \cdot \overline{\mathrm{P_2}}$, where $\overline{\mathrm{P_1}}$ and $\overline{\mathrm{P_2}}$ are binary representations of $\mathrm{P_1}, \mathrm{P_2}$. We need to show that, if $\overline{\mathrm{C_1}} \pm \overline{\mathrm{C_2}} = \overline{\mathrm{C_3}}$, then should take place the following equality: $\mathrm{P_1} \pm \mathrm{P_2} = \mathrm{D_k}(\overline{\mathrm{C_3}})$. To this end, consider the expression $(\overline{\mathrm{C_1}} \pm \overline{\mathrm{C_2}})$. According to (8) and (9), we find that $(\overline{\mathrm{C_1}} \pm \overline{\mathrm{C_2}}) = \mathrm{A} \cdot \overline{\mathrm{P_1}} \pm \mathrm{A} \cdot \overline{\mathrm{P_2}}$, but $\overline{\mathrm{C_1}} \pm \overline{\mathrm{C_2}} = \overline{\mathrm{C_3}}$. Thus, it turns out that

$$\overline{\mathrm{C_3}} = \mathrm{A} \cdot \overline{\mathrm{P_1}} \pm \mathrm{A} \cdot \overline{\mathrm{P_2}} = \mathrm{A} \cdot (\overline{\mathrm{P_1}} \pm \overline{\mathrm{P_2}})$$

And this means that $(\overline{\mathrm{P_1}} \pm \overline{\mathrm{P_2}}) = \mathrm{A}^{-1} \cdot \overline{\mathrm{C_3}} = \mathrm{D_k}(\overline{\mathrm{C_3}})$. This equality proves that this encryption is Privacy Homomorphism.

This encryption is insecure and can be broken by ciphertext only attack. But first, let us consider one interesting fact. Suppose we are given an integer x and its binary representation $\overline{\mathrm{x}} = (\mathrm{x_0}, \mathrm{x_1}, \dots, \mathrm{x_{n-1}})_2$. Moreover, suppose j is the largest index in binary representation $\overline{\mathrm{x}}$ such that $\mathrm{x_j} = 1$, i.e. $\mathrm{x_{j+1}} = \mathrm{x_{j+2}} = \dots = \mathrm{x_{k-1}} = 0$. Let z be any positive integer, then we have the following inequality: $\mathrm{z^j} \le \sum_{\mathrm{i=0}}^{\mathrm{j}} \mathrm{x_i} \cdot \mathrm{z^i} < \sum_{\mathrm{i=0}}^{\mathrm{j}} \mathrm{z^i} < (\mathrm{z}+1)^{\mathrm{j}}$. This means that $\left\lfloor (\sum_{\mathrm{i=0}}^{\mathrm{j}} \mathrm{x_i} \cdot \mathrm{z^i})^{\frac{1}{\mathrm{j}}} \right\rfloor = \mathrm{z}$. From this fact follows that this encryption is weakens to ciphertext only attack. Suppose cryptanalyst has a cipher $\overline{\mathrm{C}} = (\mathrm{c_0}, \dots, \mathrm{c_{n-1}})$ and he guess a value for j, the largest index such that $\mathrm{x_j} = 1$. Then he can compute $\left\lfloor \mathrm{c_0}^{1/\mathrm{j}} \right\rfloor = \mathrm{b_0}$ and write $\mathrm{c_0}$ in base $\mathrm{b_0}$ notation. If all coefficients are 0 and 1, then probably $\mathrm{b_0} = \mathrm{a_0}$ and x is easily found. More, values $\mathrm{c_1}, \mathrm{c_2}, \dots, \mathrm{c_{k-1}}$ can be used as an additional check. Otherwise, cryptanalyst can try a different choice for j.

There are known some security considerations for encryption systems [Fontaine, 07]. According [Fontaine, 07] the highest level of security which can be achieved by homomorphic encryption system is IND-CPA (Indistinguishability chosen plaintext attack).

In 2009 IBM (his employee Craig Gentry) presented the new Fully Homomorphic scheme - i.e. a scheme that allows one to evaluate circuits over encrypted data without being able to decrypt [Greaig, 09]. Gentry's scheme is completely impractical. It uses something called an ideal lattice as the basis for the encryption scheme, and both the size of the ciphertext and the complexity of the encryption and decryption operations grow enormously with the number of operations you need to perform on the ciphertext - and that number needs to be fixed in advance. And converting a computer program, even a simple one, into a Boolean circuit requires an enormous number of operations.

The other example of homomorphic encryption is brought in [Khan, 12], mathematical base of which is p-adic rings, but there are different opinions about the practical value of this system.

5. Complementary means of privacy computations

The need for new methods of disclosure limitation is further developed in the light of the modern advances in data mining, which can be used to undo the mask or to get narrow interval estimates for the original values. Data mining techniques aimed at the level of individual record pose a big threat to the respondents' privacy. [Rivest, 78] is one of the early referred papers the direction of privacy preserving data mining. These approaches however, focus on the secure multiparty computation. Existing literature on the protection of statistical databases against data mining attacks in the literature are very few, and can be characterized as the first steps made in this direction [Little, 93]. Current research projects aim to account the threats posed by data mining techniques by developing methodology of protection and using data mining to create masked data. Alternatively the project aims to apply data mining technique in synthetic data generation and in data distortion for general disclosure limitation.

Specifically, the research will result in the following:

- Methods of assessing attribute disclosure risk for different scenarios of data release. Intruders' actions will be simulated by the application of the relevant data mining technologies to the data sets, masked by different SDL methods;
- Methods, which reduce, attribute disclosure risk, based on the results of the previous step;
- Methods, which use local masking, approach to improve overall data utility and protect particular groups of individuals, e.g. outliers;
- Methods of masking for the data sets with special structural relationships between the variables, which should be preserved in the released, masked data.

These results will improve the quality of the released data and protect the confidentiality of individual information. Viewed more broadly, the results will facilitate better dissemination of different types of statistical data.

There is a variety of techniques in data mining, which can be used as classifiers for categorical variables and/or predictors for quantitative variables. These tools can be applied by the illegitimate data users to a particular record or a group of records with the goal of obtaining narrow interval estimates of the sensible variable. One of such tools is rule mining [Agrawal, 00; Aslan, 04].

Association Rule Mining (ARM) is the most popular data mining technique in the line of rule-based models, Incremental Reduced Error Pruning (IREP) and Frequent Fragments Mining. Assume we have a set $I =$

$\{I_1, I_2, \dots, I_n\}$ of n different attributes and let $X \subseteq I$. Given a database D with records of type X, transactions. We say that $T \in D$ supports X, if $X \subseteq T$. Consider the standard concepts of support and confidence

$$supp(X) = |\{T \in D | X \subseteq T\}|/|D|$$

$$conf(X \to Y) = supp(X \cup Y)/supp(X)$$

An association rule is the expression of form $\{X \to Y$, confidence, support$\}$, where $X \cap Y = \emptyset$. Here support is simply the proportion of records in D that contain $X \cup Y$ and confidence is the proportion of records containing X for which the rule is correct. The problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence.

The problem of discovering all association rules can be decomposed into two sub-problems:

1) Find all subsets from that have support above minimum support frequent subsets.
2) Use the found subsets to generate rules.

The first step can be accomplished using the following iterative algorithm. In the first round the algorithm counts the support of individual attributes and determines which of them have minimum support. In each subsequent pass, the algorithm starts with a set of attributes found in the previous round and uses this set and single item additions for generating new potentially frequent subsets, called candidates, and counts the actual support for these subsets. At the end of the pass, it is determined which of the candidate subsets are actually frequent, and they are used for the next pass. The process continues until no new frequent subsets are found. There is a well-known algorithm in this domain, APRIORI, which is factually the de facto standard for ARM. Algorithm is recursive. The basic intuition is that any subset of a frequent subset must be frequent which monotonicity type property is. Therefore, the candidate subset having k elements can be generated by joining frequent subsets having $k - 1$ elements, and deleting those that contain any subset that is not frequent. This procedure results in generation of a much smaller number of candidate subsets.

For solving the second sub-problem the following algorithm can be used. For every frequent subset X, find all non-empty subsets $a \subseteq X$ and output a rule of the form $a \to (X - a)$ if the ratio of support of X to support of a (the confidence of the rule) is greater than minimum confidence.

Finding all frequent item sets in a database is difficult since it involves searching almost all possible item sets and uses several passes over these data. The set of possible item sets is the power set of I and this algorithmic problem is evidently NP hard. Even the system of maximal possible frequent itemsets, which is a Sperner system, is large. There are several critical issues related to ARM.

Privacy-preserving ARM starts with information, which may contain intentionally wrong survey information items, as it is the case of masked data. The reconstructed support values cannot coincide exactly with the actual supports, and errors positive or negative may occur, having more pernicious effect than just a wrong cipher, enlarging or narrowing the set of rules mined. Often the rule sets produced are large, but most of them are uninteresting. Control of number of associations produced by change of support may reduce them to a manageable number, but this may lead to the loss of interesting rules. Support threshold alone is not enough to find interesting structures. Symmetric combinations of attributes may lead to a large set of similar or unnecessary rules mined. At this point ARM recovers many associations between the attributes, which do not have cause-and-effect relationship. There are a number of proposals, but they lack the systematic approach. It turns out that the same mathematical tool can address such behavior effectively.

The model, which will be investigated, is based on a mechanism called chain split and computations [Aslan, 09; Tonoyan, 76]. This approach allows generating the frequent subsets using the minimal possible amount of memory. Chain split is a special partition of n-cube vertices set into the growing chains. The special property provides that the 3-chain fragments with complementary vertex α may be determined by α and this is the key of chain computations. At the moment algorithms are computing and saving the maximal frequent item sets there is an inside potential to consider extended chain computation, which uses lager chain fragments, which leads to approximate results.

Moreover, thinking from point of view of monotone Boolean functions, it is important to construct the absorbing function of limited complexity, which may play the role of best approximation scaling the frequent sets density and uniting several sets of upper, zero points into the imputed values. Specifically, in chain split data mining for privacy preserving use, it may be given general limitations for frequent subsets in term of their sizes. This requires extending the chain split to these conditions, which allows finding more effective chain computation algorithms. Chain computation algorithms will be extended to serve such constructions and likely an approximate threshold will be introduced. Relative compliments on chains will be modified from distance 2 to higher distances. Expected results of this part of the project are monotone Boolean approximations. These represent a new approach with high value due to massive applications of monotone Boolean functions in different areas. Such mechanism seems to be necessary to generate frequent sets effectively when a set of input datasets are to be designed and analyzed from the point of view of a hacker actions. The approximate computation is hard computationally and heuristic in its nature.

Bootstrap and boosting approaches. To increase an accuracy of the prediction and classification the intruder can use bootstrap aggregation or bagging. In particular, for iteration i $(i = 1, 2, \ldots, k)$, a training set D_i, of d records is sampled with replacement from the original set of records, D. A classifier model M_i, is learned for each training set, D_i. To classify an unknown record, X, each classifier, M_i, returns its class prediction, which counts as one vote. The bagged classifier, M_*, counts the votes and assigns the class with the most votes to X. If the intruder's goal is the prediction of the continuous variable, bagging can be applied by taking the average value of each prediction for a given record. The bagged classifier often has significantly greater accuracy than a single classifier derived from D, the original training data. Bagged classifier is also quite robust to the effects of noisy data. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers [Aslan, 13].

For prediction, it was theoretically proven that a bagged predictor will always have improved accuracy over a single predictor derived from D.

Another approach to increase accuracy of prediction is boosting. Classical example here is Adaboost algorithm. This algorithm assigns weights to each training record. Usually on the initial stage these weights are equal to $1/d$

for all the records in the training set D with d records. Then in the k rounds, set D is sampled with replacement according to the records' weight to form a training set D_i and this set is used to derive a classifier model M_i. After a classifier M_i is learned, its error is computed and the weights for the misclassified records are increased, so that the subsequent classifier M_{i+1}, "pay more attention" to these misclassified records. The final boosted classifier M_*, combines votes of each individual classifier. Each classifier has got a weight: the lower a classifier's error rate, the more accurate it is, and therefore, the higher it's weight for voting. For each class c, the weights of each classifier that assigned the record to a class c are summed. The class with the highest sum is the "winner" and is returned as the class prediction for this record.

Application of data mining to the original data by data protector (agency) opens the door to many possibilities in Statistical Disclosure Limitation. Data mining algorithms, in particular association rule mining can be applied to discover those relationships, which are not obvious. These relationships are not necessarily strict, in the sense that they may be satisfied for most of the records, but not necessarily for all the records. Algorithms for frequent set finding can be used to discover such relationships. Our research will focus on the application of these algorithms for discovery of such relationships, but most importantly for the creation of masked data sets.

So, to be summarized, the research agenda of the second part of the project is focused on the developments of the following algorithms:

1) Algorithms of the identification of the zones (groups of records) which should be masked with different SDL methods.
2) SDL Technique which can be applied to positive variables and also to the variables than can take positive as well as negative values.
3) Combinations of SDL methods which would offer adequate protection for the records in different zones, including methods for outlier’s protection.

Acknowledgment

Authors would like to thank Gurgen Khachatryan and Anna Oganyan for many important remarks and valuable discussions.

Bibliography

[Agrawal, 00] Agrawal, R. and Srikant, R. “Privacy-Preserving Data Mining,” Proc. ACM SIGMOD Conf. Management of Data, ACM Press, 2000, pp. 439–450; http://doi.acm.org/ 10.1145/342009.335438.

[Aslan, 04] L. Aslanyan, V. Topchyan, Hierarchical Cluster Analysis For Partially Synthetic Data Generation, Transactions of IIAP of NAS of RA, Mathematical Problems of Computer Science, submitted, 2013.

[Aslan, 09] Aslanyan, L. and Sahakyan, H. "Chain Split and Computations in Practical Rule Mining", International Book Series, Information Science and Computing, Book 8, Classification, forecasting, Data Mining, pp. 132-135, 2009.

[Aslan, 13] L. H. Aslanyan, H. E. Danoyan, “On the optimality of a hash-coding type search algorithm”, Proceedings of the 9th conference CSIT, Yerevan, Armenia, pp. 55-57, 2013

[Brickell, 87] E. F. Brickell and Y. Yacobi, On Privacy Homomorphisms (Extended Abstract), Advances in Cryptology - EUROCRYPT '87, Workshop on the Theory and Application of Cryptographic Techniques, Amsterdam, The Netherlands, April 13-15, 1987, pp. 117-125.

[Cohen, 93] H. Cohen, A Course in Computational Algebraic Number Theory, Springer, 1993, 580 p.

[Dalenius, 82] Dalenius, T. and Reiss, S. P. "Data-swapping: A technique for disclosure control", Journal of Statistical Planning and Inference, volume 6, pages 7385, 1982.

[Defays, 93] Defays, D. and Nanopoulos, P. "Panels of enterprises and confidentiality: the small aggregates method", in Proceedings of the 92 Symposium on Design and Analysis of Longitudinal Surveys, pages 195204, Ottawa. Statistics Canada, 1993

[Duncan, 91] Duncan, G. T. and Pearson, R. W. Enhancing access to microdata while protecting confidentiality: Prospects for the future. Statistical Science, 6:219239, 1991

[Feinberg, 06] Fienberg, S. "Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation", Statistical Science, 21, 143-154, 2006

[Ferrer, 01a] Domingo-Ferrer, J., Mateo-Sanz, J. M., and Torra, V. "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", In Proc ETKNTTS 2001, pages 807 - 825, Luxembouorg. Eurostat, 2001

[Ferrer, 01b] Domingo-Ferrer, J. and Torra, V. "A quantitative comparison of disclosure control methods for microdata." In Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L., editors, Confidentiality, Disclosure and Data Access, pages 111133. NorthHolland, Amsterdam, 2001.

[Ferrer, 96] J. D. Ferrer, “A new privacy homomorphism and applications,” Information Processing Letters, vol. 60, no. 5, pp. 277-282, 1996

[Fontaine, 07] C. Fontaine and F. Galand, A Survey of Homomorphic Encryption for Nonspecialists, EURASIP Journal on Information Security 2007, pp. 1-10.

[Greaig, 09] Greaig Gentry, A Fully Homomorphic Encryption Scheme, 2009

[Khan, 12] Z. Khan, Quasi-Linear Time Fully Homomorphic Public Key Encryption Algorithm (ZK111), Journal of Theoretical Physics & Cryptography, vol. 1, November 2012, pp. 14-17.

[Kim, 86] Kim, J. J., A method for limiting disclosure in microdata based on random noise and transformation. In Proceedings of the ASA Section on Survey Research Methodology, pages 303-308. Alexandria V.A. American Statistical Association, 1986

[Little, 93] Little, R. J. A. "Statistical analysis of masked data", in Journal of Ofcial Statistics, 9:407426, 1993.

[Loureiro, 04] Loureiro, A., Torgo, L. and Soares, C., "Outlier Detection Using Clustering Methods: a data cleaning application" , Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector, 2004

[Miyoung, 13] Miyoung Jang, Min Yoon, and Jae-Woo Chang A k-Nearest Neighbor Search Algorithm for Privacy Preservation in Outsourced Spatial Databases, ISA 2013, ASTL Vol. 21, pp. 223 - 226, 2013.

[Moore, 96] Moore, R. "Controlled data swapping techniques for masking public use microdata sets." U. S. Census Bureau, 1996

[Pohling, 78] Stephen C. Pohling and Martin E. Hellman, An Improved algorithm for computing Logarithms over GF (p) and its cryptographic significance, IEEE trans. on Inf. th. Vol. IT-24, No. 1 Jan. 1978. pp. 160-110.

[Pollard, 74] J. M. Pollard, ‘‘Theorems on factorization and primality testing”, Roc. Cambridge Philos. SOC. vol. 76, 1974. pp. 521-528.

[Reiter, 02] Reiter, J. P., "Satisfying disclosure restrictions with synthetic data sets," Journal of Official Statistics 18 (2002) 531-544.

[Reiter, 05] Reiter, J. P., "Using CART to generate partially synthetic public use microdata.", in Journal of Official Statistics, 21, 441 - 462, 2005.

[Rivest, 78] Ronald L. Rivest , Len Adleman, Michael L. Dertouzous, On data Banks And Privacy Homomorphisms, in : R.A. DeMillo et al., eds, Foundations of Secure Computation(Academic Press, New York, 1978) 169-179.

[Sanz, 99] Mateo-Sanz J.M. and Domingo-Ferrer J, "A method for data-oriented multivariate microaggregation", in Proceedings of Statistical Data Protection'98, Luxembourg: Office for Official Publications of the European Communities, pp. 89-99, 1999.

[Tendik, 94] Tendik, P. and Matlof, N. "A modified random perturbation method for database security", ACM Transactions on Database Systems, 19(1):4763, 1994.

[Tonoyan, 76] Tonoyan, G. "Chain decomposition of n dimensional unit cube and reconstruction of monotone Boolean functions", JVM and F, v. 19, No. 6, 1532-1542, 1976

[Wallman, 04] Wallman, K. K. and Harris-Kojetin, B. A. "Implementing the confidential information protection and statistical efficiency act of 2002", Chance, 17(3):2125, 2004.

Authors' Information

Levon Aslanyan *– ITHEA ISS, Sofia, Bulgaria; Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail: lasl@sci.am*

Major Fields of Scientific Research: Discrete optimization, Artificial intelligence, NLP, WSN, Privacy preserved computation

Vardan Topchyan *– Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail:vardan.topchyan@gmail.com*

Major Fields of Scientific Research: Decision models, Homomorphic encryption, Privacy preserved computation

Haykaz Danoyan *– Institute for informatics and automation problems of NAS RA, 1, P. Sevak street, Yerevan 0014, Armenia, e-mail: hed@ipia.sci.am*

Major Fields of Scientific Research: NN Search, Discrete optimization, Coding theory, Homomorphic encryption

A NEW APPROACH TO INFORMATION SCIENCE: FRAMING A NATURALISTIC PERSPECTIVE

Jorge Navarro, Raquel del Moral & Pedro C. Marijuán

Abstract: *In recent decades, an increasing variety of research fields are converging into information science conceptualizations. They are accompanied by astounding new uses of knowledge and even more astounding social transformations that revolve around information technologies. Whether a robust information science will finally emerge may not only depend on successful discussions about the philosophy of information and the social impact of the new technologies. The most important adjustment to make is about framing a "new way of thinking" or new perspective about information science itself: an inner philosophy interconnecting research practices in fundamental areas of the new science. Like in the historical birth of other major sciences, the empirical, comparative understanding of informational phenomena and informational entities should take place first. Further, a naturalistic perspective, biologically inspired, would help in determining what model systems should be adopted for advancing the comparative study of informational entities. Rather than attempting ill-fated definitions [is information definable at all?] or closely remaining within the confines of information theory, a naturalistic, empirically oriented strategy offers room for advancement. In the end, crafting a great scientific domain around information science –rather than around computing – should be a common goal for the scholars and researchers both from natural sciences and engineering and from social sciences involved in these new studies.*

Keywords: *Information Flow, Cell Production, Communication, Information Science,*

ACM Classification Keywords: *Theory*

The informational transformation of contemporary science

The emergence of a renewed information science along the stunning *informational uprising* of contemporary societies would hardly be a surprise. The historical parallel with the emergence of thermodynamics in the generations closer to the industrial and social experimentation with steam-engines provides a sensible argument. As a matter of fact, most of scientific growth during the last two decades has been produced in information-related fields: bioinformatics and *omic* revolution, neuroinformatics, robotics, artificial intelligence, quantum information, new computing tools, models and simulations, virtual economy, virtual learning, and so on. The stunning transformation of many branches of science and technology fueled by computers, including physics itself, is leading to reconsider the natural structures and processes of information, not only in the quantum-cosmological realms or in the biomolecular, or in the human-to-machine interaction, but particularly in the human-to-human interrelationship and in the bonding structures of societies as the great drivers of global communication networks.

At the time being, the efforts to build a coherent discipline around the manifold scientific manifestations of information are only partially succeeding, but they are gaining more and more momentum and should be contemplated with high hopes and a lofty ambition. Rather than attempting the construction of another average or standard discipline, information science is about the making out of one of the “great scientific domains” of contemporary knowledge [Rosenbloom, 2013]. It is information science –rather than computer science and technology as argued by Rosenbloom– the strategic domain that has the mandate to perform the greatest cognitive interrelationships that our social system of knowledge demands: the natural sciences with the social sciences and humanities; information with meaning, knowledge, and intelligence; the individual creation of knowledge with its social organization and management; the limitations of human communication with social complexity and political structures; the anthropogenic economic flows with the geogenic planetary flows.

Providing the next generation with a more coherent and workable system of knowledge would be a valuable legacy from our generation indeed—by having launched in the chaotic sea of disciplines an *information hub* of fresh design. In this regard, the Foundations of Information Science [FIS] initiative was launched almost two decades ago with the explicit goal of developing a ‘vertical’ or ‘transdisciplinary’ science connecting the different threads and scales of informational processes [Conrad, 1996; Marijuán, 1996]. The five FIS conferences held until now represent a valuable experience to guide in the transformations that have to be made within information science itself. Actually, the FIS meeting in the 2014 ITHEA GIT Conference represents a celebration of FIS 20th anniversary.

Can information be defined? Advancing a new way of thinking

Rather than discussing on the hundreds of information definitions proposed [Lenski, 2010], let us definitely discard that. Information cannot have a universal definition because it corresponds to the open-ended interactions that a subject can engage with whatsoever elements or objects in its environment. Information pertains to an unmediated coupling between objects-subjects, and involves their contact. We demand either that a direct physical contact be established between the pair object-subject, or that a “channel” be arranged to bring to the adjacency of the subject those objects or their very effects.

Advocating the undefinability of information does not mean the ineffability of the term; rather what it means from a naturalistic perspective is that the plurality of subjects (can’t we talk about information in cells, in organisms, in individuals, or in social bodies as legitimate subjects?), plus the vastness of possible coupled objects, and the multiplicity of coupling modes between subjects and objects do not permit any universalistic definition. Achieving a universal information definition and even more a unified theory of information represents a similar attempt to the “universal non-linear theory” that John von Neumann was quipping as “the non-elephants theory”.

Changing the main analytical focus from the “name” information as an object in itself to the “adjective” informational, what entities might be attributed the quality or characteristic of being informational? The response is concise: those capable of both *communication* and *self-production*. The first trait is well known in many different fields related to information science—apparently. From our perspective, however, the communication theme demands full clarification. In the same way that the energetic-metabolic aspect of biological self-production was better understood once the *energy flow* was properly recognized and characterized [Morowitz, 1968], the *information flow* of communication demands its own general conceptualization, beyond current conceptions restricted to specialized uses.

How the informational entity organizes its openness towards the environment in the communication exchanges? The capture of the surrounding signaling flows is propitiated by *ad hoc* designs that tend to cover increased regions of space throughout special sensing structures, incorporating fractal distributions for instance. In general, the communication trait has been evolved adaptively: exploiting the maximum surrounding flows with the highest possible diversity of objects/events, using the lowest capture energy, and applying the faster processing rhythms compatible with the own structures. This can be observed in cellular signaling systems, in nervous systems, and in the communication systems of human organizations and in societies at large. The external communication flow, or better, the *information flow,* antedates and permeates the evolution of whatsoever structures of complexity.

In the other essential trait of the informational entity, *self-production*, streams or flows of objects, destined to fuel the self-production processes, are introjected or channeled inwards to be consumed—also appearing a parallel flow of extrojected material outputs. The ingoing and outgoing flows constitute the metabolism of the self-producing informational entity. As said, the *energy flow* has been well-characterized for the biosphere and ecosystems [Morowitz, 1986]; technically it is also a crucial ingredient of industrial engineering, urban planning, and economic management. From this energetic point of view, the informational entity appears as an open system, out from equilibrium, endowed with the required energy-entropy-matter flows, and often involving the generation of well-ordered dissipative structures. But these flows do not spontaneously run to encounter the entity; rather they have be searched, channeled, and internalized by following optimized behavioral and search strategies [Holland, 2012].

Although the information flow and the energy flow may partially overlap in their constitutive components, they are treated in a highly different way: "reading" the environment becomes utterly different from, and prior to, "eating" it. The living cell is the clearest instance [Marijuán, 2010]. In general, the high-energy, highly valuable self-production flows will be anticipated, detected, and captured by the faster and cheaper communication flows tended with the surrounding environment. A frequent commonality of forms occurs too, manifested in supporting structures that often display fractal forms derived from the necessity to cover a region of space and to transport the affordances of both the material "stuff" and the communicational "fluff" to a center [Bejan and Peder, 2012].

Around *the information flow*, a new conceptual framework might be established. At its core, information science should deal with informational entities that exist "in the flow", that both communicate and self-produce. By intertwining the two operational realms a seamless existential unity is generated. As Lanham [2006] has put for our societies, it is the almost immaterial *fluff* of communication which provides guidance to the constitutive material *stuff*. And vice versa, it is social life which provides the functions and goals of communication systems—always in service of social self-production. Hence an "economy of attention" emerges: how the networks of self-production processes look at the world for their survival.

This informational approach can be generalized upwards. Actually every new organization realm develops some specific information flow; at the frontiers or junctures between compartments the previous, more basic information flow is maintained. This successive composition of information layers seems valid for organisms, individuals and societies. However, a homogeneous description is out of hand—all these informational entities belong to conceptual disciplines worlds apart [Marijuán, 2010]. A parsimonious strategy, herein advocated, may consist in attempting a relatively independent description for each one of the scales, to be followed then by some tentative abstractions/conclusions interrelating them "vertically" [Conrad, 1996; Marijuán, 1996].

This new perspective is both naturalistic and empirically oriented. Its focus is in the autonomous existing informational entities, taking as their distinctive characteristic the intertwining of communication and self-production flows. Natural informational entities to study as model systems are cells, organisms [nervous systems], individuals, and societies. In spite of the disciplinary differences and conceptual obstacles, a vertical understanding of the communication and self-production flows seems feasible. The realization of multiple empirical studies on informational aspects of these natural entities will help to advance the common understanding.

The primordial informational entity: the living cell

Prokaryotic cells will be taken as a practical case (almost the single one in this brief paper) to discuss the fundamentals of the new perspective on information science. These cells contain perhaps one of the most interesting panoramas for informational analysis: where at least nearly every intervening phenomena can be molecularly described. In the description that follows the language has to switch necessarily towards the molecular-biological.

How different looks the animate from the inanimate! Just counting the number of different molecular species teeming up at the interior of a prokaryotic cell (the multiple classes of peptides, enzymes, receptors, phospholipids, RNAs, DNA, metals, nutrients, ions, etc.) the figure is staggering: in the order of 10,000 species, unthinkable of any regular physical system that magnitude - a mere cubic micron. The special properties of water, the polymerization strategies, the organization in "architectures", the coded correspondence between triplets and amino acids, the folding process, the catalytic properties of enzymes, the semi-permeable membrane, etc. should be invoked as molecular builders of the cellular order, providers of "processing power" to this basic informational entity.

From our informational perspective, it is interesting that the vocabulary developed by molecular biologists was crafted, almost from the very beginning, around the *communication* metaphor: molecular recognition, genetic and epigenetic codes, transcription, translation, processors, messengers, signalling systems, effectors, transducers, second messengers, regulators, interferences, complexes, networks, modules, etc. It is quite revealing that the founding fathers of molecular biology so heavily relied on the communication metaphor for naming the molecular events they were uncovering. Advertently or inadvertently, they were fundamentally right.

Briefly analyzing the two essential informational traits, *self-production* is basically performed by a "network society" of specialized enzyme and protein agents that are coded onto the DNA "sequential architecture" and are continuously exchanging information about their specific activities thanks to the especial solvent properties of the water matrix. In response to signals of the environment, to be described later, enzymes and proteins are synthesized (and also degraded) out from the sequential information of DNA and RNA, which are themselves incessantly subject to evolutionary combinatory games [Marijuán, 2002, 2010]. The whole productive processes culminate in the regularity of a specific cell-cycle that is open to the environment.

Nevertheless, by itself the transcription network expressing the DNA genes is closed, “blind”. In other words, the coupling between the sequential genetic architecture and the diluted architecture of enzymes and proteins needs the injection of further adaptive capability to respond to environmental demands. This is done by means of signaling guidance, so to partially deploy the genetic circuits in response to relevant happenstances of the environment and also of the cellular interior. Most of the *topological governance* of the transcription regulatory network, the decision of what parts should be activated or what particular circuits should be inhibited, is achieved by means of the cellular signaling system [Navarro, 2010].

In the implementation of *communication*, a variety of signaling systems may be found in prokaryotic cells, ranging from simple transcription-sensory regulators [a single protein comprising two domains], to those systems of multiple components and interconnected pathways that regulate key stages of the cell cycle, such as latency, pathogenesis, replication, and dispersion. A basic taxonomy of bacterial signaling systems was proposed somewhere else [Marijuán *et al*., 2010], which was centered on “the 1-2-3 scheme.” In a specific bacterium, for instance *M. tuberculosis* or *E. coli*, the number of different signaling pathways is close to one hundred, the majority belonging to the “one component systems” class, with around one or two dozen members of the “two component systems” class, and a few members of miscellaneous classes (“three component systems” and others). Every one of these signaling pathways may be present in a range of one or two orders of magnitude—some dozen to some hundred molecules. Conversely, in the eukaryotic signaling system, hundreds of different pathways and thousands of dedicated molecular agents may participate [receptors, ion channels, transducers, amplification cascades, second messengers, intermediate effectors, final effectors]; and they can be arranged differently in each tissue [Marijuán *et al*., 2013]. The flows of information crossing throughout the different receptors and further transduced along the signaling pathways are informing in prokaryotic cells about the presence of metabolic items in the environment, while in eukaryotic cells these very flows mostly relate to direct cell-to-cell communication about developmental and physiological matters. In both cases the information flows become systematically transformed into variations of the self-producing structure.

Let us emphasize the difference between signals and metabolites. Essential molecular components of the environment are continuously scanned by the signaling system: nutrients, ions, metals, peptides, amino acids, toxics, signaling hormones, etc. Once important molecular presences are detected, the system activates gene expression programs or directly induces changes in the cytoplasm and membrane. But signals themselves are left untouched: the molecules participating as “messengers” are merely recognized as signals and do not suffer further transformations as happens with metabolites along metabolic pathways.

Thereafter, *meaning* may be defined throughout *molecular mining*: as the [signal] induced changes in components and connectivity of the constitutive enzyme-protein populations and the associate metabolites and substrates. The *relevance* or *value* of the signal can subsequently be considered and gauged —this in general corresponds to second messengers and the cell cycle “checkpoints”. Completion of the cell cycle always appears as the fundamental reference. The phenomenon of *knowledge* may be appended too, once the generative codes of the protein agents implementing successful responses have been evolutionarily selected, refined, and cohered within the life cycle. The *recombination* strategy performed upon the DNA stretches emerges as an abstract problem-solving tool of far reaching evolutionary consequences, basically understood as “domain recombination” [Marijuán and del Moral, 2007; del Moral *et al*., 2011].

To conclude this brief description of the simplest informational entity [at the time being the only one we can describe molecularly with some reasonable approximation], the methodological problems found are not insurmountable. Like in other more mature disciplines, pragmatic solutions may also be found for the informational analysis of living entities: those safe guidelines under which an efficient reduction of complexity may be achieved.

The generation of new domains of complexity: social knowledge

The informational approach we have drafted for the cell can somehow be generalized upwards. The information flow – the communication sphere – exists around the successive compartments of cells and organisms, accompanied always by a parallel flow of nutrients and maintenance and repair components. Actually every new organization realm develops some specific information flow; at the frontiers or junctures between compartments the previous, more basic information flow is maintained—as happens for instance with “old” molecular diffusion in between “modern” synaptic contacts. This successive composition of information layers seems valid for organisms, individuals and societies too.

Is there any “natural” information flow we can point at in human societies? Anthropologically, the most distinguishing trait of our species is the use of language for communication among individuals. Evolutionarily, these communication activities, basically face-to-face conversation, have taken a genuine evolutionary category along the “social brain hypothesis” recently framed [Dunbar, 2004]. It means that the most plausible evolutionary hypothesis on the hypertrophic growth of human brain relates to the adaptation to bigger social groups and to the use of language as a fundamental vehicle of communication and as new form of social “grooming” [Allman, 1999; Silk, 2007]. Oral exchanges become the way to share group experiences, techniques, knowledge, culture, etc., far beyond any previous “cultural repertoires” of other anthropoids. When the open-ended combinatory structure of language is finally pictured into written “permanent” form [the parallel with the “permanent” DNA of the cell is unavoidable], a whole new world for the circulation of information and knowledge is generated in human societies.

The historical fact is that the development of social complexity has become irreversibly linked to a chain of inventions for communication and knowledge generation. There is a succession of fundamental inventions that dramatically alter the “infostructure” of societies: numbers, writing, alphabet, codices, universities, printing press, books, steam engines, means of communication, computers, etc. [Hobart and Schiffman, 1998]. Another crucial aspect related to human cognition is the *recombination of knowledge*. It is a phenomenon which has passed almost unnoticed in traditional philosophy of science notwithstanding its massive presence in contemporary scientific-technological societies [Scott, 1998; Arthur, 2009]. The term is not exclusively related to the social or to the biomolecular fields (DNA domain recombination). It is quite relevant that contemporary neuroscience has also recognized the importance of the recombinatory dynamics in the “neuronal workspace” of individuals [Dehaene, 2009]. The social creation of knowledge, and obviously the growth of science, derives from knowledge recombination processes taking place in the cerebral workspace of individuals. The strict conditions put by the scientific method would represent the ways and means to directly interconnect standardized individual perceptions and actions beyond time, space, and cultural limitations, allowing the social decomposability of problems and the occurrence of knowledge recombination dynamics at a global and intergenerational scale [Marijuán *et al*., 2012].

A new vantage point about the whole social dynamics of scientific knowledge is needed in today's "information societies"—our sciences have become a Babel Tower of more than 6,000 scientific and technological disciplines, where classical information [library] science is helpless to provide any interesting guidance to society [Marijuán *et al*., 2012]. Information science should promote a new vision to help make sense of the historical expansion of science, and of the requirements of human knowledge in action.

Information science, properly developed and linked with computer science and mathematics, should constitute one of the Great Domains of contemporary science. The informational would go together with the physical, the biological, and the social: constituting the four great domains of science. Rather than attempting the construction of another average or standard discipline, information science is about the making out of one of the "great scientific domains" of contemporary knowledge.

Bibliography

[Allman, 1999] J.M. Allman Evolving Brains. Scientific American Library. New York, NY, USA. 1999.

[Arthur, 2009] B.W. Arthur. The Nature of Technology: What It Is and How it Evolves. New York: The Free Press. 2009.

[Bejan and Peder, 2012] A. Bejan and J. Peder. Design in nature. Doubleday. New York. 2012.

[Conrad, 1996] M. Conrad. Cross-scale information processing in evolution, development and intelligence. BioSystems. 38 97-109. 1996.

[Dehaene, 2009] S. Dehaene. Reading in the brain. Penguin. New York. 2009.

[del Moral et al., 2013] R. del Moral, J. Navarro, P.C. Marijuán. New times and new challenges for information science. Proceedings of the 5th Foundations of Information Science. Information 2014, 5, 101-119. 2013.

[Dunbar, 2004] R. Dunbar. The Human Story: A New History of Mankind's Evolution. Faber & Faber Ltd. London. 2004.

[Hobart and Schiffman, 1998] M.E. Hobart and Z.S. Schiffman. Information Ages. The John Hopkins University. Baltimore, MD, USA. 1998.

[Holland, 2012] J.H. Holland. Signals and Boundaries: Building Blocks for Complex Adaptive Systems. The MIT Press. Cambridge, MA, USA. 2012.

[Lanham, 2006] R.A. Lanham. The Economics of Attention. The University Chicago Press. Chicago, IL, USA. 2006.

[Lenski, 2010] W. Lenski. Information: A Conceptual Investigation. Information. 1 [2] 74-118. 2010.

[Marijuan and del Moral, 2007] P.C. Marijuán and del R. Moral. The informational architectures of biological complexity, In: Computation, Information, Cognition –The Nexus and The Liminal. Dodig-Crnkovic G. and Stuart S. [eds.]. Cambridge University Press, Cambridge. 2007.

[Marijuán et al., 2010] P.C. Marijuán, J. Navarro, R. del Moral. On prokaryotic intelligence: strategies for sensing the environment. BioSystems. 99 94-103. 2010.

[Marijuán et al., 2012] P.C. Marijuán, R. del Moral J. Navarro. Scientomics: An emergent perspective in knowledge organization. Knowledge Organization. 39 [3] 153-164. 2012, [Marijuán et al., 2013] .C. Marijuán, R. del Moral 2012, J. Navarro. On eukaryotic intelligence: Signaling system's guidance in the evolution of multicellular organization. Biosystems. Vol. 114, Issue 1: 8–24. 2013.

[Marijuán, 1996] P.C. Marijuán. First conference on foundations of information science: From computers and quantum physics, to cells, nervous systems, and societies. BioSystems. 38 87-96. 1996.

[Marijuán, 2002] P.C. Marijuán. Bioinformation: untangling the networks of life. BioSystems. 64 11-118. 2002.

[Marijuán, 2010] P.C. Marijuán. Knowledge recombination: on the informational adaptability of cells, nervous systems, and societies. International Journal “Information Theories and Applications. Vol. 8, No. 1, pp. 3-15. 2010.

[Morowitz, 1968] H. J. Morowitz Energy flow in biology: biological organization as a problem in thermal physics. Academic press New York, London 1968.

[Navarro, 2010] J. Navarro. Transcriptional Regulatory Network of M. tuberculosis: Functional and Signaling Aspects. Master Thesis. Universidad de Zaragoza, Zaragoza, Spain. 2010.

[Rosenbloom, 2013] P.S. Rosenbloom. On Computing. The Fourth Great Scientific Domain. The MIT Press. Cambridge [MA]. 2013.

[Scott, 1998] J.C. Scott. Seeing Like a State. Yale University Press. New Haven, CT, USA. 1998.

[Silk, 2007] J.B. Silk. Social Components of Fitness in Primate Groups. Science. 317: 1347-51. 2007.

Authors' Information

Jorge Navarro López *- (jnavarro.iacs@aragon.es), Graduate in Chemical Engineering, University of Zaragoza. Master in Physics and Physical Technologies, University of Zaragoza. In 2008 he joined the Bioinformation Group of the Aragon Health Sciences Institute (IACS) as Research Assistant. His main research interest is on Systems Biology, but also cellular signalling systems, cellular intelligence and the nature of biological information.*

Raquel del Moral Bergós *- (rdelmoral.iacs@aragon.es), Graduate in Biology, Complutense University of Madrid. Master in Molecular and Cellular Biology, University of Zaragoza. In 2008 she joined the Bioinformation Group of the Aragon Health Sciences Institute (IACS) as Research Assistant. Her main research interest is on Neuroscience, but she is also keen on the nature of social information, and the parallel with the biology.*

Pedro C. Marijuán *- (pcmarijuan.iacs@aragon.es), Senior Researcher, is the leader of the Bioinformation & Systems Biology Group at the Aragon Health Sciences Institute (IACS). Engineer and Doctor in Cognitive Neuroscience (PhD Thesis on "Natural Intelligence", University of Barcelona, 1989). During more than 20 years he has advanced research on the nature of biological (cellular) information, communication, knowledge, and intelligence. He has covered both the intracellular molecular mechanisms and the systemic, behavioural (brain), and social realms. He was co-founder with Michael Conrad of FIS (Foundations of Information Science).*

ANALYSIS AND PROCESSING OF THE TEXT INFORMATION AIMED AT EXTRACTING BASIC KNOWLEDGE

Kryvyi Sergii, Bibikov Dmitriy

***Abstract**: The problems extraction knowledge from natural language text is considered. An automatization approach to extraction knowledge is proposed.*

***Keywords**: automatically analysis of natural languages text, extraction of knowledge.*

***ACM Classification Keywords**: I.2 ARTIFICIAL INTELLIGENCE – I.2.4 Knowledge Representation Formalisms and Methods.*

Introduction

The rapid development of science and technology during the last decades of the 20th century and early 21st century led to enormous information growth that one person (even highly skilled in science and technology) is unable to learn, understand and use to conduct research. Due to such a situation there is a need to automate the search and processing the necessary information for its subsequent efficient usage. To do this a few problems should be solved [1].

The first and one of the main problems is the analysis of natural language text information (morphological, syntactic, semantic and logical analysis) to extract knowledge.

The second problem is the issue of search engine design and extraction of knowledge, construction of its architecture and development of tools to help the user.

Third issue - is the integration of knowledge from multiple (or two) domains to ensure the effectiveness of interdisciplinary research, using existing algorithms, facts, theoretical principles and practical solutions.

The third problem is closely linked with such problems as the effective usage of automatic search of theorem proof in formal logic and problems similar to it. Application of the prover can only be successful when all the necessary information will be at its disposal. For example, if the prover needs to prove the Lagrange theorem which concerns the divisibility of the order of a finite group by the index of its subgroups, it’s not enough for the prover to have axiomatically group theory. It also needs the axiomatic of divisibility theory, and perhaps axiomatic of Peano along with some additional facts from other areas. The solution to this problem is easier to find if prover has an integrated system that includes relevant information from other areas of knowledge. Then the prover finds the necessary information and uses it to conduct successful proof.

Certain process formalization is discussed in this paper, like the analysis of natural language texts (NLT), getting primary information from the NLT, finding logical conclusions from these facts and making sure they are correct.

Short overview of research methods of natural language texts

Formalization of natural language to automate the analysis of natural language texts was initiated in the early 30s of the 20th century by A. Tarski and his students, although this need was expressed long before - by Aristotle, Leibniz and Euler. In particular, Aristotle has provided four types of statements:

A - "all X is Y"; E - "no X is not Y"; I - "is Y some X", O - "some X is not Y".

These types of statements were called syllogisms, and the approach of Aristotle - syllogistics. Later Euler outlined his understanding of Aristotle syllogistics using the geometric interpretation of syllogistics as circles (this interpretation was called Euler's circles). Euler's ideas were developed further in the works of French mathematician and astronomer J.D. Zherhon who introduced types of relations and syllogistic interpretation of Aristotle in terms of these relations. The main types of relations imposed by Zherhon are: G1 – "the same or equivalent"; G2 – "left-side inclusion"; G3 – "special case of shrinkage"; G4 – "right-side inclusion"; G5 – "incompatibility" [9].

Zherhon showed that each type of a syllogism of Aristotle can be expressed as a set of some possible options of such relations. In particular A: (G1, G2), E: (G5), I: (G1, G2, G3, G4), O: (G3, G4, G5). For example, the statement I means that some non empty subsets of the set or class X is included in Y. The main difficulty in using Zherhon's relations is that almost all types of relations in a complex sentence require a large number of test options. Therefore, a more appropriate approach was based on the usage of formal mathematical logic.

Much more serious attempts were made by A. Tarski [5, 6], which led to the emergence of the notion of satisfaction of formulas – a more general concept than the notion of truth. This notion Tarski applied to open and closed formulas (in sentences of natural language a closed formula means a phrase), and this helped to formulate the concept of truth of natural language sentences and impose it on every open atomic formula that consists of a primitive predicate (subject constant) and as many variables, as correspond to the predicate arity. As the set of such formulas is finite, this approach is constructive.

The next attempt to improve formalization of A. Tarski was made by D. Davidson [7]. He proposed to add the recursive definition of truth to the concepts of truth and satisfiability. Thus the T theory, which includes a recursive definition of truth, does not explain how the meaning of phrases depends on the meaning of words of these phrases. But since the word "value" is not synonymous with "truth", the definition of truth is not always the definition (meaning). Thus, Davidson's thesis is not quite obvious, but it's easy to justify. Hence, it's necessary to compare the meaning of the narrative sentences with the terms of their correctness.

If we accept the equation "value of a narrative sentence = the condition of truth of this sentence", a condition must be set: if the definition describes how conditions of truth of a compound sentence depend on the conditions of validity of its constituent simple sentences, then the definition should describe how the value of a complex sentence depends on the values of its constituent simple sentences.

Montague also believed that the methods of formal semantics can be applied to the study of natural language semantics. But, unlike Davidson, he rejected the application of first order predicate logic, opting for categorical grammars. These grammars include those categories that specialists in traditional grammars use in definitions of natural language, e.g. such categories as subject or predicate. This is an opposite approach to the Davidson's. This made it possible to replace the notion of Montague's absolute truth with the notion of relative truth in the model, because in one model one sentence can be true, and in the other - false. This expansion helped to define the notion of logical truth and logical consequence of a larger fragment of natural language [8]. So Montague highlighted two elements: intention (meaning) and extention (denotation) and applied them to subjects, predicates and phrases. There are other approaches to the analysis of natural language texts based on the notions of semantic networks, frames, etc.

If we briefly characterize the ideas that dominated the last decade of the 20th century, they can be reduced to the following.

In the beginning of the 1970s the studies of natural and artificial (formal) languages certain attempts to construct a theory dominated. This theory would consider both natural and artificial languages. Syntax is only considered in terms of semantics. The goal of semantics is to explain concepts of truth and logical imitation. The purpose of syntax is to characterize syntactic categories that form expressions.

This paper considers an approach which combines aspects of algebraic analysis of natural language and also the logical aspects of such analysis. The following text is structured as follows.

Formal statement of the problem of knowledge extraction from NLT is formulated. From this formulation follows concretization of texts of certain restrictions. As an examples of this type of constraints the syllogistics of Aristotle. In particular, for the syllogistics of Aristotle we suggest to build a set-theoretic interpretation of the extended system of rules of inference and analysis of possible situations that may arise in the application of this system of rules.

Later texts of definitions are considered that concern relations of subordination.

A formal statement of the problem of knowledge extraction from NLT

Before we beguine the review of the system of processing and extraction of knowledge contained in the NLT, we shall define the notion of knowledge and knowledge extraction from NLT. For this purpose, we use the concepts used in programming with constraints [2].

Let this set D, which is identified in a finite set of n-arity relation R on D, so $R_i \subseteq D^n$, where $R_i \in R \subseteq D$, i = 1, ..., k. The language restrictions L on D we call some non empty set $L \subseteq R \subseteq D$. The problem of satisfiability of constraints is formulated as follows.

Definition 1. For any set D and any constraint language L over D the constraint satisfaction problem (CSP (L)) is the solution of this combinatorial problem:

Instance: A triple P = (V, D, C), where V is a finite set of variables; C is a set of constraints ($C_1, ..., C_q$); each constraint $C_i \in C$ is a pair (s_i, R_i), where s_i is a n-element sequence of V, which is called the domain restrictions, $R_i \in L$ is a n-ary relation over D, called the constraint relation.

Question: Does there exists a function $\varphi : V \to D$ such that for each constraint $(s, R) \in C$, whith $s = (v_1, v_2, ..., v_n)$, the tuple $(\varphi(v_1), \varphi(v_2), ..., \varphi(v_n) \in R$?

Set D in this case is called the domain of the problem, the set of all solutions CSP when P = (V, D, C) is indicated by Sol (P).

In case with NLT analysis, in order to extract knowledge a set D, field of the problem, is interpreted as a set of objects extracted from the input text T, which is factorized according to some equivalence relation R (we call this relation synonymous relation), which has "coded" relations $R_i, i = 1, 2, \ldots, k$. Variables of the set of variables $V = \{v_1, v_2, \ldots, v_m\}$ take their values in this factorized set of objects that appear in the text T (these can be lexico-grammatical levels, specific objects like people, dates, objects, etc.)

The problem of knowledge extraction from NLT is a problem of search of an interpretation $\varphi : V \to D$, while building obvious R_i relations from a set of $L \subseteq R$. The ratio $R_i \in L,\ i = 1, 2, ..., k,$ extracted from the text T, we shall call knowledge. This interpretation we are building in iterative manner.

While analyzing the NLT, our primary task is to build two fundamental relations which are present in virtually every NLT. This equivalence relation and partial order are known as generally valid. The first of these relations has been already discussed and it defines classes of synonymous objects, the second relation explains the hierarchy of equivalence classes. Both of these form the basis for building ontology, while the knowledge gained at this stage - will be called basic. With respect to partial order a different semantic meaning can be included: it may be relevant taxonomy ("belong" to the set, class, group, etc.), attitude of patrimony (“consist of”), related genealogy (“father-son”), cause-related relations ("if - then"), an attribute relation, etc.

The above definition of knowledge extraction from NLT is quite general and needs refinement. Let’s consider some of the concrete definitions. Refinement can be performed in different directions depending on subject area and purpose pursued by analysis of NLT. We shall illustrate them with examples:

1) Aristotle’s syllogism and its set-theoretic interpretation. If we analyze the types of generally meaningful relations, one can notice that they are associated primarily with the relation of partial order. This type of relation has the distributive grating which has useful properties and these properties can be used to generate effects, e.g. generate new knowledge. Moreover, it is easy to notice that the Aristotle’s syllogisms are interpreted in the algebra of sets and relations with such dependencies [3, 10]:

$$A(X,Y) \Leftrightarrow X \subseteq Y,\ E(X,Y) \Leftrightarrow X \cap Y = \varnothing, I(X,Y) \Leftrightarrow X \cap Y \neq \varnothing,\ O(X,Y) \Leftrightarrow X \setminus Y \neq \varnothing.$$

And from the algebra of sets and relations it is known that operations of union, intersection and complement, this algebra is a Boolean ring, carriers of which are partially ordered sets against set-theoretic inclusion. Laws of algebra and properties of partially ordered sets can be used as inference rules in such a formal system. More details of these opportunities are considered in [3]. In particular, these are the laws of algebra of sets and relations (commutativity, associativity, distributivity, idempotensy, acquisitions and De Morgan's laws), three basic properties of the inclusion relation (transitivity, and antisymmetry contraposition) and the law of double complementation. Thus the first two properties act as inference rules. We shall illustrate them with examples.

Example. The following facts are set (taken from the book by Carol L. "The History of nodules”): "All the little children are foolish"; "Anyone who can tame a snake deserves respect"; "All the stupid people do not deserve respect".

Let us find out what consequences follow from these facts. Note that this type of facts logic sometimes calls polysyllogisms or sorites. And syllogism is called a system which has only two replaces.

Now we define key terms that make up the system of facts, denote them and select the universe U. In this example, basic terms are: "little children" (C), "smart people" (P), "those who tame snakes" (T), “those who deserves respect "(Π). Clearly, these terms represent some sets in the universe "people". Their negations will be under the following terms: "grownups" ($\neg$C), "foolish people" ($\neg$P), "those who cannot tame snakes" ($\neg$T), "those who do not deserve respect" ($\neg$Π). Now take our facts look as follows: C $\subseteq$ $\neg$P, T $\subseteq$Π, $\neg$P $\subseteq$ $\neg$Π.

Thus, the grating is identified (as a universal set), which consists of elements ($\varnothing$, U, P, T, Π, $\neg$C, $\neg$P, $\neg$T, $\neg$Π), where U is the universe. Thus, the first effects of these facts are the following based on the rule of contraposition (rule of contraposition in this interpretation of the form "A $\subseteq$B follows from $\neg$ B $\subseteq$ $\neg$ A, where the sign $\neg$ means set complement):

$$(C^1): P\subseteq\neg C, \quad (C^2): \neg П \subseteq \neg T, \quad (C^3): П \subseteq P.$$

If we transfer the obtained effects in natural language, they respectively mean the following facts: "All smart people are not little children", "those who do not deserve respect, do not tame snakes", "he, who is a clever person, deserves respect". Using transitivity rule, we get the following consequences:

$$(C^4): C \subseteq \neg П, \quad (C^5): T \subseteq P, \quad (C^6): \neg P\subseteq \neg T, \quad (C^7): П\subseteq \neg C.$$

From these effects by the same rule of transitivity we get two more results:

$$(C^8): C \subseteq \neg T, \quad (C^9): T \subseteq \neg C.$$

If we translate the last consequences into natural language, they will sound like this: "All small children cannot tame snakes", "all who tame snakes are not small children».

From this example it follows that the problem of finding the effects is reduced to a problem of constructing a contra positive, transitive and ant-symmetrical closure (CTA circuit (closure)) of a basic set of formulas. When a closure is transitive, such situations may arise:

K1) the following formula is obtained $A \to \neg A$ or $\neg A \to A$;

K2) in the process of building a transitive closure, at least one cycle is obtained.

We will now consider what these situations in our case mean. The first formula in the case of K1) corresponds to the situation $A' \subseteq A$. From the properties of intersection and its complement we have $A \cap A' = \varnothing$, so the inclusion is true only if A is an empty set. The second formula in the case of K1) means the complement A' for set A should be a universal set. In terms of algebra of sets such situations cannot be described as contentious, but in our case, this situation means that some object must exist and at the same time does not exist. This situation is interesting for at least two reasons. The first reason is that the situation is catastrophic, and the second reason makes it possible to exclude certain terms that have led to controversy from consideration. We analyze these cases in detail. The first cause of such contention $A \to \neg A$ is the appearance of formulas like $A \to B$ and $A \to \neg B$ in the set of consequences. Another cause of controversy $A \to \neg A$ is the emergence of two formulas $A \to B$ and $\neg A \to B$.

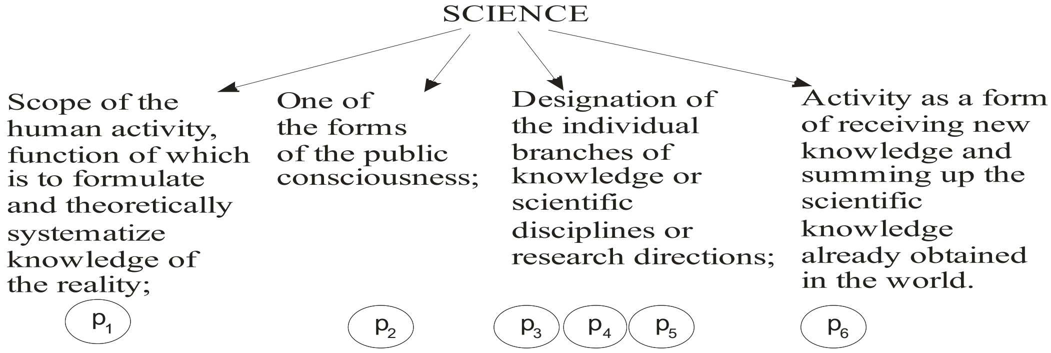
Note that the existence of such contradictions $A \to \neg A$ do not always lead to catastrophic consequences. Sometimes the appearance of such conflicts makes it possible to recognize the facts that lead to conflicts and delete conflicts.

The situations for K1) and K2) another extremely important point follows. Contradictions such as K1), K2) are formal because they appear only as a result of logical analysis given set of facts, but there is still a kind of contradiction, which differs significantly from the contradictions K1) and K2). Suppose that as a result of construction of ACT-circuit with good sound and proven facts that are not contentious, such as known and grounded theories, received conflicting effects, i.e. effects that are contrary to the facts of the original theories. It is said that controversy exists between the initial theories, and this is a sign of the emergence of new knowledge or at least the impetus for the analysis and search for causes of the appearing controversies. All of these gives us the ground to introduce such a definition.

Definition 2. Information system is called correct if contradictions like K1) or K2) don't appear.

It is known that in the process of building ACT-closure, according to different sets of initial facts, you can get one and the same set of consequences. This allows us to write the following equivalence relation on the sets of facts: two sets of facts F and F 'are called equivalent if the ACT (F) = ACT (F'). Based on this relation the structure of information system and initial set of facts can be simplified (by elimination).

2) Processing definitions. We consider the text that can be called conventionally structured. These are natural language definitions. Below are examples of such definitions, drawn from the dictionary.



SCIENCE

Scope of the human activity, function of which is to formulate and theoretically systematize knowledge of the reality;

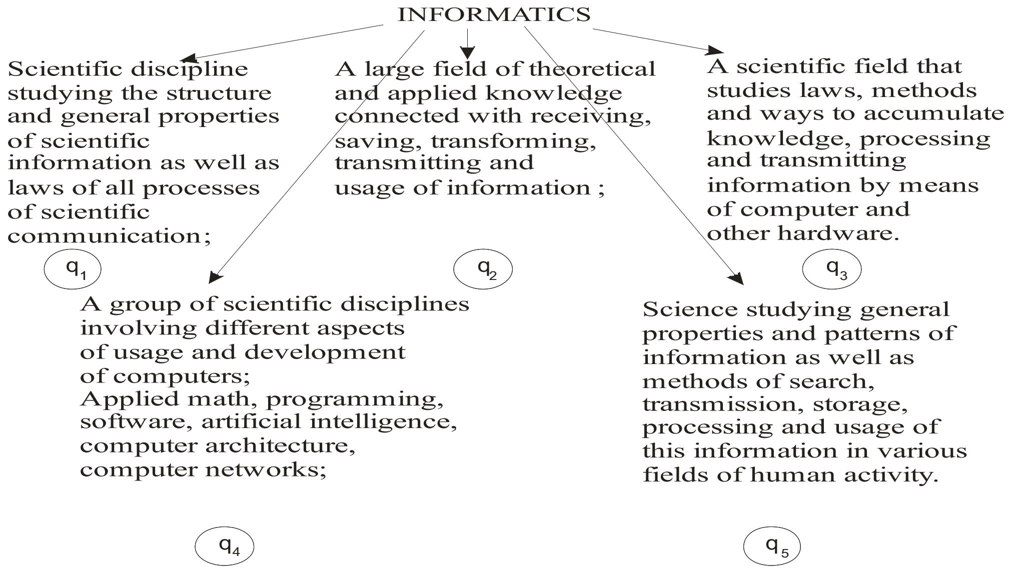
One of the forms of the public consciousness;

Designation of the individual branches of knowledge or scientific disciplines or research directions;

Activity as a form of receiving new knowledge and summing up the scientific knowledge already obtained in the world.

$K(a)=\{p_1,p_2,p_3,p_4,p_5,p_6\}$ - equivalence class

The following definition is also taken from the dictionary.



INFORMATICS

Scientific discipline studying the structure and general properties of scientific information as well as laws of all processes of scientific communication;

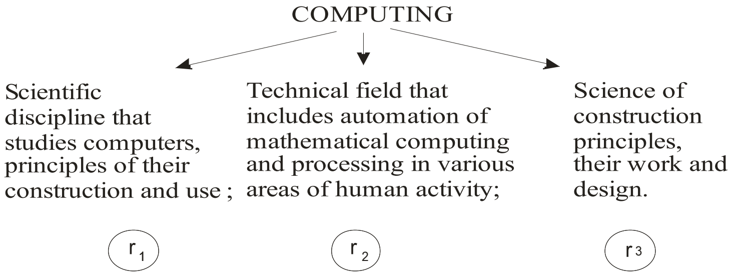
A large field of theoretical and applied knowledge connected with receiving, saving, transforming, transmitting and usage of information ;

A scientific field that studies laws, methods and ways to accumulate knowledge, processing and transmitting information by means of computer and other hardware.

A group of scientific disciplines involving different aspects of usage and development of computers;
Applied math, programming, software, artificial intelligence, computer architecture, computer networks;

Science studying general properties and patterns of information as well as methods of search, transmission, storage, processing and usage of this information in various fields of human activity.

$K(b)=\{q_1,q_2,q_3,q_4,q_5\}$



COMPUTING

Scientific discipline that studies computers, principles of their construction and use ;

Technical field that includes automation of mathematical computing and processing in various areas of human activity;

Science of construction principles, their work and design.

$K(c)=\{r_1,r_2,r_3\}$

These examples show that the construction of equivalence classes is not difficult. As a result of construction of equivalence classes of objects such results appear:

$$K(a) = \{p_1, p_2, p_3, p_4, p_5, p_6\}, \; K(b) = \{q_1, q_2, q_3, q_4, q_5\}, \; K(c) = \{r_1, r_2, r_3\}.$$

The problem appears when computing the second relation, that determines the ratio of subordination (hierarchy) between the equivalence classes found. But such an abstract representation of classes $K_i(x)$ of this relation cannot be determined (not enough information). It is necessary to know the structural characteristics of the elements of the classes $K_i(x)$. So, naturally the necessity to structure elements of equivalence classes appears. For example, if you return to the previous example, every element of the class $K(a)$ takes form:

$$p_1 = (p_{11}, p_{12}, p_{13}), \; p_2 = (p_{21}), \; p_3 = (p_{31}, p_{32}, p_{33}), \; p_4 = (p_{41}, p_{42}), \; p_5 = (p_{51}, p_{52}),$$

$$p_6 = (p_{61}, p_{62}),$$

where p_{11} = "sphere of human activity", p_{12} = "development of knowledge of the objective reality,"

p_{13} = "system of knowledge of the objective reality", p_{21} = "form of social consciousness",

p_{31} = "industry knowledge", p_{32} = "discipline", p_{33} = "scientific direction"

p_{41} = "activities to obtain new knowledge", p_{42} = "summation of knowledge of the SMW".

Similarly, other elements in the equivalence classes are structured.

$$r_1 = (r_{11}, r_{12}, r_{13}, \ldots), \; r_2 = (r_{21}, r_{22}, r_{23}, \ldots), \; r_3 = (r_{31}, r_{32}, r_{33}, \ldots),$$

$$q_1 = (q_{11}, q_{12}, q_{13}, \ldots), \; q_2 = (q_{21}, q_{22}, q_{23}, \ldots), \; q_3 = (q_{31}, q_{32}, q_{33}, \ldots), \; q_4 = (q_{41}, q_{42}, q_{43}, \ldots)$$

$$q_5 = (q_{51}, q_{52}, q_{53}, \ldots)$$

From the structure follows this formulation. If the equivalence class belongs to the object a, it looks like a formal definition of disjunction of elements that make up this class. Each item that is part of a class equivalence is described by the corresponding equivalence predicate, so if $K(a) = \{p_1, p_2, p_3, p_4, p_5, p_6\}$, then $p(a) \Leftrightarrow p_1(a) \vee \ldots \vee p_6(a)$, where p_i are predicates characterizing element of the class $K(a)$, and their disjunction characterizes a class concept a.

Further, if $q_i \in K(a)$ and $q_i = (q_{i1}, q_{i2}, \cdots, q_{ik})$, then element p_i (or object p_i), characterized by attributes p_{ij}, are represented as conjunction $p_i(a) \Leftrightarrow p_{i1}(a) \wedge \ldots \wedge p_{ik}(a)$, where $p_{ij}(a)$ - is a predicate that characterizes the concept of separate attribute $a, \; i = 1, \ldots, l; j = 1, \ldots, k.$

Thus, each class $K(a)$ is described by disjunctive form, like:

$$p(a) \Leftrightarrow (p_{11}(a) \wedge \ldots \wedge p_{1m_1}(a)) \vee \ldots \vee (p_{l1}(a) \wedge \ldots \wedge p_{lm_l}(a)).$$

Noted formalization defines a partial order relation, which is found the following way:

$$K(a) \leq K(b) \Leftrightarrow (\exists p_i(a))(\exists q_j(b))(q_j(b) \leq p_i(a)),$$

where $q_j(b) \le p_i(a)$ means that $q_j(b)$ included as a member of the conjunctive in $p_i(a)$.

So related partial order naturally requires a predicate-relational representation of the objects of equivalence classes and most of these classes [4].To illustrate the information, we go back to the example above. Let's consider, how the fact that class of "SCIENCE" subordinates the class of "INFORMATICS".

Class "SCIENCE", K(a), is described by a formula: $p(a) \Leftrightarrow p_1(a) \vee p_2(a) \vee \ldots \vee p_6(a)$, where $p_1(a) \Leftrightarrow$ SCOPE-HUMAN-ACTIVITY(a),

$p_2(a) \Leftrightarrow$ FORMS-PUBLIC-CONSCIOUSNESS(a),

$p_3(a) \Leftrightarrow$ SCIENCE-DISCIPLINE(a), $p_4(a) \Leftrightarrow$ SCIENCE-DIRECTION(a),

$p_5(a) \Leftrightarrow$ ACTIVITY-RECEIVING-KNOWLEDGE(a),

$p_6(a) \Leftrightarrow$ ACTIVITY-SUMMING-SCIENTIFIC-KNOWLEDGE(a),

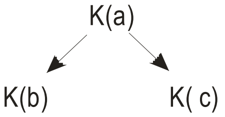
Class "INFORMATICS", noted as $K(b)$, is described by a formula:

$q(b) \Leftrightarrow q_1(b) \vee q_2(b) \vee q_3(b) \vee \ldots \vee q_l(b)$,

where $q_1(b) \Leftrightarrow p_3(a) \wedge$ GENERAL-PROPERTIES-SCIENCE-INFORM(b) $\wedge$ ...

Using the above definition of partial order relations, we find that $p_3(a)$ belongs to a class $K(a)$ definition, and $q_1(b) \vee p_3(b) = p_3(b)$ (based on the law of absorption), which means that $K(b) \le K(a)$.

Similarly we define the subordination of class $K(c)$ to the class $K(a)$, resulting in a graph:



So, the received hierarchy can be changed or modified through the dialogue with the user in order to achieve a more correct representation.

To sum up,, we formalize what was discussed in the above article about the processing of text definitions in the following way.

Suppose T is the set of texts definitions. Under this set of T a set of classes is constructed, determined by the equivalence relation R and is factor set D. Obtained in this manner set D of special relations are defined $R_2, \ldots, R_k$, which describe the characteristic properties of elements D, i.e. elements of equivalence classes. These relations, which we will call primary knowledge are presented in the form of predicates that define the partial order relation R_1. This relation is the second relation on which ontology is based. More precisely, ontology is built on transitive closure R_1^* by relation R_1, which is correlated with the relation R.

Definition 3. Relations R and R_1^* we will call consistent if $\forall a,b \in D$ there is inclusion $(a,b) \in R * R_1^*$, where $R * R_1^*$ - and the superposition relations R and R_1.

From this definition the primary ontology follows logically: $O = (D = T / R, \Re = \{R_1, R_2, \dots, R_k\}, \varphi, A)$ where $\varphi : D \to T$ - interpretation - A set of axioms, defined by predicates that describe the characteristic properties of elements from D, where $R_2, \dots R_k$ - corresponding relations, and R_1 partial order relation.

Conclusion

The proposed ways of automatic processing NLT are the basis of both theoretical and practical process of the analysis of the process of extraction of basic knowledge from NLT and their representation in the form of ontology. Using this framework, and especially its implementation, we will increase its capacity by building new meta relations over the built relations which are separate parts of basic knowledge, that are in the given NLT.

As a basis for building system analysis NLT, algebraic system of listed entities is used with the prospect of further hardware implementation. And the construction of a relevant ontology is performed on the basis of available tools and systems of construction of ontologies.

Bibliography

1. Palagin A.V., Kryvyi S. L., Petrenko N.G. Knowledge-oriented information systems with natural languages processing: foundation of methodologies and architecture-structured organization. – journ. USiM. – 2009. - №3. – C.42-55 (in russian)

2. Cohen D. Jeavons P. The Complexity of Constraint Languages. In “Handbook of Constraint Programming - Edited by F. Rossi, P. van Beek and T. Walsh. -2006. – P. 245 - 280.

3. Kulik B.A. Lodic of natural reasoning - S.-Petersburg: Newskij dialect.- 2001.- 127 c. (in russian)

4. Rubashkin V.S. Representation and analysis of sense in informational systems. – M.: Nauka.- 1989.-188 c. (in russian)

5. Tarski A. The semantic conception of truth. Philosophy and phenomenological Research. – v.4. – 1944. – P. 241-375.

6. *Tarski A.* Logique, Semantique and Metamathematique (1923-1944). Colin. – Paris. – 1972.

7. *Devidson D*. Proceedings of Philosofical Logic. – Reidel. – Dordrecht. – 1969.

8. *Montague R*. Universal grammars. Theoria. Formal Phylisophy: Selected Papers of R. Montague. – Yale University Press. -1974. - vol. 36. – PP222-246.

9. *Thause A, Gribomont P., Hulin G. and other.* Logical approach to artificial intellect. From modal logic to logic of data bases. - M.: Mir.-1998. – 494 c. (translated in russian)

10. Kolmogorov A.N., Dragalin A.G. Introduction to mathematical logic. M.: Publ. MGU,1982, 118 c. (in russian)

Author information

***Kryvyi Sergii** – professor of Kiev national university; Ukraina, Kiev; str. Vladimirskaja, 40;e-mail: krivoi@i.com.ua*

Area of scientific activity: Discrete mathematics, analysis, verification and program development

***Bibikov Dmitriy** – post graduate student of Glushkov’s institute of cybernetics of NASU; e-mail: bb_coff@mail.ru*

Area of scientific activity: Artificial intelligence, automatization of analysis NLT

AUTOMATED TRANSLATION FROM INFLECTIONAL LANGUAGES TO SIGN LANGUAGES

Olexandr Barmak, Iurii Krak, Sergii Romanyshyn

***Abstract**: The article describes the algorithmic implementation of information technology for translation from inflectional languages to sign language. For example info logical model of Ukrainian dictionary and sign language, related generalized grammatical constructions for automatic translation are built. The experimental results to verify the effectiveness of the proposed information technology are represented.*

***Keywords**: Automated translation, sign language, inflectional language.*

Introduction and problem statement

The current state of the international community produces certain attitudes towards people with disabilities. The problem of increasing the degree of people with impaired hearing disabilities participation in society is of great importance. The main obstacle to resolving this problem is the difficulty in communication between people with impaired hearing and hearing people. One of the ways to solve this problem is to create an advanced information technologies for non-verbal communication. These areas of research are involved in many of the leading organizations in the world: the Zardoz system [Veale, 1998], ViSiCAST project [Marshall, 2003], system developed at Dublin City University [Morrissey, 2010], informational technology for non-verbal communication for people with impaired hearing [Krak, 2008].

Sign language is a natural language that conveys information through movements of hands and fingers, facial expressions, position of the body. It is used as part of communication for people and serves as the primary means of communication for people with impaired hearing. Sign languages are not the visual interpretation of ordinary language. They have their own grammar that can be used to discuss a variety of topics from simple and specific to the sublime and abstract. The lexicology, phraseology, morphology of sign language is still poorly understood.

The aim of this work is to develop informational technology of translation verbal inflectional languages to a natural sign language. When the data is translated from one language to another pair of grammatical constructions is obtained. These grammatical constructions convey meaning: sentences in the input language → appropriate sentence in the source language. We suppose that a pair of data structures can be represented as a certain generalization. By analyzing certain amount of pairs obtained by translation the word order in sentences can be fixed, and allow us to build a generalization, in which instead of specific words in the sentence we will get the set of words that can be used in these fixed locations. The amount of grammar constructions obtained this way is fairly small (relative to the total number of sentences).

We offer a solution to this problem with the following restrictions:

1) The system works only with simple sentences limited to a fixed list of topics and situations;

2) The system translates (with no subsequent corrections) only trained structures of sentences, without distorting the meaning; translation prediction is possible for other structures of sentences;

3) Permanent expansion of a list of translation structures is possible.

Models of dictionaries for automated translation

Automated translation from inflectional languages to sign language involves the creation of plural model for the dictionary of the language [Shirokov, 1998] for modeling of generalized grammatical constructions of inflectional and sign languages.

In inflectional languages grammatical meaning are transmitted through flexions. The words of inflectional language are modeled as a combination of two components: constant component (base) and a variable component (flexion).

$$x = c(x)\,\&\,f(x), \tag{1}$$

where $c(x)$ – constant part of lexeme x, $f(x)$ – variable part of lexeme x, & – concatenation.

Inflectional languages have a formal model of inflection. An inflection expresses one or more grammatical categories with a prefix, suffix or infix, or another internal modification such as a vowel change. Inflexion expresses different grammatical categories such as tense, mood, voice, aspect, person, number, gender and case. Considering these features of inflectional languages words are modeled as:

$$W = \{W_i : W_i = \{I_{i_1} \in I, F_{i_2} \in F,\ k,\ In_{i_3} \in In\}\} \tag{2}$$

where W_i are parameters of inflectional language ($i = 0,\cdots,N-1$, N – amount of words in dictionary); F – set of all possible flexes; k – position in base word form, from which it is concatenated with flexion ($k = 0$ – if inflected word if completely different from base word form); In – set of inflectional parameters of a language (tense, mood, voice, aspect, person, number, gender and case), I - a set of base word forms:

$$I = \{I_i : I_i = \{word\,\mathrm{inf}, p \in P\}\} \tag{3}$$

where P - a set of parts of speech; $word\,\mathrm{inf}$ - base word form.

Sign language dictionary structure is simpler due to the absence of inflection in it. In non-verbal communication mimic component plays a very important and sometimes crucial role. It should be noted that the syntax of sign language is characterized by non-manual marker: the questioning sentences with general questions – raising eyebrows; in a separate (private) question – omission of eyebrows and head tilted forward; in denial – negative head movements, the corresponding expression; in narrative sentences it is characterized by facial emotional color that corresponds to the meaning of the information transmitted.

The set of gestures is modeled as:

$$Ges = \{Ges_i : Ges_i = \{word\,des, pges \in PGes, em \in Em\}\} \tag{4}$$

where $word\,des$ – description of gesture, $PGes$ – corresponding part of speech, $Em \in \{em_1, em_2, em_3\}$ – a set of emotional coloring: em_1 .– narrative emotion, em_2 .– questioning emotion, em_3 – other emotions.

Models of grammatical constructions for automated translation

Will consider the syntactic features of sign language using three typical structures of sentences: subject-object-verb, subject-verb-object, verb-subject-object. Subject and predicate in these proposals are related by predicative bond. We define simple sentences as sentences that have one predicative bond. Note that the order of words in sentences with one predicative connection in most spoken languages of the world is described by one of three types of structures [Tomlin, 1986]. In sign language simple sentences serve as the primary means of communication and are divided into declarative, interrogative and incentive.

Sentences								
1	2	…	n	→	k1	k2	…	km
Word	Word		Word		Gesture	Gesture		Gesture

Grammatical constructions								
1	2	…	n	→	k1	k2	…	km
{Set of words}	{Set of words}		{Set of words}		{Set of gestures}	{Set of gestures}		{Set of gestures}

Generalized grammatical constructions								
1	2	…	n	→	k1	k2	…	km
([±]PP, [±]II), [±]V	(…)		(…)		([±]PP)	(…)		(…)

Figure 1. Building generalized grammatical constructions

After building the inflectional and sign language dictionaries we need to build a model of grammatical constructions of sentences. By sentence will understand the sequence of words and punctuation (fig. 1). We consider working only with simple sentences (any complex sentence can be represented as a composition of simple sentences). By grammatical sentence construction we understand the sequence of words of language relating to parts of speech that convey meaning (we discard function words as they don't convey meaning). We distinguish grammatical constructions by the number of words they contain. Grammatical constructions (as opposed to sentences) contain the set of words that meet in certain position of sentences:

$$\begin{gathered} Gr = \{Gr_i = \{word_j \in W \| seq_j \in Seq \| p \in P,\ In_{i_1} \in In, num\}, \\ GStr_i = \{ges_j \in Ges \| gseq \in Gseq \|\ pges \in PGes, num, gesnum\}\} \end{gathered} \tag{5}$$

where *num* – number of element in structure of inflectional language, *gesnum* – number of element in structure of sign language, In - set of inflectional parameters of a language, *P* – a set of parts of speech of inflectional language, *PGes* – a set of parts of speech of sign language, *W* – a set words of inflectional language, *Ges* – a set words of gestures,

$$Seq = \{Seq_i : Seq_i = \{word_j \in W : word_j \in W, n\}\},$$

where *n* – number of word in set of words,

$$Gseq = \{Gseq_i : Gseq_i = \{ges_j \in Ges : ges_j \in Ges, n\}\},$$

where *n* – number of gesture in set of gestures.

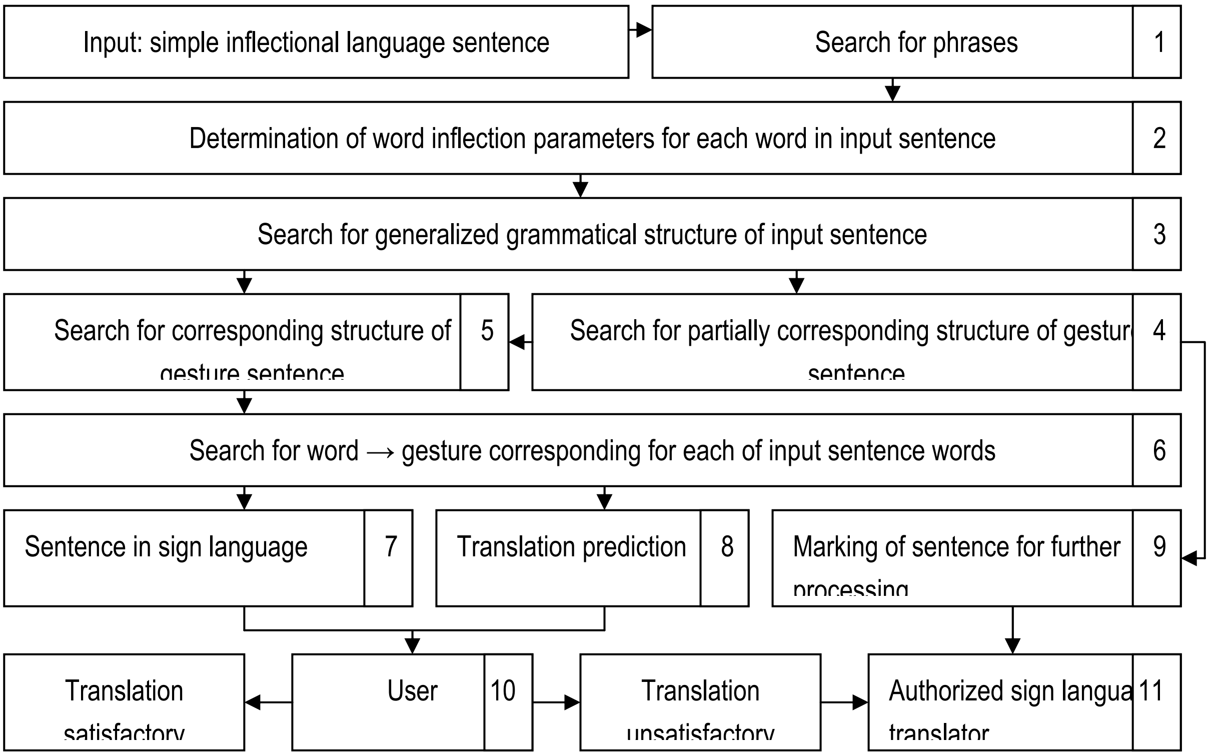
After receiving a set of grammatical constructions we analyze each of the sets of elements that go into it. By generalized grammatical constructions we understand a construction that contains a combination of inflectional parameters of a language instead of set words:

$$\begin{gathered} GGr = \{GGr_i = \{p_j \in P, In_j \in In, num\}, \\ GStr_i = \{pges \in PGes, num, gesnum\}\} \end{gathered} \qquad (6)$$

Generalized grammatical constructions can contain parameters which may or may not be used at certain sentence position, such as part of speech $PP = \{PP_i : PP_i = \{p_i \in P\}\}$ and inflectional parameters of word (depending on language specification it can contain tense, mood, voice, aspect, person, number, gender and case) $II = \{II_j : II_j = \{In_j \in In\}\}$. In addition to inflectional parameters generalized grammatical constructions can also contain set of certain words $V = \{V_i : V_i = \{w_i \in W\}\}$.

Translation algorithm

The following algorithm (fig. 2) for automatic translation from the inflectional language to sign language is proposed:



Input: simple inflectional language sentence
Search for phrases
1
Determination of word inflection parameters for each word in input sentence
2
Search for generalized grammatical structure of input sentence
3
Search for corresponding structure of gesture sentence
5
Search for partially corresponding structure of gestur sentence
4
Search for word → gesture corresponding for each of input sentence words
6
Sentence in sign language
7
Translation prediction
8
Marking of sentence for further processing
9
Translation satisfactory
User
10
Translation unsatisfactory
Authorized sign langua translator
11

Figure 2. Automated translation algorithm

1. Input sentences are checked for existence of phrases.

2. Using inflection language vocabulary inflection parameters for each word are determined.

3. Based on inflection parameters the generalized grammatical construction is determined.

4. If generalized grammatical construction if not found the search of partially corresponding constructions is performed.

5. If generalized grammatical construction of input sentence is found we determine corresponding generalized grammatical construction of gesture sentence and gestures corresponding to input words.

7. If the corresponding generalized grammatical construction and corresponding gestures are found.

8. In case if partially corresponding grammatical construction an attempt to predict translation is made.

9. If no partially corresponding grammatical construction is found the sentence is marked for further processing by authorized sign language expert. The expert can create new grammatical constructions or corresponding between words and gestures.

10. After receiving the translation user chooses one of the following:

- The translation is satisfactory and is stored to statistical database as correct;
- The translation is unsatisfactory and is stored to statistical for further processing. The result of translation can be manually edited.

Creation of the automated translation system

To implement the proposed approach, on an example of the Ukrainian language, a set of gestures and a set of sentences used in everyday communication were formed based on the educational program of the "Ukrainian Sign Language" which is used in schools for the people with impaired hearing to master sign language. Sentences were translated to sign language with the help of sign language experts. Each of the sentences was put in correspondence sign language sentence. The sentences in the data model were merged into the grammatical constructions obtained by generalization according to the formula (5). For example sentences “He walks”, “She walks”, “Man walks”, etc were merged into grammatical construction “{He, She, Man…} walks”. By analyzing the sets of words on their inflectional parameters generalized grammatical constructions were built according to the formula (6). For example set of words in first position of grammatical structure “{He, she, man…} walks” contains only singular pronouns and nouns in masculine and the feminine gender in nominative case followed by singular verb in present tense. Based on that information generalized grammatical construction which contains singular numerals and nouns in masculine and the feminine gender in nominative case on first position and singular verb in present tense in second position is built.

Discussion

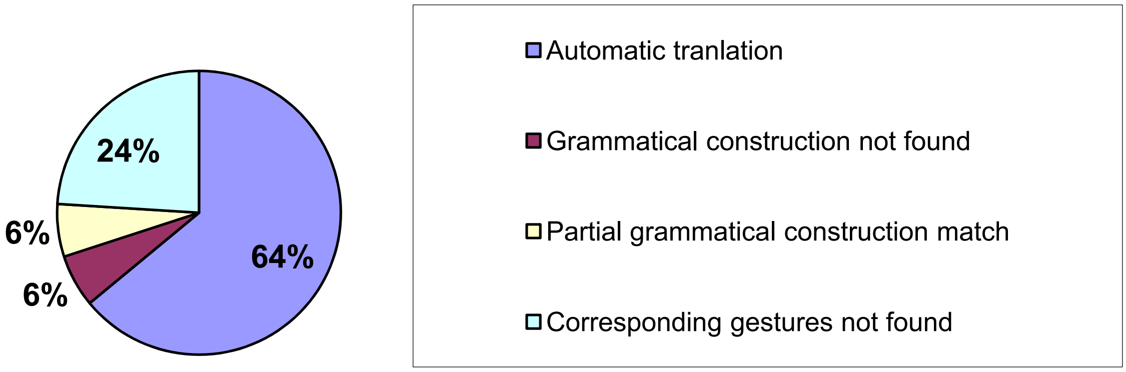


Figure 3. Automated translation system testing results

Using the proposed technology a dictionary of 2 millions words (140 thousand base word forms and 2 thousands flexes) of Ukrainian language was built. Based on the educational program of the "Ukrainian Sign Language" used in schools for the people with impaired hearing to master sign language the set of 10 thousand sentences and a set of 3 thousand gestures were formed. As a result 300 generalized grammatical constructions were built. To test the technology 500 simple sentences were taken from periodical literature for people with impaired hearing. As a result (fig. 3), 64% of sentences were translated automatically, 24% were not translated due to lack of word → gesture corresponding, 12% were not translated due to absence of generalized grammatical constructions (for 55% of these sentences generalized grammatical constructions which partially correspond to sentence were proposed).

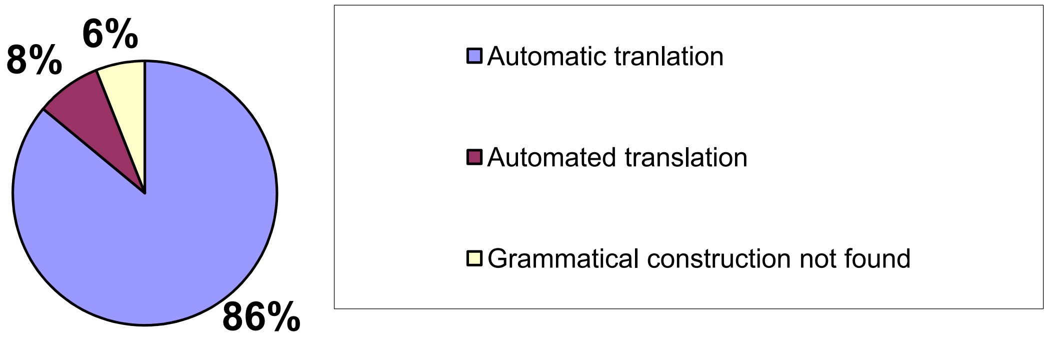


Figure 4. Automated translation system testing results with updated gesture dictionary

After adding the lacking word → gesture dependences we got the result (fig. 4) of up to 94% satisfactory translations (86% automatic and 8% automated translation based on partially corresponding grammatical constructions).

Further research aims to specify satisfactory gesture vocabulary based on corpus of texts used by people with impaired hearing in everyday communication and update generalized grammatical constructions list based on testing results.

Bibliography

[Brinton, 2000] Brinton, Laurel J. (2000). The structure of modern English: a linguistic introduction. – Amsterdam, Philadelphia: John Benjamins, 2000. – 104 p.

[Krak, 2008] Krak Iu., Kryvonos Iu., Barmak O., Ternov A. (2008) Informational Technology for Non-verbal Communication for People with Impaired Hearing // Artificial Intelligence. – 2008. №3. – P.325-331. (in Ukrainian)

[Marshall, 2003] Marshall, I. and Sáfár, E. (2003) A prototype text to British Sign Language (BSL) translation system. // 41st Annual Meeting of the Association of Computational Linguistics, Sapporo, Japan. – 2003 – P.113–116.

[Morrissey, 2010] S. Morrissey, H. Somers, R. Smith, S. Gilchrist and S. Dandapat. Building a Sign Language corpus for use in Machine Translation. // Proceedings of the 4th Workshop on Representation and Processing of Sign Languages: Corpora for Sign Language Technologies. Valetta, Malta. – 2010 – P.172-177.

[Shirokov, 1998] Shirokov, V. Informational theory of lexicographical systems. – Kyiv: Dovira, 1998. – 331 p. (Ukrainian)

[Tomlin, 1986] Tomlin, R. Busic word order. Fundamental principles. – London: Croom Helm, 1986. – 308 p.

[Veale, 1998] Veale, T., Conway, A., Collins, B. The challenges of cross-modal translation: English to Sign Language translation in the Zardoz system. // Machine Translation. – 1998. №13. – P.81–106.

Authors' Information

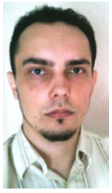
Iurii Krak – *V.M.Glushkov Cybernetics Institute of NASU, senior scientist,*

e-mail: *yuri.krak@gmail.com*

address*: 40 Glushkov ave., Kiev, Ukraine, 03680*

Olexander Barmak – *Khmelnytsky National University, docent,*

e-mail: *alexander.barmak@gmail.com*



Sergiy Romanyshyn *- Khmelnytsky National University, post graduate student,*

e-mail*: serg.romanyshyn@gmail.com*

MODELING OF REASONING IN INTELLIGENT DECISION SUPPORT SYSTEMS BY INTEGRATION OF METHODS BASED ON CASE-BASED REASONING AND INDUCTIVE NOTIONS FORMATION

Alexander Eremeev, Marina Fomina

***Abstract**: Modeling of reasoning in intelligent systems on the example of intelligent decision support system of real time by means of integration of methods based on case-based reasoning (accumulated experience) and inductive notion formation in the presence of noisy data are considered.*

Keywords: *intelligent decision support system, real time, plausible reasoning, modeling, case-based, inductive notion formation, noisy data.*

Introduction

Modern research and development concerning perspective intelligent (expert) decision support systems (IDSS), in particular, IDSS of real-time (IDSS RT) [Eremeev and Vagin, 2011] are directly related to the problem of the modeling plausible reasoning (so called "common sense" reasoning) [Vagin et al., 2008]. The presence of such reasoning modeling methods (inductive, abductive, fuzzy inference, plausible, argumentation, and those based on analogies and cases) in IDSS RT designed for monitoring and management of complex objects (systems) and various processes allows to diagnose of problem situations and aids decision making persons (DMPs) in finding effective managing effects aimed at normalizing the situation.

In this paper, the main attention is given to methods of case–based reasoning and inductive notion formation. The last is applied to situations when a suitable case for a current situation absents in the case library and a corresponding hypothesis must be formed that could constitute a new precedent in the case of its justification.

The methods of case–based reasoning and case retrieval from a system case library (CL) for further usage are considered. The possibility of using different algorithms to retrieve cases is discussed.

Methods and generalization algorithms for searching the regularities are suggested. The problems of dealing with noisy data under searching hidden regularities and choosing control effects in IDSS RT are considered.

Features of an IDSS RT

Now very actual problem in the Artificial Intelligence area is the problem of the construction of intelligent systems, whose typical representative is an IDSS RT oriented to open subject areas and dynamic subject areas [Eremeev and Vagin, 2011].

IDSS RT systems are based on the integration of knowledge representation and knowledge operation models that are capable to adaptation, modification, and learning. Such models are oriented to specific problem areas and respective uncertainty types, what reflects the ability to develop and modify their states.

The generalized structure of an IDSS RT is shown in Fig. 1.

By realizing methods of reasoning modeling in an IDSS RT one should take into consideration the features of these systems:

- the necessity to take a decision under time constrains defined by an actually controlled process;
- the need to consider a time factor in the description of a problem situation in process of finding the solution;
- the impossibility of obtaining all the objective information necessary for decision making and in this connection the usage of subjective expert information;
- the multivariate character of search;
- the necessity of applying methods of plausible reasoning and the active participation of a DMPs in decision making;
- the presence of incomplete, fuzzy and even inconsistent data for description of situations.

The methods of case-based decision search can be used in many IDSS RT units (analyzer, problem solving unit, modeling unit, and prognosis unit) and allow to increase the effectiveness of the DMPs activity in some problematic (irregular) situations.

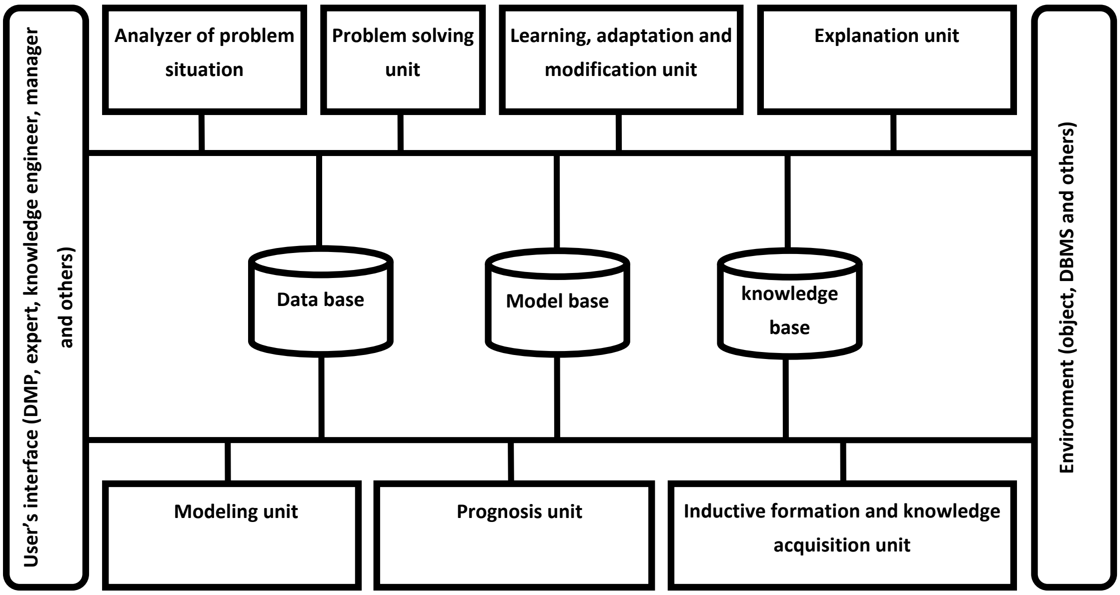

User's interface (DMP, expert, knowledge engineer, manager and others)
Analyzer of problem situation
Problem solving unit
Learning, adaptation and modification unit
Explanation unit
Data base
Model base
knowledge base
Modeling unit
Prognosis unit
Inductive formation and knowledge acquisition unit
Environment (object, DBMS and others)

Fig. 1. General architecture of an IDSS RT

Case–Based Reasoning

A case can be defined as a particular situation that has occurred in the past and can serve as an example or justification for subsequent cases of a similar kind.

Case-Based Reasoning (CBR) is an approach that allows solving a new unknown problem using or adapting the solutions of known problems, i.e., using the experience gained in solving similar problems.

By search the solution in IDSS is reasonable to apply the plausible inference methods that allow to find an applicable solution (that is not optimal). One of the approaches is based on the fact that at the first stage of the solution search of a new unknown problem a person (expert or DMP) tries to use the decisions that were taken previously in similar cases and if necessary adapt them to the problem (the current situation).

This approach on the basis of the saved previous experience became a basis for the modeling case-based reasoning methods.

As a rule, case-based inference includes four main stages that form the so-called cycle of case-based reasoning, or CBR cycle [Aamodt and Plaza, 1994]. The CBR cycle is also called the learning cycle by precedents (examples).

The main CBR cycle stages are:

- Retrieval of the most adequate (similar) case (or cases) for the target situation from CL;
- Reusage of the retrieved case in order to try to solve the target problem;
- Revision and adaptation of the solution if it is necessary to match the target problem;
- Saving (memorizing) of a new solution as part of a new precedent.

The main goal of using the case-based tools within IDSS RT consists in giving a ready solution to an decision making person for the current situation on the basis of precedents which already took place in the past in case of control of this or similar objects.

At the first stage of CBR- cycle (under case acquisition) similarity degrees of a current situation with cases from a case library are performed and subsequent case extraction with the goal of solving a new problem situation is produced. For successful implementation of reasoning on the basis of cases, it is necessary to provide correct extraction cases from a case library.

Commonly, a case includes the following components [Alterman, 1989, David, 1991]:

- The problem description (target situation);
- The problem solution (diagnostics of the target situation and recommendation to DMP);
- The result (or prognosis) of solution application.

The result can include the list of actions to be executed, additional comments, and references to other cases. The case can have both positive and negative outcome solution application; also in some cases the choice of the proposed solution can be substantiated and possible alternatives can be given.

The main methods of case presentation can be divided into the following groups:

- Parametric;
- Object-oriented; and
- Special (graphs, trees, logic formulas, etc.).

In most cases simple parametric presentation is enough to present cases, i.e., presentation in the form of the set of parameters with particular values and solutions (diagnosis and recommendations to DMP): CASE$(x_1,\ldots, x_n, R)$.

In the given description of a case, select a feature constituent $\{x_1,\ldots, x_n\}$, where $x_1,\ldots,x_n$ are the parameters of the situation describing the given case.

In the given case each object is characterized by n parameters (attributes): $A_1, A_2, \ldots, A_n$. Attributes can accept numerical, logical or symbolic values. Denote by $Dom(A_1), Dom(A_2), \ldots, Dom(A_n)$ the sets of admissible values of attributes. For attribute A_k $1 \le k \le n$, $Dom(A_k)$={x_1, x_2, ... x_{q_k} }, where q_k is the number of different values of the attribute A_k. Thus, each situation s_i is represented as a set of attributes values, i.e., $s_i = x_{i1}, x_{i2}, \cdots, x_{in}$, where $x_{ik} \in Dom(A_k)$, $1 \le k \le q_k$. Such a description of a situation connected with case, is called a parametric description.

Commonly, a case includes also the information R i.e. the diagnosis and recommendations to the DMP. Additionally, the description of the results of the solution and additional comments can be present [Eremeev and Varshavskiy, 2008 (1), Eremeev and Varshavskiy, 2008 (2)].

Case-based inference (reasoning) is related first of all with searching the situations in CL analogues to a situation in question, their extraction, assessment and treating.

Methods of case extraction

There are many methods for case extraction and modification. Some methods are based on search cases by similarity: we need to compute measuring the similarity degree between the case and the target situation. For defining a similarity degree, it is necessary to introduce a metric in the parameter space (attributes and properties) to describe cases and the current situation. Then, consequently to chosen metric the distance between the points corresponding to the cases and the point corresponding to the target situation is determined, and a point (a case) that is the nearest to the target situation one should choose.

The approach based on forming inductive descriptions for classes of similar situations is very important. Such approach is related with solving the problem of inductive notion formation or the generalization problem.

Let us give the formulation of feature-based concept generalization. Let S be the set of all situations, represented in a certain IDSS. There is the set V of situations, in which identical or similar decisions were accepted. We can call V class of situations. All situations included in V form the set of positive objects related to some class (concept) and let W be the set of negative objects. We will consider the case where $S = V \cup W$, $V \cap W = \varnothing$. Let K be a non-empty set of objects such that $K = K^+ \cup K^-$, where $K^+ \subset V$ and $K^- \subset W$. We call K a learning sample. Based on the learning sample, it is necessary to build a rule separating positive and negative objects of a learning sample.

Thus, the class description is formed if one manages to build a decision rule which, for any example from a learning sample, indicates whether this example belongs to the class (concept) or not. The algorithms that we use form a decision in the form of rules of the type "IF condition, THEN the class description." The condition is represented in the form of a logical function in which the Boolean variables reflecting the feature values are connected by logical connectives. Further, instead of the notion "feature" we will use the notion "attribute". The decision rule is correct if, in further operation, it successfully recognizes the objects which originally did not belong to the learning sample.

Generalization algorithms

The decision tree *T* is a tree in which each non-final node accomplishes checking of some condition, and in case a node is finite, it gives out a decision for the element being considered. In order to perform the classification of the given example, it is necessary to start with the root node. Then, we go along the decision tree from the root to the leaves until the final node (or a leaf) is reached. In each non-final node one of the conditions is verified.

Depending on the result of verification, the corresponding branch is chosen for further movement along the tree. The solution is obtained if we reach a final node. Decision tree may be transformed into a set of production rules.

Let us consider two algorithms C4.5 and CART, which are based on a procedure of decision tree building.

The algorithm C4.5 as its predecessor ID3 suggested by J.R.Quinlan [Quinlan, 1986, Quinlan, 1996] refers to an algorithm type building the classifying rules in the form of decision trees. However, C4.5 works better than ID3 and has a number of advantages:

- Numerical (continuous) attributes are introduced;
- Nominal (discrete) values of a single attribute may be grouped to perform more effective checking;
- Subsequent shortening after inductive tree building based on using a test set for increasing a classification accuracy.

The algorithm C 4.5 is based on the following recursive procedure:

An attribute for the root edge of a tree *T* is selected, and branches for each possible values of this attribute are formed.

The tree is used for classification of learning set examples. If all examples of some leaf belong to the same class, then this leaf is marked by a name of this class.

If all leafs are marked by class names, the algorithm ends. Otherwise, an edge is marked by a name of a next attribute, and branches for each of possible values of these attribute are created, go to step 2.

The criterion for choosing a next attribute is the gain ratio based on the concept of entropy [Quinlan, 1996].

In the algorithm CART [Breiman et al., 1984], building a binary decision tree is performed. Each node of such decision tree has two descendants. At each step of building a tree, the rule that shares a set of examples from a learning sample into two subsets is assigned to a current node. In the first subset, examples are entered where a rule is performed, and the second subset includes examples where a rule does not perform. Accordingly for the current node, two descendant nodes are formed and the procedure is recursively repeated until a tree will be obtained. In this tree the examples of a single class are assigned to each final node (tree leaf).

The most difficult problem of the algorithm CART is a selection of best checking rules in tree nodes. To choose the optimal rule, there is used the assessment function of partition quality for a learning set introduced in [Breiman et al., 1984].

The important distinction of the algorithm CART from other algorithms of building the decision trees is the use the mechanism of tree cutting. The cutting procedure is necessary to obtain the tree of an optimal size with a small probability of erroneous classification.

Formed by one of generalization algorithms a decision tree can be used under finding the required cases in a CL. For such searching it is necessary to go along a decision tree from a root up to final nodes (leafs of a tree). Such path from a tree root to a final node (a leaf) corresponds to sequence of checking for attribute values describing a current problem situation. A final node corresponds to one or several cases. If a final node is related with some subset of cases then for choosing the most suitable from them, the method of “nearest neighbours” can be used. Such approach is useful for large CL because the time of decision search is significantly reduced.

Noise models

Assume that examples in a learning sample contain noise, i.e., attribute values may be missed or distorted. Reasons of noise arising are described in [Mookerjee et al., 1995]. Our purpose is to study noise effect on the functioning C 4.5 and CART algorithms.

One of basic parameters of research is a noise level. Let a learning sample K $(|K| = m)$ be represented in the table with m rows and r columns, such table has $N=m\cdot r$ of cells. Each table row corresponds to one example and each column – to certain informative attribute. A noise level is a magnitude p_0, showing that an attribute value in a learning or test set will be distorted. So, among all N cells, $N\cdot p_0$ of cells at the average will be distorted. Modeling a noise includes noise models and ways of their entering as well.

For research, two noise models were chosen: “absent values” and “distorted ones”. In the first case for the given noise level with probability p_0, a known attribute value is removed from a table. The second variant of entering a noise is linked with substitution of a known attribute value for another one that may be wrong for the given example. Values for replacement are chosen from domains $Dom(A_k)$, $1 \le k \le r$, where p_0 sets up a probability of such substitution.

At entering a noise of the type “absent values”, it is necessary also to select a way of treating absent values. In the paper two ways are considered: omission of such example and restoring absent values on the “nearest neighbors” method [Vagin and Fomina, 2011].

There are several ways of entering a noise in learning sets [Quinlan, 1986]. Let us consider three ways of entering a noise into a table.

Noise is entered evenly in the whole table with the same noise level for all attributes.

Noise of the given level is entered evenly in one or several explicitly indicated attributes. Entering a noise into the single table column, the content of which is the most important attribute (root node), is an extreme case here.

The new way of irregular noise entering in a table was offered. Here a noise level for each column (informative attribute) depends on a probability of passing an accidentally selected example through a tree node marked by this attribute.

We have:

- A sum noise entered into a table corresponds to the given noise level;
- All informative attributes, values of which are checked in nodes of a decision tree, are put on distortions;
- The more “important” an attribute the higher a distortion level of its values.

Principles of noise level account for the third irregular model are proposed. Let the decision tree T have been built on the basis of the learning sample K. Evidently, an accidentally selected example will passes far from through all nodes. Hence, our problem is to efficiently distribute this noise between table columns (attributes) in correspondence with statistical analysis of DBs having a given average noise level p_0.

For each attribute A_k, find a factor of the noise distribution S_k according to a probability of passing some example through the node marked A_k . Clearly, each selected example from K will pass through the root node of a decision tree. Therefore the value 1 is assigned to the factor S_k of the root attribute.

All other tree nodes which are not leaf have one ancestor and some descendants. Let one such node be marked by attribute A_i and have the ancestor marked by A_q. The edge between that nodes is marked by the attribute value x_j where $x_i \in Dom(A_q)$. Let m be the example quantity in K and m_j be the example quantity in K satisfying to the condition: attribute value for A_q is equal to x_j.

Then the factor of noise distribution

$$S_{A_i} = S_{A_q} \frac{m_j}{m}.$$

The value 0 is assigned to all factors for attributes not using in a decision tree. Introduce the norm

$$S = \sum_{i=1}^{r} S_{A_i}$$

Thus, each attribute A_i, will be undergone to influence of a noise where a noise level is

$$d_{A_i} = \frac{S_{A_i}}{S} \cdot p_0 \cdot r$$

Here p_0 is a given noise level, r is an attribute quantity.

It is easy to see that $\left(\sum d_{A_i}\right)/r = p_0$, i. e. the average noise level is the same as the given one.

Further, we consider the work of the generalization algorithm in the presence of noise in original data. Our purpose is to assess the classification accuracy of examples in a test sample by increasing a noise level in this sample.

Modeling the algorithms of forming generalized notions in the presence of noise

The above mentioned algorithms C 4.5 and CART have been used to research the effect of a noise on forming generalized rules and on classification accuracy of test examples. It should take into account that using the decision tree for classification of an example with absent values can lead to multivariate decisions. Therefore it is necessary to find a possibility of restoring these absent values. To restore unknown values the methods of nearest neighbours (kNN) and choice of average (MORM) are used [Vagin and Fomina, 2010].

To develop the generalization system, the instrumental environment MS Visual Studio 2008, program language C# has been used. The given environment is a shortened version MS Visual Studio. DBMS MS Access was used to store data sets.

The program IDTUV3 performs the following main functions:

- Loads the original data from DB;
- Enters different variants of noise in learning and test sets;
- Builds the classification model (a decision tree, or binary decision tree) on the basis of the learning sample;
- Forms production rules in accordance with the constructed tree;
- Recognizes (classifies) objects using a classification model;
- Statistics on classification quality is formed.

We present experiment results fulfilled on the following three data groups from the known collection of the test data sets of California University of Informatics and Computer Engineering "UCI Machine Learning Repository" [Merz and Murphy, 1998]:

1. Data of Monk's problems;
2. Repository of data of the StatLog project:

– Australian credit (Austr.credit);

3. Other data sets (from the field of biology and juridical-investigation practice).

We can make the following conclusions. A noise in DBs influences essentially on the classification accuracy and on generalization algorithms as a whole.

The noise entered into a test set has essentially larger influence on the classification accuracy than a noise entered in a learning set (on the average up to 5 – 6% at entering a noise up to 30%).

With increasing a noise level, the irregular way of entering a noise has essentially larger influence on the classification accuracy than the uniform way of entering a noise (on the average up to 3 – 4% at entering a noise up to 30%).

Under growth of a noise level, “distortion model” sometimes is able to increase the classification accuracy.

Table 1. Classification results for examples with noise (“distorted values”) by noise entering to test sample.

Data set	**Method of entering a noise**	**Classification accuracy of “noisy” examples, %**				
		No noise	**5% Noise**	**10% Noise**	**15% Noise**	**20% Noise**
MONKS1	*uniform*	82,3	83,53	82,74	83,06	78,14
	root attribute		81,63	81,12	79,73	76,71
	irregular		83,49	81,89	82,01	76,98
MONKS2	*uniform*	88,54	84,43	82,15	79,35	73,5
	root attribute		83,15	79,36	75,28	65,98
	irregular		82,71	80,82	74,68	68,12
MONKS3	*uniform*	85,44	82,35	83,78	79,2	75,89
	root attribute		82,24	80,46	79,81	70,52
	irregular		82,13	81,59	81,37	71,77
GLASS	*uniform*	70,35	68,93	67,03	62,93	59,15
	root attribute		65,34	63,71	61,26	55,74
	irregular		64,48	64,52	63,68	56,61
AUSTRALIAN CREDIT	*uniform*	83,31	82,73	80,57	73,19	69,41
	root attribute		79,34	75,61	73,33	62,01
	irregular		82,14	77,49	74,07	63,71

The method of “nearest neighbours” gives better classification accuracy in comparison with exclusion from a sample of examples with unknown values (on the average up to 8% under a noise level up to 30%).

The dependence of classification accuracy on a noise level at different variants of entering a noise is close to linear.

From three ways of entering a noise, the most influence on the classification accuracy has entering a noise in the root node.

Conclusion

The question of using the methods off modeling plausible reasoning on the basis of nontraditional logic is discussed. For modeling reasoning on the basis of cases in IDSS RT, the basic ways of representing and extracting the cases from case libraries are considered. For effective extraction of cases from a case base, methods of forming generalized descriptions of situation classes on the basis of decision trees building algorithms are used. The ways of solving the information generalization problem under the noise presence in the original data are researched. The new model of irregular noise insertion in informative attributes of a learning sample is offered. The machine experiments on research of noise influence on the work of generalization algorithms C4.5 and CART are produced. It is shown that the new modal of irregular noise insertion significantly influences on the classification accuracy of test examples and is perspective for further research.

The paper has been written with the financial support of RFBR (No. 14-01-00427, 12-01-00589a).

Bibliography

[Aamodt and Plaza, 1994] Aamodt A. and Plaza E., Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, AI Communications, IOS Press, 1994, vol. 7, pp. 39–59.

[Alterman, 1989] Alterman R., Panel Discussion on Case Representation in: Proc. of the 2nd Workshop on Case-Based Reasoning, Pensacola Beach, Fl, US: 1989.

[Breiman et al., 1984] Breiman L., Friedman J. H., Olshen R. A., Stone C. T. «Classification and Regression Trees».— Wadsworth, Belmont, California, 1984.

[David, 1991] David B.S., Principles for Case Representation in a Case-Based Aiding System for Lesson Planning. In: Proc. of the Workshop on Case-Based Reasoning, Madison Hotel, Washington, May 8–10, 1991.

[Eremeev and Vagin, 2011] Eremeev Alexander P. and Vagin Vadim N. Common Sense Reasoning in Diagnostic Systems. In: Practice and Challenges from Current to Future, Chiang Jao (Ed.), ISBN: 978-953-307-326-2, InTech, 2011, pp. 99-120. Available from: http://www.intechopen.com/articles/show/title/common-sense-reasoning-in-diagnostic-systems

[Eremeev and Varshavskiy, 2008 (1)] Eremeev A., Varshavskiy P. Case-based reasoning method for real-time expert diagnostics systems. In: International Journal “Information Theories & Applications”, 2008, Volume 15, Number 2, pp. 119-125.

[Eremeev and Varshavskiy, 2008 (2)] Eremeev A., Varshavsky P. Reasoning by structural analogy taking into account the context for intelligent decision support systems. In: International Book Series ‘Information Science & Computing’, Number 3, ITHEA Sofia, 2008, pp. 9-16.

[Merz and Murphy, 1998] C.J.Merz, P.M.Murphy. UCI Repository of Machine Learning Datasets, (1998) Information and Computer Science University of California, Irvine, CA 92697-3425 http://archive.ics.uci.edu/ml/

[Mookerjee et al., 1995] V. Mookerjee, M. Mannino, R. Gilson: Improving the Performance Stability of Inductive Expert Systems under Input Noise. In: Information Systems Research 6(4), 1995, pp. 328-356

[Quinlan, 1986] Quinlan J. R. The effect of noise on concept learning. In Machine Learning Vol. II (Michalski R. S., Carbonell J. G. and Mitchell T. M., eds.) Chapter 6. Palo Alto, CA: Tioga, 1986

[Quinlan, 1986] Quinlan J.R.: Induction of Decision Trees. In: Machine Learning 1, 1986, pp. 81-106

[Quinlan, 1996] Quinlan J.R.: Improved Use of Continuous Attributes in C 4.5. In: Journal of Artifical Intelligence Research 4, 1996, pp. 77-90

[Vagin and Fomina, 2010] V. Vagin, M. Fomina. Methods and Algorithms of Information Generalization in Noisy Databases. In: Advances in Soft Computing. 9th Mexican Intern. Conference on AI, MICAI 2010, Pachuca, Mexico, November 8-13, 2010, Proceedings, Part II. / G. Sidorov, A.H. Aguirre, C.A.R. Garcia (Eds). Springer Verlag Berlin, 2010, pp. 44-55

[Vagin and Fomina, 2011] V.Vagin , M. Fomina. Problem of Knowledge Discovery in Noisy Databases. In: International Journal of Machine Learning and Cybernetics. Vol. 2, Number 3, 2011, pp. 135-145

[Vagin et al., 2008] Vagin, V.N., Golovina, B.Yu., Zagoryanskaya A.A., Fomina M.V. Exact and Plausible Reasoning in Intelligent Systems./Eds. Vagin, V.N. and Pospelov, D.A., Moscow; FIZMALIT, 2008 (in russian).

Authors' Information

Alexander Eremeev – Head of the Applied Mathematics Department, Professor, National Research University „Moscow Power Engineering Institute“, 111250 Krasnokazarmennaja str., 14, Moscow, Russia; e-mail: eremeev@appmat.ru

Major Fields of Scientific Research: Intelligent (expert) decision support systems

Marina Fomina – Computer Science Department, Docent, National Research University „Moscow Power Engineering Institute“, 111250 Krasnokazarmennaja str., 14 , Moscow, Russia; e-mail: m_fomina2000@mail.ru

Major Fields of Scientific Research: Inductive notion formation, Knowledge discovery in Databases.

FORESIGHT PROCESS BASED ON TEXT ANALYTICS

Nataliya Pankratova, Volodymyr Savastiyanov

Abstract: *The significant increase in number of information sources unfavorable affects on traditional foresight techniques not directly adapted to big data era. Without automation of knowledge processing the quality of final foresight product is significally dependent on human (experts, analytics) abilities. In the article the new process workflow is proposed using text analytics tools to support all stages of foresight. The proposed advanced model of fact extraction with modified rules is based on new workflow, which includes marking data with additional metadata, using automated classification and sentiment extraction techniques, data quality improving steps in addition to quantitative and qualitative analysis of data. The modified rule based model of knowledge extraction adapted to used toolkit is presented. Given approach were tested on supporting of foresight process in domain of agricultural development of Crimea region.*

Keywords: *foresight, decision making, textual analytics, sentiment analysis, knowledge society, data mining, DSS.*

ACM Classification Keywords: *H.5.3 Group and Organization Interfaces - Computer-supported cooperative work*

Introduction

We are living in knowledge society now [Eurofound, 2003]. The knowledge about short and long-term trends is the major driving forces not only for technologies but society. The knowledge about markets, user preferences and requirements is vital for strategy decision making [Eurofound, 2003; Pillkahn, 2008]. The scale how to utilize this knowledge is very varied: from studies how to make short, middle and long-term future less unpredictable till how to get future more favorable to human, organization, region or even country initiatives.

The one of the importunes issue which is another side of knowledge basis - we have enough information to make any decision using all sources in the only case we are capable of to grab, separate, aggregate, clean and process all this available knowledge [Pillkahn, 2008]. Foresight study is the known approach how to partly deal with this problem [UNIDO, 2005]. There are numerous agencies well known for their foresight studies and various authors or scientific groups leading their own foresight management strategies. The most part of these studies was founded in the period of data availability, access and aggregation technologies development (i.e. storage, databases software, OLAP technologies). So, today all mentioned technologies lead to the problem of big data in the world scale, what is really insuperable obstacle even with foresight methodology on the way to make decision horizon both desirable and trustworthy [Zgurovsky and Pankratova, 2007].

The problem is even harder due to widespread of communication technologies and social networks, blogs technologies and data mining tools. All this technologies drives not only the rising speed of data and knowledge interchange and processing, but also the morbid dependency of even small people’s actions from information flows. Receiving false (opinion spam), inequality or specially wrong formed (misinformation) knowledge from huge number of sources forces key figures, decision makers and experts select week strategy or act without strategy under uncertain condition and time shortage due to panic or inadequate reaction of community [Gladwell, 2001; O'Reilly, 2013]. Also fixed-date absence of any key figures actions is encouraging instability of situation [Martin et al, 2006], causing money, time, resources, power or even human losses [Zgurovsky and Pankratova, 2007; Gladwell, 2001].

So, taking into account information above, the big supply to any decision support methodology is the strategy to stay as long as possible well informed using advanced knowledge representations and modern data mining techniques during the time flow before critical time threshold set in.

1. Current information model of foresight process

In this paper the information model of foresight process implemented in the information platform of scenario analysis (IPSA) is using, which combines classical theory of foresight adopted to Ukrainian conditions and advanced mathematics in expert's opinion processing [UNIDO, 2005; Zgurovsky and Pankratova, 2007]. The analytical level of IPSA consists of numerous organizational stages, methods of quality and quantity analysis, and special mathematical methods of expert's opinion aggregation for all cases of typical input and combination of initial information for particular domain [Zgurovsky and Pankratova, 2007]. The process of current IPSA operation is shown on Figure 1.

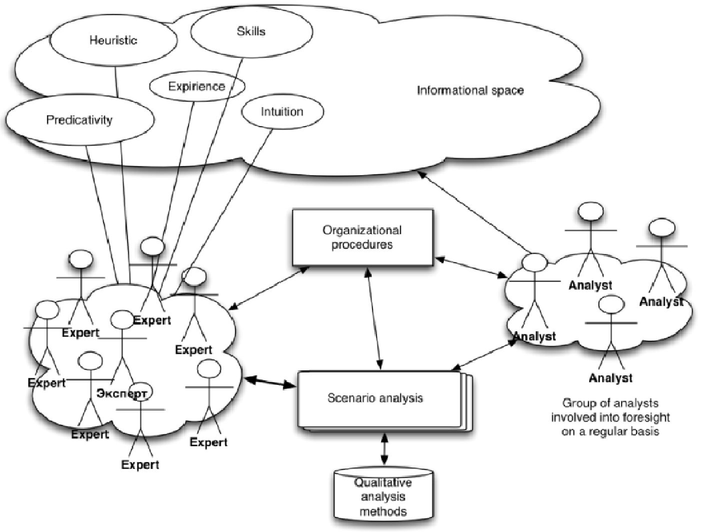


Figure 1. Operation of current foresight information platform of scenario analysis

The process of operation is based on scenario analysis involving expert's knowledge processing through the mathematical instrumentality of qualitative and quantitative methods conforming expert judgements in a number of steps guided by organizational procedures. These judgements are reflecting researched subject area with trends and estimates [Zgurovsky and Pankratova, 2007]. A group of analysts involved into foresight on a regular basis are managing the plan of foresight process and supply expert's panels with initial information. The group of analysts also interacts with key figures that are the end users of foresight. The final product of foresight is a set of scenario alternatives with a priory chosen time scale in the future and concerning a priory chosen subject area.

All steps of foresight are performing with a help of IPSA, which is both an information portal for end users and interface to the processing libraries with qualitative and quantitate mathematical methods.

The integrity of knowledge base and researched subject domain coverage in foresight process is a question of the group of analysts' proficiency and is not supported by any computer aided technology on the mentioned above stage of IPSA development.

2. Metadata used in current information model of foresight process

Current foresight process reflected in information platform of scenario analysis utilizes some special information entities for intersteps knowledge exchange. These entities are described in table below (Table 1).

Table 1. Metadata of foresight process

№	Metadata name	Metadata description and source
1	Time scale	On organizational stage
2	Goal	Goals of foresight
3	Idea	Sources: environmental scanning, brainstorm
4	Cluster of ideas	Sources: environmental scanning
5	Expert's estimate	Sources: Delphi, Saaty method, cross-impact, morphological analysis
6	Key technology	Sources: environmental scanning, brainstorm, morphological analysis
7	Driving force	Sources: environmental scanning, brainstorm, morphological analysis
8	Scenary	Sources: scenarios, STEEPV (STEEEPVA)
9	Roadmap	Sources: Roadmapping

The current metadata are descriptive by nature and allow only building knowledge storage system. Some of produced by qualitative methods entities are semistructured [Buneman, 1997] and physically are stored as textual blocks with numbers, facts, conclusions or other knowledge. Further in the article this metadata would be referred to as metadata of type I (or just *metadata-1*).

3. Big-data-ready information process of foresight

Big data readiness in foresight is mentioned in sense of how to deal with rapidly increasing number of new knowledge sources under time pressure. There is a strong need to enforce all participant of foresight process (i.e. analysts, experts, key figures and end users) to stay at equal level of being kept informed during foresight process. At the same time humans are limited in their decision making ability due to limitations in their ability to retain and process information [Simon, 1956]. So there is a strong need in the automated tools for relevance analysis of new knowledge generated in every current period of time: previously generated scenario, extracted trends, facts, key figures, drivers and other knowledge; scenario alternatives which were not chosen by decision makers; hypothetical events from data scientists, future studies, science fiction and other sources.

We should take into account the text-based nature of mostly (up to 80%) current and future sources of qualitative business data (news, social networks, blogs, reviews, reports, transcripts, legal papers, emails, etc.) [Berry and Linoff, 2011; Ryan and Bernard, 2000] and currently existing experience storage types in information platform of scenario analysis (see Table 1). The most appropriate tools for processing that information sources is text analytics toolkit. Using advanced automated toolkit for knowledge mining brings new data and metadata into the foresight process for sake of improving scenario trustworthiness and quality.

Text analytics toolkit today is reach on various technics of data aggregation, classification, fact extraction, topic mining, ontology mining, features or aspects extraction and opinion mining [Berry and Linoff, 2011; Chakraborty et al, 2013]. Some of these tools could directly improve certain qualitative methods of foresight some involved into other methods could support the foresight process to improve the quality of results.

4. Advanced information model of foresight process

Advanced information model of foresight strategy additionally includes specially designed knowledge database, module of information quality assessment, and module of foresight process supporting (Figure 2).

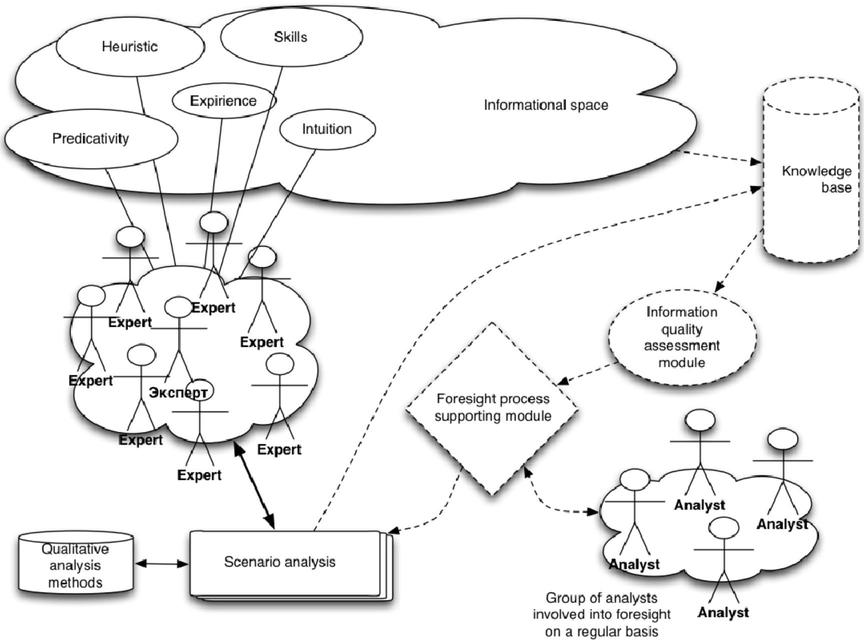


Figure 2. Operation of advanced foresight information platform of scenario analysis

The knowledge base is supposed to store both primary and processed structured data, such as sets of knowledge and facts, trends, metadata, structural and functional relationships, etc.

Information quality assessment module provides an assessment of the consistency and relevancy of knowledge representation in database regarding to subject domains, which are formed by all gathered knowledge: data from real objects and systems, hypothetical objects, systems and notion about them in the representation of the experts, also the signals from the external environment in certain time slices of dynamic changes in the behavior of the studied system and its environment during the process of foresight.

Foresight process supporting module combines forming of organizational procedures by schedule (queue of procedures) with automated notifications to the participants that would have been sent on a base of automated information quality assessment. Forming a queue of procedures is necessary: to enhance on regular basis the reliability and trustworthiness of scenarios by improving the quality of a priori mined information and knowledge; to adjust the foresight process to new knowledge utilizing during its course.

5. Strategy of foresight process support

Information platform of scenario analysis receive structured and semistructured data objects *Obj* reflecting real world on their input from the informational space (i.e. all available digital sources of knowledge):

$$Obj=< Obj_{St<i>}, Obj_{SSt<j>} >, i=1,I, j=1,J.$$

On the data standardization step all data is storing into the relational database as raw data. Structured Obj_{St} data would be useful for various statistical processing and direct representation with OLAP or visual analytics technologies on later stages. All semistructured Obj_{SSt} data, mostly in text representation (or XML with metadata), is also storing as RAW data into the relational database with minimal metadata if available:

$$Obj_{SSt<j>} = < ID, Body, CrawlDte, Author, SourceID, SourceDte >,$$

where *ID* is unique document identifier, *Body* - is raw data, *CrawlDte* - is crawl date, *Author* is author of source information, *SourceID* is unique document identifier in source stream (i.e. URL, etc), *SourceDte* - date extracted from source.

In the next step stored semistructured data is processing by *information quality assessment module* to retrieve additional metadata of type I and II. The scheme of the strategy is shown on Figure 3.

According to the mentioned strategy on the first stage all data is processing by *information quality assessment module* for sake to enrich the mined knowledge with additional metadata could be extracted from raw data and combine it coupled with statistical data into frame based hierarchy which reflects the examining subject area(s) [Liu, 2012]. The *information quality assessment module* is combining text analytics framework based on classification, fact extraction and sentiment analysis tools. There are a lot of tools and approaches how to do this tasks with help of rule based algorithms or machine learning [Liu, 2012]. Currently the mentioned strategy had been tested with help of SAS(R) Text analytics toolkit [Chakraborty et al, 2013].

Two level classifications are used: based on general classifiers to expose the density of subject domains covering by sources and foresight advanced metadata classifiers with fact extraction step.

As a general classifier could be used IPTC classifier from the standard SAS(R) package. There are also a number of others classifiers according to the different subject areas available or could be built [Ryan and Bernard, 2000; Moldovan and Girju, 2001]. General classification is very important step to identify possible interrelations between examining subject areas in available sources. All extracted metadata is storing to the knowledge base.

A special approach to combining classification with fact extracting step is using to identify special metadata of type II (metadata-II) (Figure 4) in advanced foresight process. It is using to separate following entities from input data:

- Goal phrases;
- Structure and interrelations of entities;
- External environment influence and interaction points with external trends;
- Time horizon of trends and gathered knowledge;

- Problem identification;
- Effect detection with sorting on future effects and past facts, suggestion.

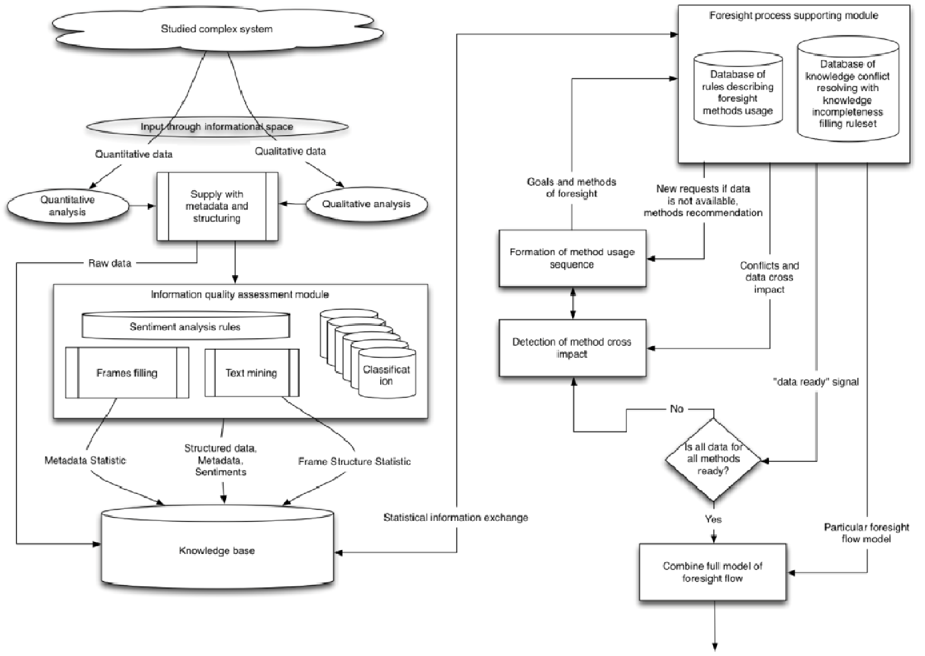

Studied complex system
Input through informational space
Quantitative data
Qualitative data
Quantitative analysis
Supply with metadata and structuring
Qualitative analysis
Raw data
Information quality assessment module
Sentiment analysis rules
Frames filling
Text mining
Classification
Metadata Statistic
Structured data, Metadata, Sentiments
Frame Structure Statistic
Knowledge base
Statistical information exchange
Foresight process supporting module
Database of rules describing foresight methods usage
Database of knowledge conflict resolving with knowledge incompleteness filling ruleset
Goals and methods of foresight
New requests if data is not available, methods recommendation
Formation of method usage sequence
Conflicts and data cross impact
Detection of method cross impact
"data ready" signal
No
Is all data for all methods ready?
Yes
Particular foresight flow model
Combine full model of foresight flow

Figure 3. Scheme of foresight process support strategy

The mentioned approach is built on a base of general model of facts extraction from natural language texts [Simakov et al, 2006], supplemented by a set of rules for sentiment extraction and adopted to using SAS(R) text analytics toolkit. The modified model with extended set of rules is given below:

$$E=<T , V , a> , \qquad (1)$$

where T are all text objects (documents) from input data, V - all rules, a - logical function $a(t_i,v_j)$ which is «True» if t_i satisfies v_j.

The difference of the proposed model from general model of facts extraction from natural language texts [Simakov et al, 2006] is that text is not a sequence of words, but sequence of paragraphs, sentences and then words:

$$t = par_1par_2\ldots par_{|N|},$$

$$par = sent_1sent_2\ldots sent_{|M|},$$

$$sent = wrd_1wrd_2\ldots wrd_{|K|},$$

$$wrd_k \in\{Wrds, Pnkt, POS\_tag, * , Aa\},$$

where *Wrds* are list of words, *Pnkt* is punctuation, *POS_tag* are part-of-speech tags, * is any single word, *Aa* is any word starting from capital letter. As in [Simakov et al, 2006] we could present a text fragment in a view of:

$$t=t_1 + t_2 + \ldots +t_j.$$

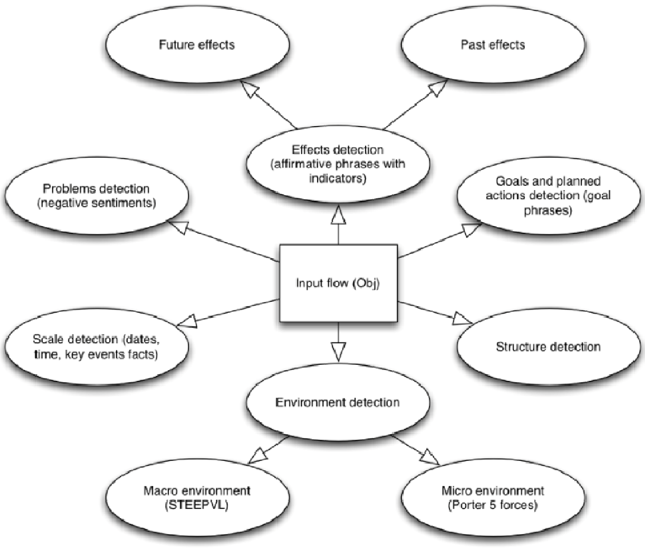

Future effects
Past effects
Effects detection (affirmative phrases with indicators)
Problems detection (negative sentiments)
Goals and planned actions detection (goal phrases)
Input flow (Obj)
Scale detection (dates, time, key events facts)
Structure detection
Environment detection
Macro environment (STEEPVL)
Micro environment (Porter 5 forces)

Figure 4. The structure of additional metadata in advanced foresight process

Needed set of fragments $T_{ri}^{q} = \{t\}$ in every text is covered by pattern r_i^q and all text is covered by all possible fragments $T_r = U\ T_{ri}^{q}$. Pattern $r_i^q = <c, e, d>$, where *c* is lexical restriction, *e* is exception from lexical restriction, *d* is rule scope and $d \in \mathbf{N}$, $q \in \{1,2,3,4,5,6,7,8\}$. For convenience we limit $0 < d \leq 170$. Exception from lexical restriction *e* is implicitly equal empty set in some rules to make the patterns compatible to the SAS(R) text analytics toolkit.

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{1} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \forall j\ w_{ij} \in \{t\},\ j \geq 2 \\ |\,t\,| \geq 2 \\ \forall j\ w_{ij} \in c, \forall j\ w_{ij} \notin e,\ e = \varnothing \end{cases}$$

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{2} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \exists j\ w_{ij} \in \{t\},\ j \geq 1 \\ |\,t\,| \geq 1 \\ \forall j\ w_{ij} \in c, \forall j\ w_{ij} \notin e,\ e = \varnothing \end{cases}$$

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{3} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \exists j\ \exists sent_k w_{ij} \in sent_k,\ k \geq 1, j \geq 2 \\ t \in \{sent\}, |\,sent_k| \geq 2 \\ \forall j\ w_{ij} \in c, \forall j\ w_{ij} \notin e,\ e = \varnothing \end{cases}$$

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{4} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \forall j\ w_{ij} \in \{t\},\ j \geq 2 \\ |\,t\,| \geq 2, \forall a,b \in \{j\}\ dist(w_{ia}, w_{ib}) \leq d \\ \forall j\ w_{ij} \in c, \forall j\ w_{ij} \notin e,\ e = \varnothing \end{cases}$$

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{5} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \forall j\ w_{ij} \in \{t\}, j \geq 2 \\ \left| t \right| \geq 2, \forall a,b \in \{j\},\ a < b \\ \forall j\ w_{ij} \in c, \forall j\ w_{ij} \notin e,\ e = \varnothing \end{cases}$$

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{6} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \forall j\ w_{ij} \in \{t\}, j \geq 2 \\ \left| t \right| \geq 2, \forall a,b \in \{j\}\ dist(w_{ia}, w_{ib}) \leq d,\ a < b \\ \forall j\ w_{ij} \in c, \forall j\ w_{ij} \notin e,\ e = \varnothing \end{cases}$$

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{7} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \forall j\ w_{ij} \in \{t\}, j \geq 3 \\ \left| t \right| \geq 3 \\ \forall a,\ b,\ c \in \{j\},\ a < c < b,\ w_{ia},\ w_{ib} \in c,\ w_{ic} \notin e \end{cases}$$

$$\forall p\ s_p \in s\ \forall t \in T_{ri\ s}^{8} \rightarrow \begin{cases} \exists w_{ij} \in \{w_i\}, \forall j\ w_{ij} \in \{t\}, j \geq 1 \\ \left| t \right| \geq 1 \\ \forall j\ w_{ij} \notin e,\ e = \varnothing \end{cases}$$

$T_{ri}{}^{8}{}_{s}$ pattern currently not presented in the used toolkit (e is explicitely set to Ø), but it is added to the model to theoretically fulfill also possible negation pattern. Also in other rules e is explicitely set to Ø to satisfy used software toolkit limitations. The final extraction rule in set of rules V has a form:

$$v_i = (\{<p_j, arg_j>\}, s, w), j \geq 1,$$

where argument name $arg_j \in \{Ø, \{a \div z\}\}$, *s* is sentiment, *w* is weight of the rule, $s \in \{-1;0;1\}$, $w \in \boldsymbol{R}$ (taken $w \in (0;10]$). Also, there could be modifications: $s \in <\{+,=,-\},\{-2;-1;0;1;2\}, \{-3;-2;-1;0;1;2;3\}>$ according to human recognition ability [Miller, 1956]. When arg_j = Ø there is no facts to be extracted, only matching is needed. The main advantage of proposed model on [Simakov et al, 2006] is the principle how the sequence of fact could be extracted. According to the definition, sequence of facts could be returned into different arguments or sequentially extracted and concatenated into the one single argument. There could be present or absent prefixes and postfixes around any fact should be extracted.

With a help of sentiment colored rules it is possible to identify positive, negative or neutral trends in external environment, effects of possible action plans of authorities, identify problems. To identify them six predefined conceptual categories for sentiment rules classification with fact extraction ability could be used: simple sentiment word or phrase; decreased and increased quantity of an opinionated item; high, low, increased and decreased quantity of a positive or negative potential item; desirable or undesirable fact; deviation from the norm or a desired value range; produce and consume resource and waste [Liu, 2012]. This classification is also useful for automated opinion extraction from expert's conclusion or scenario essay in free form (text form).

The heart of any knowledge, sentiment, fact extraction and classification system is the set of available subject domain's feature hierarchies (product features) could be used with extraction rules. There is a known technique to extract product features, which imply opinions [Liu et al, 2011]. It allows automatically extract possible features from subject domain, which are declaring desirable or undesirable facts. This advanced technique is not utilizing in current work, only automated extraction of noun phrases with manual sorting and filtering for Russian and Ukrainian languages was done instead. The words and phrases mined as products and features were used in sentiment rules for subject domains processing. The rules were written using SAS(R) sentiment analysis studio according to the proposed model (**100**) with wide range of special Boolean rule modifiers («OR», «AND», «SENT», «ORD», «DIST», «ORD_DIST», «UNLESS») and special rules types available [Reckman et al, 2013].

Processed and stored into the knowledge base data then enters on the input of foresight process supporting (FPS) module. As a part of automated tools of FPS there are *database of rules describing foresight methods usage* and *database of knowledge conflict resolving with knowledge incompleteness filling rule set.*

Database of rules describing foresight methods usage consist of hierarchy of frames, which slots are describing input, output and control of every method and rules of process activation which are required to fulfill the slot to complete every particular frame. In other words this database consists of workflow and all metadata of particular foresight process, in this case according to Ukrainian Foresight Program [Zgurovsky and Pankratova, 2007]. Currently all methods of that foresight approach get strong formalization, allowing doing the automation [Zgurovsky and Pankratova, 2007].

Database of knowledge conflict resolving with knowledge incompleteness filling rule set help to complete the frame hierarchy, resolve conflicts of knowledge and discover new knowledge and relations. One part are just «if … else» rules which reflect the process to complete all slots of frames. If there is no possibility to find some knowledge the process ends forming the message to particular expert (experts group) to fulfill incomplete knowledge with help of analysts' personnel. It should be recalled that foresight is not isolated from human computational process, so recall to expert judgement or utilizing human brain resources is normal foresight workflow.

Another part of rules are knowledge conflict resolving rules. There could be extracted group of facts that consist of opinions, subjective information, not proven by authorities or government instances numbers, expectations, external trends and so on. All this fact could conflict or strengthen each other. This conflict could be resolved manually by experts or due to rule sets, in comparison with qualitative data and forecasts. Also there are known number of techniques to support or reduce expensive brainwork. In the work of Lerman and McDonald [Lerman and McDonald, 2009] shown the technique how to summarize mined opinion in case of different product features, that is mean we could summarize mined data to show the expert group «bird view» picture in form of colored radar chart with marked divergences, exclusive features, numbers and polarities of sentiments. Other technique to summarize mined opinion and resolve conflicts is direct usage of foresight methods: Delphi [Pankratova and Malafeeva, 2012], Saaty [Pankratova and Nedashkovskaya, 2013] and others [Zgurovsky and Pankratova, 2007]. In addition the mined by *information quality assessment module* into knowledge base products and features could be compared with product and features from the ontologies with structured information that are available online. Lu at al. [Lu et al, 2010] proposed methods for selecting a subset of aspects from the ontology that can best capture the major opinions, including size-based, opinion coverage-based, and conditional entropy-based methods. They also have done the approach to order aspects, give measures for quantitative evaluation of both aspect selection and ordering and give the way to discover new aspects. In this paper all mined products and features are only compared with industry ontology by set intersection to find out coverage of features with subject domain. In future work implementing of mentioned techniques and approaches is planned.

Step by step by discovering all black holes and eliminating conflicts in knowledge base the foresight process through sequences of foresight method usage is heuristically forming. In this algorithm two important cycles to form final schedule should be separated: sequence of foresight method usage is forming, interaction of foresight method depending on the input data is forming. To avoid infinite loops of knowledge mining with open set of knowledge horizon expert methods are used to terminate the loop. Within every chosen method a priory relevancy level is given or could be accepted during the process.

6. Foresight of agricultural development of Crimea region

Given approach were tested on supporting of foresight process in domain of agricultural development of Crimea region. The source documents from Crimea authorities, expert opinions, interviews and other provided information sources were collected, structured and processed. In addition supplementary data was retrieved from reviews of agricultural development central parts of Ukraine, Russia and Belorussia, official programs of development from Russia and Belorussia, existing official programs of some agricultural branch development from Ukrainian ministry of agriculture, transcripts from Ukrainian parliament sessions, reviews of agricultural technologies and news sources. The CTEA-classifier[1] were used as a general classifier on a first step of classification. Initial rules for CTEA classifier were written by group of analysts manually. The processed data were put into knowledge base as hierarchy structured facts (Table 2).

Table 2. Knowledge base statistic (domain of agricultural development of Crimea region)

Parameter	Q-ty	Hierarchy type for storage
All objects	25360	Static (Structural)
Objects relevant to CTEA	406	Static (Structural)
Objects in trends in agricultural domain (Ukraine + Crimea)	11991	Functional
Objects in problems in agricultural domain (Ukraine + Crimea)	2000	Functional
Objects in goals in agricultural domain (Ukraine + Crimea)	1862	Functional
Non-unique objects in agricultural domain	7060	Static (Structural)
Technologies in agricultural domain	378	Functional
Problems in agricultural domain	225	Functional
Trends in agricultural domain	1385	Functional
Goals in agricultural domain	112	Functional

After all stages in knowledge base were formed Trends group, Problems group, Goal group, Technologies group, Key Factors group with given statistic:

- Final Trends group consist of 12 major trends;
- Final Problems group consist of 143 significant problems;
- Final Goals group consist of 82 goals;
- Final Technologies group consist of 253 technologies;
- Key Factors group consist of 35 factors.

[1] State Standard of Ukraine approved a new classifier DK 009:2010 “Classification of types of economic activities” (CTEA), which corresponds to the European standards and requirements.

Taking a part in similar foresight researches of the same scale as a member of foresight analysts group the author has noticed that new approach saves up to 35% of time on knowledge processing in comparison with traditional foresight process. Also the modified foresight process was given new metrics in mean of knowledge extraction dynamic progress. It is possible to introduce new foresight process's KPI based on knowledge quality and knowledge coverage in addition to traditional organizational KPIs of traditional foresight. The integrity of knowledge base and researched subject domain coverage in foresight process is a question of the group of analysts' proficiency and is not supported by any computer aided technology on the mentioned above stage of IPSA development.

Conclusion

The significant increase in number of information sources unfavorable affects on traditional foresight techniques not directly adapted to big data era and need to implement knowledge mining tools into the process. The organizational nature of foresight combined with utilizing of experts' opinion leads to unpredictable increase of time and costs needed to knowledge convergence in organization with foresight process in big data era. To contradict described problem modified structure of foresight process were proposed. Strategy of foresight process support using text analytics tools was implemented.

The proposed advanced model of fact extraction with modified rules is based on new strategy, which includes marking data with additional metadata, using automated classification techniques in addition to quantitative and qualitative analysis of data. Classification, fact extraction and sentiment analysis tools help to find new interrelations of available knowledge and significantly increase and equalize the level of awareness of foresight participants which leads to increasing confidence level of final foresight product.

The approaches were tested on supporting of foresight process in domain of agricultural development of Crimea region. Implementing foresight-supporting tools based on text analytics could save up to 35% of time on knowledge processing. In addition it is possible to introduce new foresight process's KPI based on knowledge quality and knowledge coverage complementary to traditional organizational KPIs of traditional foresight.

Bibliography

[Berry and Linoff, 2011] Berry, Michael J. A., and Gordon S. Linoff. Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, 3rd edition. Wiley Computer, 2011.

[Buneman, 1997] Peter Buneman Semistructured data. In: Proc. ACM Symposium on Principles of Database Systems, pp. 117-121, Tucson, AZ., Abstract of invited tutorial, 1997.

[Chakraborty et al, 2013] Goutam Chakraborty, Murali Pagolu, Satish Garla. Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, SAS Institute Inc., 2013.

[Eurofound, 2003] European Foundation for the Improvement of Living and Working Conditions. Handbook of Knowledge Society Foresight. 2003. <http://www.eurofound.europa.eu/pubdocs/2003/50/en/1/ef0350en.pdf>.

[Gladwell, 2001] Gladwell, M.: The Tipping Point. How Little Things Can Make A Big Difference. Boston: Little, Brown & Company, 2001.

[Lerman and McDonald, 2009] Lerman, Kevin and Ryan McDonald. Contrastive summarization: an experiment with consumer reviews. in Proceedings of NAACL HLT 2009: Short Papers. 2009.

[Liu, 2012] Liu B., Sentiment Analysis and Opinion Mining, ISBN-10: 1608458849, ISBN-13: 978-1608458844, Morgan & Claypool Publishers, 2012

[Liu et al, 2011] Zhang, Lei and Bing Liu. Identifying noun product features that imply opinions. in Proceedings of the Annual Meeting of the Association for Computational Linguistics (short paper) (ACL-2011). 2011b.

[Lu et al, 2010] Lu, Yue, Huizhong Duan, Hongning Wang, and ChengXiang Zhai. Exploiting Structured Ontology to Organize Scattered Online Opinions. In Proceedings of Interntional Conference on Computational Linguistics (COLING-2010). 2010.

[Martin et al, 2006] Martin, B., Cashel, C., Wagstaff, M., & Breunig, M. Outdoor Leadership: Theory & practice. Champaign, IL: Human Kinetics, 2006.

[Miller, 1956] George A. Miller, "The Magical Number Seven," Psychological Review (March 1956), vol. 6 j , no. 2., 1956.

[Moldovan and Girju, 2001] Dan Moldovan and Roxana Girju, An Interactive Tool For The Rapid Development of Knowledge Bases. In International Journal on Artificial Intelligence Tools (IJAIT), vol 10., no. 1-2, March 2001.

[O'Reilly, 2013] O'Reilly O'Reilly Media, Inc., Big Data Now: Current Perspectives from O'Reilly O'Reilly Media, Inc., O'Reilly Media, 2013, ISBN: 978-1-449-37420-4.

[Pankratova and Malafeeva, 2012] N.D. Pankratova, L.Y. Malafeeva Formalizing the consistency of experts' judgments in the Delphi method // Cybernetics and Systems Analysis: Volume 48, Issue 5 (2012), Page 711-721, 2012.

[Pankratova and Nedashkovskaya, 2013] Pankratova N., Nedashkovskaya N. Estimation of Sensitivity of the DS/AHP Method While Solving Foresight Problems with Incomplete Data // Intelligent Control and Automation, v.4, №1. – 2013. - P. 80-86

[Pillkahn, 2008] Ulf Pillkahn, 'Using Trends and Scenarios as Tools for Strategy Development', Siemens, 2008.

[Reckman et al, 2013] Hilke Reckman, Cheyanne Baird, Jean Crawford, Richard Crowell, Linnea Micciulla, Saratendu Sethi, and Fruzsina Veress. Rule-based detection of sentiment phrases using SAS Sentiment Analysis, Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, pages 513-519, Atlanta, Georgia, June 14-15., 2013.

[Ryan and Bernard, 2000] Ryan, G., & Bernard, R. Data management and analysis methods. In N. Denzin& Y. Lincoln (Eds.), Handbook of Qualitative Research (pp. 769–802). Thousand Oaks, CA: Sage, 2000.

[Simakov et al, 2006] Andreev A, Berezkin D., Simakov K. The model of fact extraction from natural language texts and the learning method. RCDL 2006, <http://www.ixlab.ru/pub/docs/RCDL_2006_1.pdf>

[Simon, 1956] Simon, H. A. Rational choice and the structure of the environment. Psychological Review, 63,129-138.,1956.

[UNIDO, 2005] UNIDO, TECHNOLOGY FORESIGHT MANUAL, Vienna, 2005, ISBN 978-3-89578-304-3

[Zgurovsky and Pankratova, 2007] Zgurovsky, Mikhail Z., Pankratova, N.D., System Analysis: Theory and Applications, Springer, 2007, ISBN 978-3-540-48880-4.

Authors' Information

***Nataliya Pankratova** – DTs, Professor, Deputy director at Institute for applied system analysis, National Technical University of Ukraine "KPI", Av. Pobedy 37, Kiev 03056, Ukraine; e-mail: natalidmp@gmail.com*

Major Fields of Scientific Research: System analysis, Theory of risk, Applied mathematics, Applied mechanics, Foresight, Scenarios, Strategic planning, information technology

***Volodymyr Savastiyanov** – Researcher at Institute for applied system analysis, National Technical University of Ukraine "KPI", Av. Pobedy 37, Kiev 03056, Ukraine; e-mail: vvs@iasa.com.ua*

Major Fields of Scientific Research: system analysis, foresight, future studies, text analytics, sentiment analysis, data mining, knowledge mining, statistical analysis, artificial intelligence.

IMPROVED ALGORITHM FOR VIDEO SHOT DETECTION

David Asatryan, Manuk Zakaryan

Abstract: *Currently digital video are widely used in various fields of science and technology and in human daily activities. Intensively growing and already existing huge amount of digital video data need to be managed, so the shot boundary detection is the first and important step for content-based video retrieval and indexing. Each algorithm aimed for this approach should accurately detect boundaries between camera shots and do a segmentation of a video. In this paper a new method of abrupt transitions detection is proposed, based on Weibull distribution of each frame's gradient magnitude. Experimental results successfully show that it can effectively detect hard cuts and has certain advantages against widely used other methods, which are using image pixels point-by-point comparison method.*

Keywords: *segmentation, cut detection, similarity measure, Weibull distribution.*

ACM Classification Keywords: *Image Processing and Computer Vision*

Introduction

Internet, telecommunications and digital television made explosive rate of video recordings. With such enormous video data resources there are rising the issues for the efficient browsing, searching and retrieval of digital videos. However, the traditional video indexing method, which uses human beings to manually annotate or tag videos with text keywords, is time-consuming and not efficient.

For video data to be useful, its content must be represented so that it can be stored, queried, and displayed in response to user information needs. The process of shot detection is a fundamental component in automatic video indexing, editing, archiving and browsing [Hanjalic, 2002]. Any approach to indexing and archiving video for retrieval requires parsing the video and extracting key frames to generate an indexed database. A typical video indexing technique is to segment a video sequence into shots and then select representative key-frames [Fernando, 2000]. A shot is defined as an unbroken sequence of frames from a single camera. In a video sequence there can be a number of different types of transitions between shots, such as a cut (abrupt shot change between two frames) or gradual transitions such as fades, dissolves and wipes [Borecsky, 1996]. Here, we focus only on cut detection problem, using a new approach to determine the similarity or dissimilarity of the sequential frames.

The major techniques that have been used for shot boundary detection are pixel differences, statistical differences, histogram comparisons, edge differences, compression differences, and motion vectors [Borecsky, 1996].

Pixel Differences: The easiest way to detect if two frames are significantly different is to count the number of pixels that change in value more than some threshold. This total is compared against a second threshold to determine if a shot boundary has been found. This method is sensitive to camera motion and somewhat slow [Borecsky, 1996].

Statistical Differences: Statistical methods expand on the idea of pixel differences by breaking the images into regions and comparing statistical measures of the pixels in those regions. This method is reasonably tolerant of noise, but is slow due the complexity of the statistical formulas.

Histograms: Histograms are the most common method used to detect shot boundaries. The simplest histogram method computes gray level or color histograms of the two images. If the bin-wise difference between the two histograms exceeds a threshold, a shot boundary is assumed.

Compression Differences: It used differences in the discrete cosine transform (DCT) coefficients of JPEG compressed frames as their measure of frame similarity, thus avoiding the need to decompress the frames.

Edge Tracking: The method aligned consecutive frames to reduce the effects of camera motion and compared the number and position of edges in the edge detected images. The percentage of edges that enter and exit between the two frames is computed. Shot boundaries were detected by looking for large edge change percentages. This is more accurate at detecting cuts than histograms and much less sensitive to motion than chromatic scaling.

Motion Vectors: Method uses motion vectors determined from block matching to detect whether or not a shot was due to zoom or a pan. Because shots with camera motion can be incorrectly classified as gradual transitions, detecting zooms and pans increases the accuracy of a shot boundary detection algorithm.

The most of existing methods for video cut detection use some inter-frame difference metric. In frame pair where this difference is greater than a predefined threshold is deemed to be a shot boundary or cut location. Probably the simplest of these methods is based on pair-wise pixel comparison, and the widely used modification of them is the mean-squared error (MSE). Despite of many advantages of MSE-based methods there are certain problems of image processing technique, where application of MSE measure for two images dissimilarity assessment comes to conflict with the Human Visual System (HVS) perception [Smoliar, 1994] . HVS, in contrast to MSE, is less sensitive to camera or object motion because it is extracting basically the structural and intentional but not pixel-by-pixel information from an image. Last two decades bring many new metrics for images similarity assessment, beginning from [Wang, 2002 - Wang, 2009], which give the results more coherent with HVS perception.

As it is mentioned above the shot boundary detection algorithms mainly uses similarity or dissimilarity measures between consequent frames. The comparison value is called a threshold, and it is one of important elements in the shot change detection algorithm. We can divide threshold setting into two groups: the fixed threshold method and the adaptive threshold method [Zhi, 2005].

The fixed threshold method determines optimal thresholds from repeated experiments. However, they require much experimental iteration and must find other optimal threshold for other video sequences. Most of them iterate adjustment of thresholds until they get the best results. These methods may have long processing time. In general, variation of thresholds is relatively large to use a fixed threshold for all video sequences. Thus, some algorithms for shot detection were improved by analyzing the whole video sequences for setting multiple thresholds instead of a fixed threshold [Cheng, 2002]. These methods may also have long processing time. Thus, it is difficult to apply them to actual applications requiring real-time operations. Meanwhile, the adaptive threshold based segmentation algorithms get sub-optimal threshold according to [Kim, 2009].

This paper introduces a novel video cut detection technique using the similarity measure [Asatryan, 2009] based on comparing the structural properties of images. The approach proposed in [Asatryan, 2010] successfully applied to some problems of image registration even the images are rotated or scaled.

Shot Detection Algorithm

Let's consider a sequence of frames $f_1, f_2, ..., f_i, ...$, where f_1 is the first frame of video sequence or the first frame which follows the previous cut.

The simplest decision rule for cut detection is based on consecutive comparison of contiguous frames using certain similarity measure. When the level of similarity measure exceeds some predefined threshold t_c , then the corresponding frame is considered as a cut frame. Of course, there are other decision rules which use the information of previous frames [Cheng, 2002] or use a few subsequent frames before decision making [Kim, 2009]. However, in any case the quality of a decision rule depends on used similarity measure. In this paper, for simplicity, we choose the first type of cut detection algorithm and compare it with the new measure based on structural properties of an image.

For simplicity we consider the Gray Scale (8 bit) format image $\mathrm{I} = \{\mathrm{I}(\mathrm{m},\mathrm{n})\}$ with $\mathrm{M} \times \mathrm{N}$ sizes, $\mathrm{m} = 0,1,..., \mathrm{M}$; $\mathrm{n} = 0,1,..., \mathrm{N}$.

The standard algorithm of inter-frame comparison by using a similarity measure μ can be presented as follows

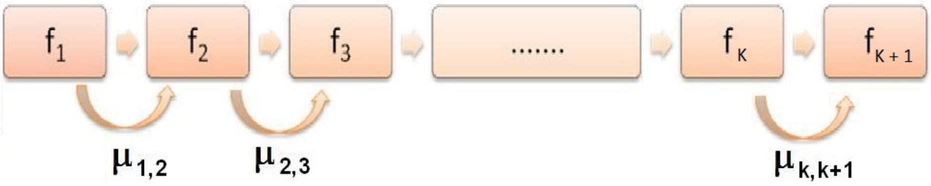


Fig.1 Algorithm of inter-frame comparison

Let consider the frames $f_1, f_2, ..., f_k, f_{k+1}$ and the sequence of corresponding values of the similarity measure $\mu_{1,2}, \mu_{2,3}, ..., \mu_{k,k+1}$. When $\mu_{\mathrm{i,i+1}} \le \mathrm{t_c}$ for $\mathrm{i} = 1,2,..., \mathrm{k}-1$ and $\mu_{\mathrm{k,k+1}} > \mathrm{t_c}$ then point k is assumed as a cut point.

To demonstrate the advantages of shot detection algorithm with using new similarity measure we have to choice some popular, simple and interpretable measure for inter-frame similarity estimation. As it is noted in the Section 1 the simplest measure is MSE-based similarity measure PSNR, which is done according the formula (1) as follows [Fernando, 2000]

$$\mathrm{PSNR} = 10\log_{10}\frac{\max\limits_{\mathrm{m,n}}\left|\mathrm{I}_1(\mathrm{m},\mathrm{n}) - \mathrm{I}_2(\mathrm{m},\mathrm{n})\right|^2}{\mathrm{MSE}^2}, \quad MSE^2 = \frac{1}{MN}\sum_m\sum_n\left[I_1(m,n) - I_2(m,n)\right]^2 \tag{1}$$

New measure is based on structural properties of an image, therefore it can be expected that the new measure will provide the detection of content changes more adequately than any point-by-point measure.

The mentioned measure is described below. We consider a model of image structure based on the set of edges which are determined by the gradient field of the image [Asatryan, 2009]. Here we rest upon the fact that the HVS uses the edge information for understanding and analyzing the structure of an image [Wang, 2002], [Wang, 2009]. It is also very important that the edges are invariant to image rotation, scaling and other transformations, so they provide more adequate extraction of structural information from any image [Wang, 2004]. In [Asatryan, 2009] a measure is proposed for images similarity assessment based on using the gradient magnitude distribution.

Let $\|G_H(m, n)\|$ and $\|G_V(m, n)\|$ at a point (m, n) of an image be the horizontal and vertical gradients, determined by one of known gradient methods, and the matrix of gradient magnitude $\|M(m, n)\|$, where

$$M(m,n) = \sqrt{G_H^2(m,n) + G_V^2(m,n)} \tag{2}$$

Following to [Asatryan, 2009] we suppose that the gradient magnitude (2) is a random variable with Weibull distribution density

$$f(x;\eta,\sigma) = \frac{\eta}{\sigma}\left(\frac{x}{\sigma}\right)^{\eta-1}\exp\left[-\left(\frac{x}{\sigma}\right)^{\eta}\right], x \geq 0, \eta > 0, \sigma > 0 \tag{3}$$

As a measure of structural similarity of two images with probability distribution functions of gradient magnitude $f_1(x;\eta_1,\sigma_1)$ and $f_2(x;\eta_2,\sigma_2)$ accordingly, we take

$$W^2 = \frac{\min(\eta_1,\eta_2)\min(\sigma_1,\sigma_2)}{\max(\eta_1,\eta_2)\max(\sigma_1,\sigma_2)},\ 0 < \mathrm{W}^2 \leq 1, \tag{4}$$

where the corresponding parameters are represented as statistical estimations gotten from the corresponding samples of gradient magnitude.

As it has been shown in [Asatryan, 2009 - Asatryan, 2012] the measure (4) has certain advantages against widely used other methods, which use image pixels point-by-point comparison method.

Results of Experiments

We tested our method with several video clips of different themes and varying nature. The videos fluctuate widely in content and length. But in this section we include analyzes and results on one exact video. The purpose of this section of our work is to illustrate how our algorithm works and compare the results with widely known PSNR method mentioned above.

We take manually detected positions of the shot boundaries as the ground truth, defining in that way the number of missed detections and false detections.

The frame sequence fragment of video under test is demonstrated in Fig.2.

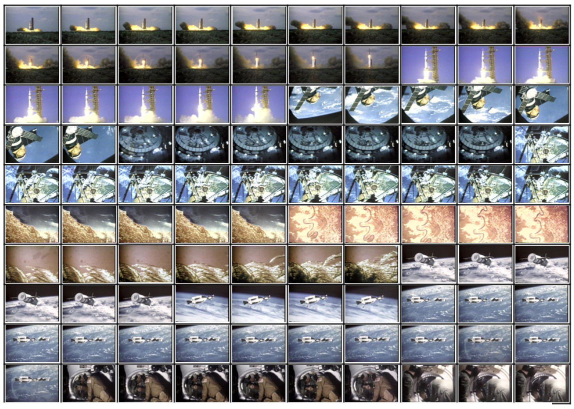


Fig.2 Consecutive frames from video sequences presenting abrupt cuts

The video was divided into 130 frames and contains 9 cuts. The values of PSNR and W^2 have been calculated for all adjacent frames. The experimental results are shown in Fig.3 and Fig.4. From graphical representation it is clearly visible that the threshold value for decision making regarding presence of cuts is of 0.8. But this value can vary for other type and content videos. Our experiments show that acceptable threshold t_c for W^2 vary between 0.6 and 0.8.

It can be noted that a reasonable algorithm for cut detection using PSNR curve can be based on choosing some extremal values of PSNR instead of using a threshold. However, we can propose a statistical model for determination of the threshold for cut detection by using PSNR. The model is based on assumption that the adjacent frames consists of pixels with gradient magnitudes, which are samples from independently and normally distributed random variable. Let Δ be the dynamic range and σ^2 be the variance of the magnitudes related to a shot. Then $\Delta = 6\sigma$, $\mathrm{MSE}^2 \approx 2\sigma^2$, $\mathrm{PSNR} = 10\log\frac{\Delta^2}{2\mathrm{MSE}^2} \approx 12.5\,\mathrm{dB}$, so we can put $t_c = 12.5$.

The values of specified thresholds are shown in Fig. 3 and Fig.4. Solid circles are real cuts; gray circles are false cuts and white circles - cuts that have not been detected. One can see that our proposed method correctly detect all the cuts, while PSNR gives missed hits and false hits.

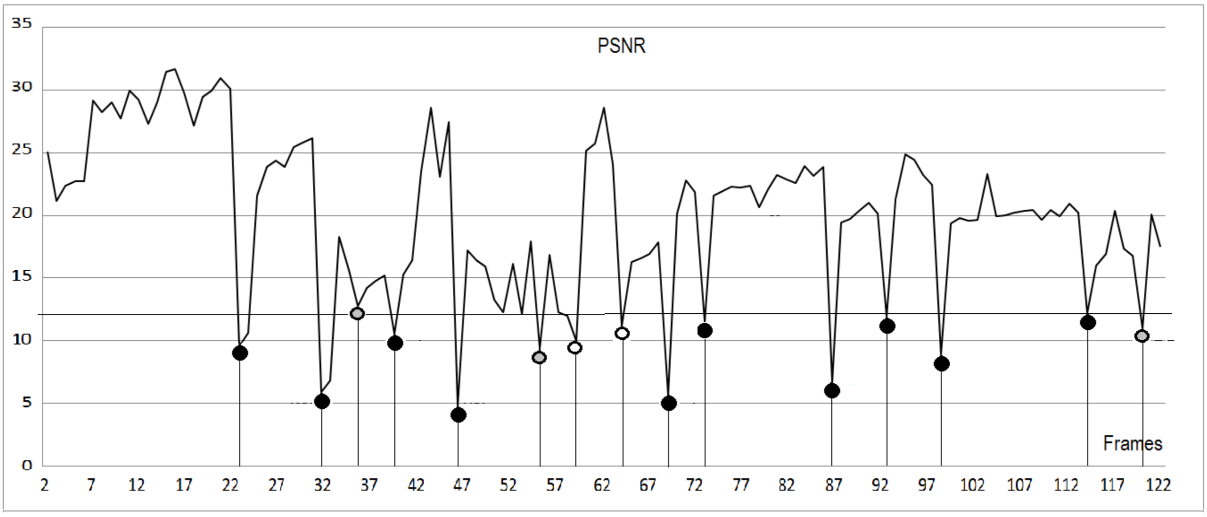


Fig.3 Results of cut detection by using PSNR

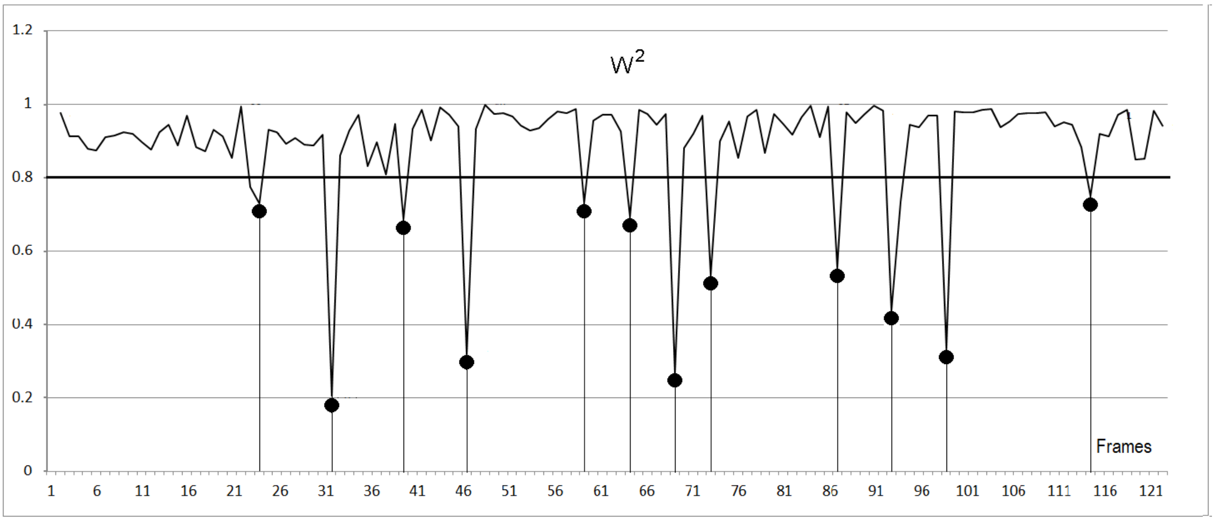


Fig.4 Results of cut detection by using W^2

As it was already mentioned to compare similarity of two frames with our method, we need to estimate two parameters of corresponding Weibull distributions. It is interesting to investigate the behavior of content of the frames within each frame of a shot by graphical analyzing the (η,σ) - scatter.

Fig.5 shows that corresponding shot did not change actively, while Fig. 6 shows that the sense was dynamically changed from the beginning to end. This kind of analyzes may help to analyze some physical process, which is controlled by video camera.

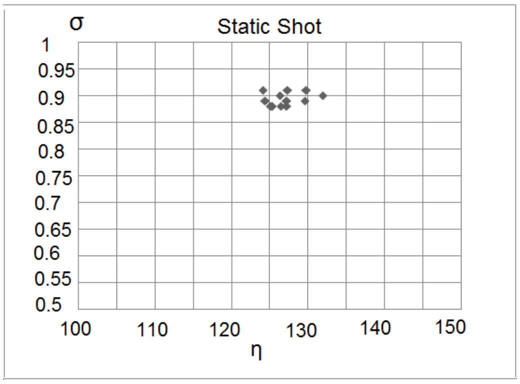


Fig.5 Distribution of σ and η parameters for static shot

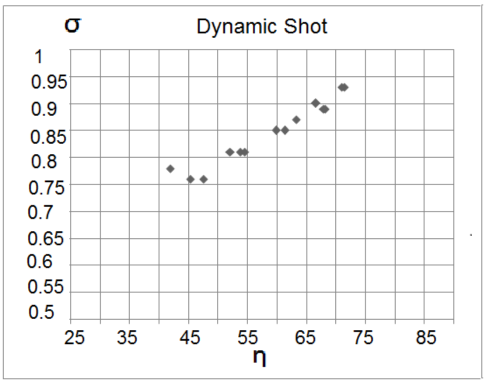


Fig.6 Distribution of σ and η parameters for dynamic shot

Conclusion

Shot boundary (cut) detection is the main and important step of searching and browsing the digital video. In this paper we propose a novel method for shot detection based on Weibull distribution model. This is a new technique for shot detection not based on existing methods or their combinations. It is using structural properties of the image and gives results which are closer to human visual system perception. Earlier the method was successfully compared to some of widely using methods reported in literature. The experimental results show the effectiveness of proposed method of cut detection, in comparison with method based on widely used mean-square measure. Convenient algorithm used for images similarity measurement, which is using only two parameters of an image, may be easily and effectively used to do more deeper analyzes inside and also between different shots, which can give general information about type of whole video or separate segments on it. We believe that this method can be successfully applied to detection of gradual transitions, such as fade in/ fade out, dissolves and etc., and our farther works will be devoted to them.

Acknowledgment

This work was supported by the State Committee of Science of Armenia (13A-1b25). http://scs.am/

Bibliography

[Asatryan, 2009] D. Asatryan and K. Egiazarian. Quality Assessment Measure Based on Image Structural Properties. International Workshop on Local and Non-Local Approximation in Image Processing, Finland, Helsinki, pp. 70-73, 2009.

[Asatryan, 2010] D. Asatryan, K. Egiazarian and V. Kurkchiyan. Orientation Estimation with Applications to Image Analysis and Registration. International Journal "Information Theories and Applications", vol. 17, no. 4, pp. 303-311, 2010.

[Asatryan, 2012] D. Asatryan, V. Kurkchiyan and M. Bagramyan. A Method for Quality Assessment of Image Resizing Algorithms. Mathematical Problems of Computer Science, vol. 36, pp. 128-132, 2012.

[Borecsky, 1996] J. S. Borecsky and L. A. Rowe. Comparison of video shot boundary detection techniques. Journal of Electronic Imaging, vol.5, no.2, pp.122–128, 1996.

[Cheng, 2002] Y. Cheng, X. Yang, and D. Xu. A method for shot boundary detection with automatic threshold. Proceedings of IEEE TENCON, vol. 1, pp. 582-585, 2002.

[Fernando, 2000] W.A.C. Fernando, C.N. Canagarajah, and D.R. Bull. A unified approach to scene change detection in uncompressed and compressed video. IEEE Transactions on Consumer Electronics, vol. 46, no. 3, pp. 769–779, 2000.

[Hanjalic, 2002] A. Hanjalic. Shot-Boundary Detection: Unraveled and Resolved? IEEE Transactions on Circuits and Systems for Video Technology, vol. 12, no. 2, pp. 90-105, 2002.

[Kim, 2009] W.Kim, K. Moon and J.Kim. An automatic shot change detection algorithm using weighting variance and histogram variation. 11th International Conference on Advanced Communication Technology, vol. 2, pp.1282-1285, 2009.

[Liu, 2010] W. Liu. An Effective Method For Abrupt Scene Change Detection. 2010 Sixth International Conference on Natural Computation , vol. 2, pp. 850-852, 2010

[Smoliar, 1994] S.W. Smoliar and H.J. Zhang. Content-based video indexing and retrieval. IEEE Multimedia, vol. 1, no. 2, pp. 62-72, 1994.

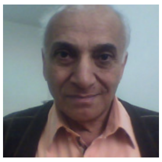
[Wang, 2002] Z. Wang, A.C. Bovik. A universal image quality index. IEEE Signal Processing Letters, vol. 9, no. 3, pp. 81-84, 2002.

[Wang, 2004] Z. Wang, A.C. Bovik, H.R. Seikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.

[Wang, 2009] Z. Wang and A.C. Bovik. Mean Squared Error: Love It or Leave It?. IEEE Signal Processing Magazine, vol. 26, no.1, pp. 98-117, 2009.

[Zhi, 2005] M. Zhi and A. N. Cai. Shot change detection with adaptive thresholds. IEEE International Workshop on VLSI Design and Video Technology, pp. 147-149, 2005.

Authors' Information



David Asatryan *– Professor, Head of group of the Institute for Informatics and Automation Problems of NAS Armenia, 1, P.Sevaki Str., 0014, Yerevan, Armenia; e-mail: dasat@ipia.sci.am*

Major Fields of Scientific Research: Digital signal and image processing.

Manuk Zakaryan *– Ph.D. student at Russian – Armenian (Slavonic) University, Software Developer at EGS Armenia; e-mail: zakaryanmanuk@yahoo.com*

Major Fields of Scientific Research: Digital signal and image processing, Software developing.

MPLS NETWORK STRUCTURAL SYNTHESIS WITH APPLICATION OF MODIFIED GENETIC ALGORITHM

Yuriy Zaychenko, Helen Zaychenko

***Abstract**: The problem of MPLS networks structural synthesis is considered. The various modifications of Genetic Algorithms (GA) are investigated for this problem solution which differs in implementation of crossover and mutation procedures. The most adequate version of GA was elaborated and its investigations were carried out. The application to suggested GA for real problem of topological optimization of MPLS network is presented.*

Keywords: MPLS network, Genetic algorithms, structure synthesis

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings National Information Technologies NIT 2013*

Introduction

One of the most perspective network technologies is the MPLS (Multiple Protocol Label Switching). This technology uses different classes of service (CoS) for transmission of various information types (data, audio and video) and assures the corresponding quality of service (QoS) for corresponding CoS. QoS indices include: Packets Delay Time (PDT), Packets delay Variance (PDV) and Packets Loss Ratio (PLR). The important feature of MPLS is that it's easily integrates with Internet protocol stack TCP/IP. The appearance of MPLS technology required the necessity of optimal design methods of MPLS networks which would take into account the specificity of MPLS technology.

The main goal of this paper is to consider the problem of MPLS networks structure optimization, to elaborate and investigate the adequate method for its solution (modified genetic algorithm of structure synthesis under uncertainty).

Problem statement

Let's consider the problem of computer network structure optimization with MPLS technology [Зайченко Е., 2008].

A set of networks nodes (so-called LSR (Label Switching Routers) $X = \{x_j\}$ $j = \overline{1,n}$ - is given, their locations over territory, a set of channels $D = \{d_1, d_2, ..., d_k\}$ and their costs per unit length $C = \{c_1, c_2, ..., c_k\}$, CoS (Class of Service) are defined, matrices of incoming demands are known for each Class of Service $H(k) = \|h_{ij}(k)\|$ $i, j = \overline{1,n}$; $k = 1, 2, ..., K$, where $h_{ij}(k)$ is the k-th class flow rate which is to be transmitted from node I to node j (Mbits/sec).

Besides, desired values of QoS are set as the constraint on Mean Packet Time Delay for each CoS k $T_{зад,k}$, $k = \overline{1,K}$

It's required to determine MPLS network structure as a set of channels, to find their capacities $\{\mu_{rs}\}$ and flows distribution for each class $F(k)=[f_{rs}(k)]$ so that to enable to transmit all the classes demands $H(k)$ with mean packets time delay T_{cp} not exceeding given values $T_{зад,к}$ and Packets Loss ratio (PLR) not exceeding limitation on this value PLR$_k$, and total network cost should be minimal [Зайченко Е., 2008]. Let's construct a math model of this problem. It's necessary to find such network structure E, for which:

$$\min_{E\{\mu_{rs}\}} C_{\Sigma}(M)=\sum_{(r,s)\in E} C_{rs}(\{\mu_{rs}\}) \tag{1}$$

under constraints

$$T_{cp}(\{\mu_{rs}\};\{f_{rs}\})\le T_{зад,k},\ k=\overline{1,K} \tag{2}$$

$$f_{rs}<\mu_{rs} \text{ for all channels } (r,s) \tag{3}$$

$$\mu_{rs}\in D \tag{4}$$

$$PLR_k(\{\mu_{rs}\};\{f_{rs}\})\le PLR_{k\,çàä} \tag{5}$$

where $PLR_k(\{\mu_{rs}\};\{f_{rs}\})$ is a rate of lost packets for k-th class of flow $PLR_{k\,çàä}$ is a given constraint on this value.

This problem belongs to a class of so-called NP- difficult optimization problems. For its solution modified genetic algorithm is elaborated and investigated [Зайченко Е., 2008].

Modified Genetic Algorithm for Network structure Synthesis

As it's well known GA consists of three procedures: crossover, mutations and selection [Згуровский, 2013]. But in this problem crossover and mutation procedures are made adjustable, so that strategic parameters are adapted in the process of algorithm run.

Define a channels matrix $K=\|k_{ij}\|$, where $k_{ij}=\begin{cases}1,\exists(i-j)\\0,\neg\exists(i-j)\end{cases}$, for each structure (network). Then generate initial population of n different structures in a given structures class – multi-connected structures with connectivity factor 2. For synthesis we'll use semi-uniform crossover. Parent structures for crossover we select randomly with probability inverse-proportional to a cost function

C_{Σ} (Ei(k)), to each parent matrix corresponds $K^i, i=1,2$. In the process of semi-uniform crossover each offspring gets one-half of parent's genes.

Crossover mask is represented as a matrix of the following form $M=\|m_{ij}\|$, где $m_{ij}=\begin{cases}0,p\ge p0\\1,p<p0\end{cases}$,

$p0=0.5$ - is a parameter, and a $p\in[0,1]$ is random.

Formally this process of crossover may be written as follows $E(k)'=\|e(k)'_{ij}\|=\begin{cases}(i-j)^1,k_{ij}^{\ 1}=1,m_{ij}=0\\(i-j)^2,k_{ij}^{\ 2}=1,m_{ij}=1\end{cases}$.

During crossover we generate only one offspring due to goal of maximization of algorithm productivity. In case of getting isolated subgraphs connect them with direct channels to a root. Further for obtained offspring – structure E(k)' solve the problems channels capacities assignment and flows distribution (problem CA-FD) and Fiona new channels capacities and flows distributions for all classes of service [Згуровский, 2013]. Then after comparison cost value for offspring and parents the we decide whether to introduce or not offspring structure E(k)' in a sequence of locally efficient structures 9 current population) П.

After crossover it's necessary to define mutation procedure. Note that basic algorithm suggested in [Зайченко Е., 2008] used the unconditional mutation procedure. Mutations consist of deleting or introducing some new channels in network structure. In the process of algorithm improvement the following schemes of mutations probability changes were suggested:

- Deterministic and adaptive change.

In deterministic version mutation probability is defined with application of time-variable function. As an example we change probability as follows

$$\sigma(t) = 1 - ct/T \, .$$

Note as time passes the probability decreases.

Note that main properties of such approach are:

1) Mutation probability change does not depend on the success of its application in genetic search;
2) A designer fully controls the probability changes due to certain formula;
3) Mutation probability change is fully predictable.

For the implementation of adaptive approach of mutation probability change we use Rechenberg rule [Згуровский, 2013]. In this case the rule for the mutation probability change will be as follows

$$\sigma(t) = \begin{cases} \sigma(t-1)/\lambda, \varphi(t-1) > 1/5 \\ \sigma(t-1)\lambda, \varphi(t-1) < 1/5 \\ \sigma(t-1), \varphi(t-1) = 1/5 \end{cases} ,$$

where $\varphi(t)$ is a per cent of good mutations, and $\lambda = 1.1$ is a learning factor.

Note that the main properties of this approach are the following:

1) Mutation probability change depends on the successfulness of its application in the process of genetic search;
2) Mutation probability change is non-predictable.

- Self-adjustable mutation.

Self-adjustable mutation may be implemented on the level of chromosome (network structures) and on the level of genes (channels) for should be given) mutation probability change rule (law) $\sigma(t)$, and then $\sigma(t)$ is coded in chromosome as:

$$\{Ek, \sigma(t)\} \text{ or } \{Ek = \|ek_{ij}\|, \|\sigma_{ij}(t)\|\} \, .$$

But this approach leads to crucial decrease in algorithm productivity and for our problem is not good alternative. Note that main properties of such approach are the following:

1) Mutation probability change is a result of natural choice;

2) The designer practically doesn't control this process;
3) Mutation probability change is non-predictable.

As a contra version to scheme with unconditional crossover and dynamic mutation the scheme with unconditional mutation and dynamic crossover was implemented.

In the process of algorithm improvement with unconditional mutation the following schemes of crossover probability change were investigated:

– Deterministic;

The implementation of deterministic scheme is based on hypotheses that on various search stages crossover may be more or less significant/ that's why as a function of crossover probability change is reasonable to choose non-monotonic function like such:

$$\sigma(t) = |\sin(t)|, \text{ где } 0 \le t \le T .$$

– Adaptive.

Define adaptive crossover as an operator probability of which decreases if a population is homogenous and increases if the population is sufficiently heterogenous one. As a measure of homogeneity/ heterogeneity take

$$C_\Delta = \max(C_\Sigma(E_i(k)) - C_\Sigma(E_j(k)),\ i \in [1,...,n], j \in [1,...,n], i \ne j ,$$

where $n = 3$ is a population size.

It's reasonable to suppose that in case of very like species in population crossover will be inefficient and vice versa. Thus in adaptive approach the rule of crossover probability change takes the form

$$\sigma(t) = \begin{cases} \sigma(t-1)\lambda, C_\Delta > C^* \\ \sigma(t-1)/\lambda, C_\Delta < C^* \\ \sigma(t-1), C_\Delta = C^* \end{cases},$$

where C^* is a threshold value, and $\lambda = 1.1$ is a learning factor.

Self-adjusting crossover. The implementation of self-adjusting crossover is not reasonable due to substantial decrease of algorithm productivity like self-adjusting mutation.

Experimental Investigations

The experimental investigations of various modifications of GA were carried out in which the efficiency of different variants of crossover and mutation procedures were explored and compared. The problem to be solved is a National Ukrainian MPLS network design. In process of experiments were varied sets of channels capacities, costs of unit channel length, demands matrices, given QoS values (PDT, PLR).

After series of experiments were carried out the following results were obtained:

1) A combination of unconditional crossover and dynamic deterministic mutation - this implementation proved to be one of the most successful - the increase of productivity up to 15%). This experiment confirmed the hypothesis that mutations play the essential role at the initial phase of search while at final stage the most efficient is to use crossover for finding optimal (Quasi-optimal) solution on the base of earlier obtained solutions.

2) The combination of unconditional crossover and dynamic adaptive mutation; this combination did not allow to reach stable decrease of algorithm run time.
3) The combination of unconditional crossover and dynamic self-adjusting mutation – this implementation is unreasonable as it essentially complicates the process of genetic search and leads to decrease of algorithm productivity – combination of unconditional mutation and:
4) Dynamic deterministic crossover: this implementation did not allow obtaining the stable increase in productivity.
5) Dynamic adaptive crossover: this implementation proved to be the most successful - the productivity increase is 20 - 22%.

By this the hypothesis that crossover operator has some positive properties which mutation operator does not have been confirmed. But its worth to note application of crossover operator is efficient only if the species in population are quite different.

As the illustration of experiments on Fig.1 the initial structure of MPLS network in Ukraine is presented while on Fig. 2 one of the optimal structures is presented obtained by modified genetic algorithm which uses the combination of dynamic adaptive crossover and unconditional mutation.

These results were obtained with test data close to real data. Note that after the application of modified GA total cut in network costs for optimized network structure comprised:

14250 thousand$ - 10023thousand. $ = 4227 thousand $,

that is by 30% less than the costs of initial network structure. It's very important that algorithm productivity was increased: this result was obtained with 22% less time than by basic GA.

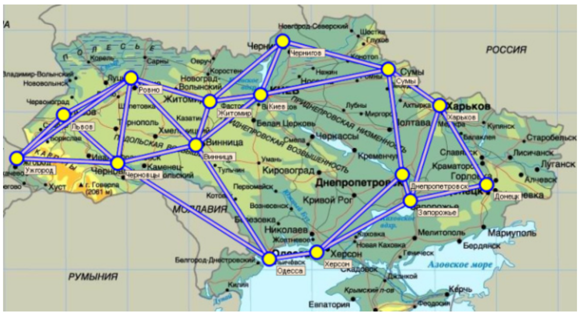


Figure 1 Initial MPLS network structure

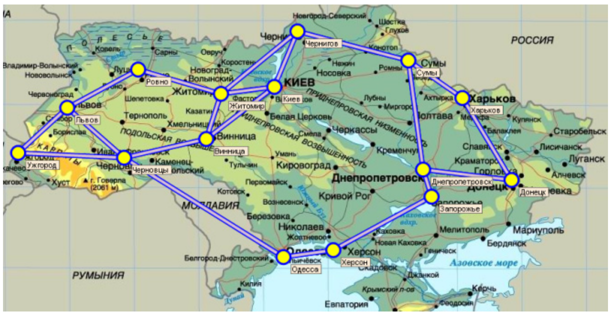


Figure 2 Optimized MPLS network structure

Conclusion

The problem of MPLS computer networks structure design is considered.

Different genetic algorithms for its solution with various modifications of crossover and mutation procedures were investigated for its solution.

The most efficient GA for MPLS structure synthesis was determined and its application for Ukrainian MPLS network topological design is presented.

Bibliography

[Зайченко Е., 2008] Е.Ю. Зайченко, Ю.П. Зайченко. Сети с технологией MPLS: моделирование, анализ и оптимизация. –К.: НТУУ «КПИ», 2008ю -240 с.

[Згуровский, 2013] Згуровский М.З., Зайченко Ю.П. Основы вычислительного интеллекта. –К.: Изд. «Наукова Думка», 2013.- 406 с.

Acknowledgement

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

Authors' Information

Zaychenko Yuriy *– Dr. of Science, professor Institute for Applied System Analysis NTUU"KPI", P.O. Box: 03056, Kiev-56, Peremogy ave. 37, Ukraine; e-mail: baskervil@voliacable.com*

Major Fields of Scientific Research: Decision- making Theory under uncertainty, Computational Intelligence C computer networks modeling and design

Zaychenko Helen *- Dr. of Science, professor Institute for Applied System Analysis NTUU"KPI", P.O. Box: 03056, Kiev-56,Peremogy ave. 37, Ukraine; e-mail: syncmaster@bigmir.net*

Major Fields of Scientific Research^ Computer networks optimal design and modeling, Operations Research

FORMAL VERIFICATION OF THE SEQUENCE DIAGRAM

Vitaliy Lytvynov, Irina Bogdan

Abstract: *The article describes three different approaches to verification of one of the most frequently used UML-diagrams – the sequence diagram. It indicates that these methods allow estimating its correctness only in certain aspects, and complex application of these approaches is the most effective way.*

Keywords: *verification, digital machine, driver, UML-diagram, record, correctness, sequence diagram.*

Introduction

One of the characteristic features of systems of various nature and destination is the mutual interaction of separate elements that these systems are made of. It means that various constituent elements of the systems do not exist in isolation, but have the influence on each other, and that is the fact that distinguishes the system as a unique formation from the simple complex of elements. [1]. The appropriate interaction diagrams are used for modeling the interaction of objects in UML language. The interaction of objects is frequently considered in time, and then the sequence diagram is used to represent the temporal characteristics of the transmission and the receiving of the messages between the objects.

The message sequence diagram shows the exchange of messages between the objects, the object creation and the response to the messages. The sequence diagrams are used at the analysis stage to illustrate the process in the object domain, so how the tasks are solved in general case due to the interaction of various objects within the script [2].

The message sequence diagram is used at the design stage to describe the algorithms, that is which objects are involved in data processing in order to find a particular aspect of the decision and how these objects interact with each other in order to have an impact on this process. It underlines the importance of the verification of sequence diagram. The basic meaning of the verification of sequence diagram is to make sure that there is a correct exchange of the messages between the objects, which classes have already been, verified previously [3].

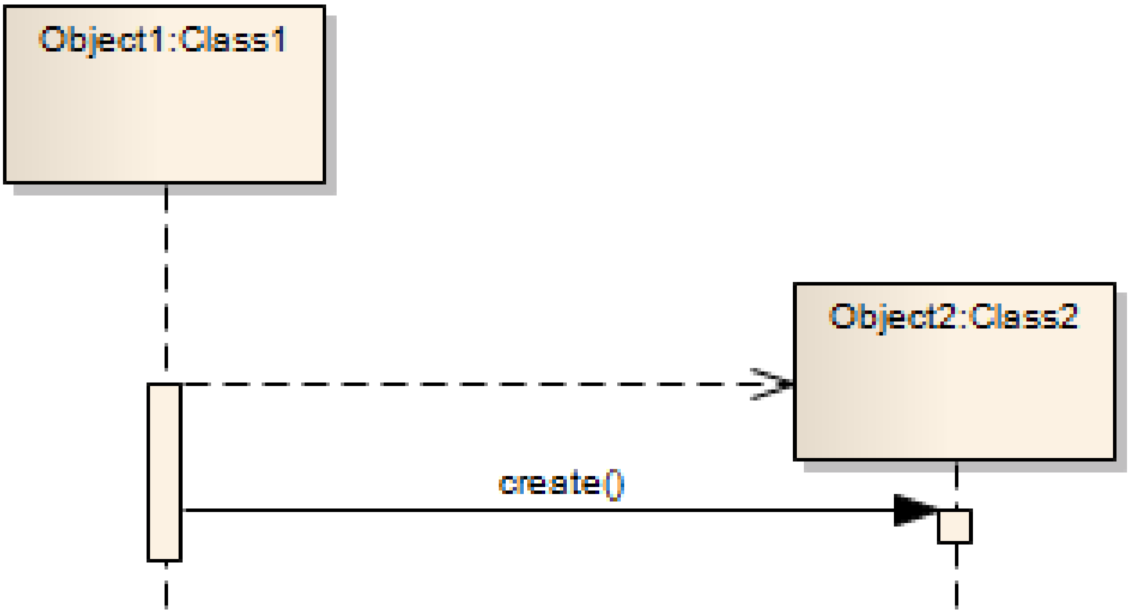
At the moment, there are several different approaches to the verification of sequence diagrams. Most of them allow assessing the correctness of this diagram from different points of view.

The construction of the driver

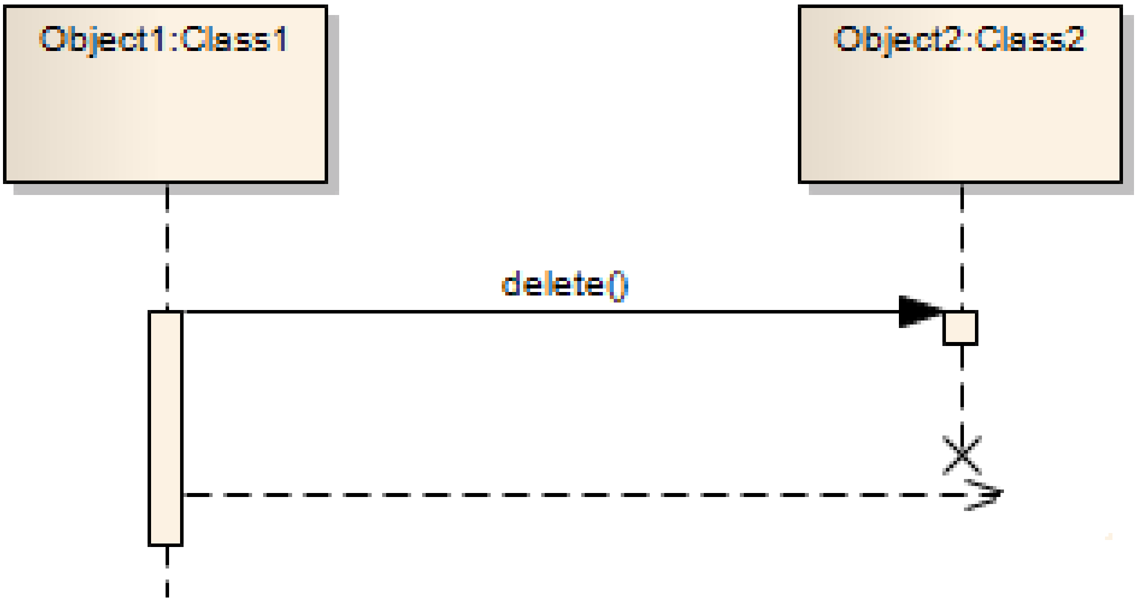
One of the simplest approaches is the one that is associated with the construction of the driver. This approach allows assessing the correctness of the creation of the sequence diagram formally and identifying the most typical mistakes.

As some objects in the sequence diagram will exist permanently, and some - temporary (only during the performance of the required actions), so, first of all, you should assure that the destruction or creation of the objects, that are created for the time of the performance of their actions, is done correctly and explicit messages are provided for them.

Picture 1 shows an example of sending a message to an object before its creation, that is not correct, and Picture 2 shows an example of sending a message to an object after its destruction, that is also not correct.

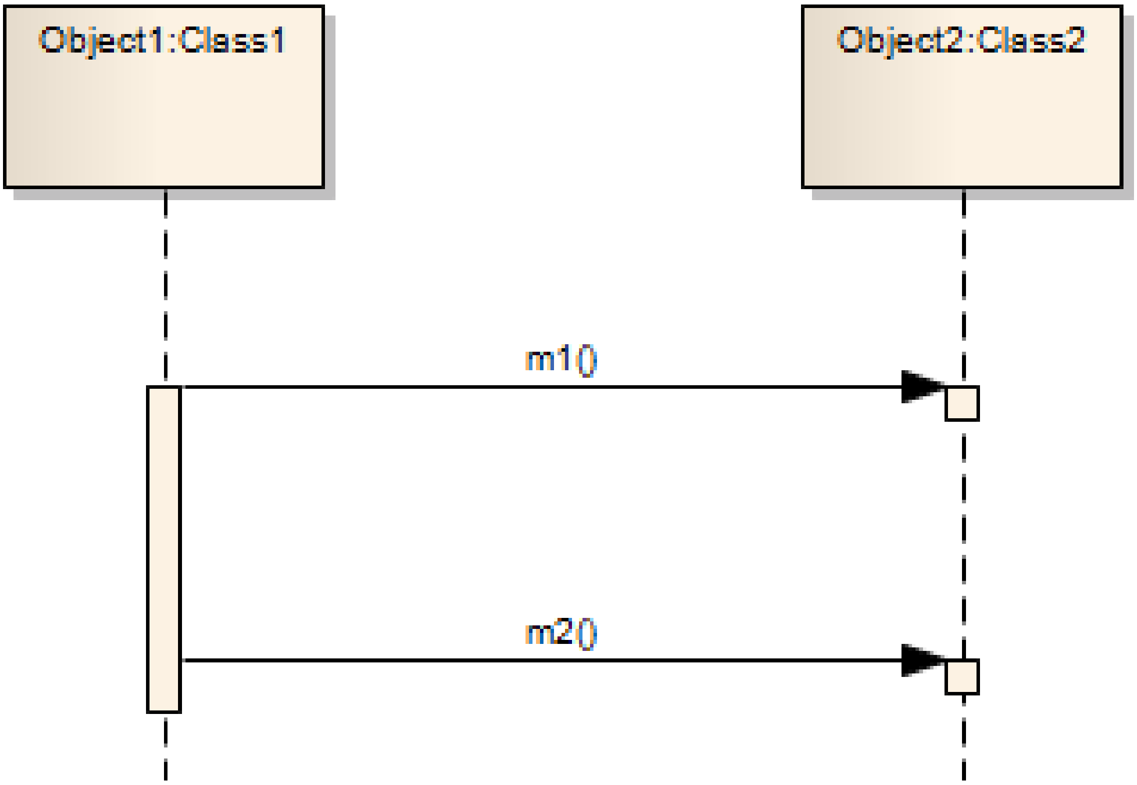


Picture 1. An example of typical error: sending a message to an object before its creation



Picture 2. An example of typical error: sending a message to an object after its destruction

The next step is to perform the actual verification of interaction between the objects. During the verification of interaction under the conditions of contact approach, the attention should be paid to checking, whether the preconditions of methods of the object-recipient are performed by the object-sender. The verification does not take place if these preconditions are violated. The meaning of such checking is to determine whether the object-sender executes the testing of the preconditions of the object-recipient, before posting an unacceptable from the very beginning message. At the same time, it is checked whether the object-sender stops its activity correctly, possibly, generating an exception. So, for example, according to the definition of synchronous message, the actions of the sender are blocked until it receives the reply. Picture 3 shows an example, according to which the actions of the sender are not blocked after the sending of synchronous message before getting the response, which is not correct.

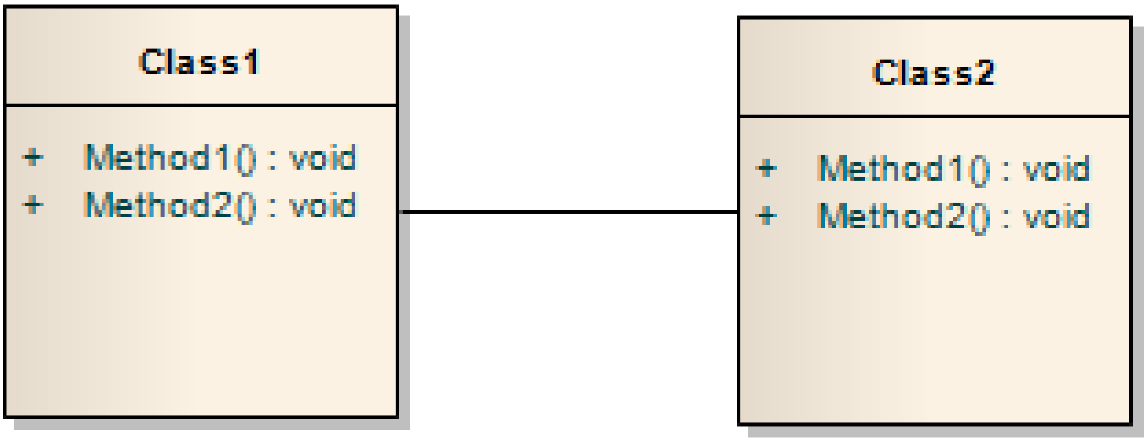


Picture 3. An example of typical error: the actions of the sender are not blocked after sending of synchronous message before getting the response.

Method of protocols

Verification of the sequence diagrams can also be operated with the method of protocols [2]. As some object starts interaction with other objects the number of messages it gets increases. All these messages are regulated into an intended sequence. Verification on protocol controls fitting of installation into this sequence. Different protocols, that one or another object is involved in, can be inferred out of preconditions and post-conditions which institutionalize the performance of alone operations declared in the class of this object. Identifying sequence of method calls, in which the methods whose post-conditions satisfy the preconditions of the next method are aggregated, build a protocol. Such sequences are picked out on the diagram of classes by analysis of the methods of every specific class. In the case of this approach every alone protocol contains methods of alone class.

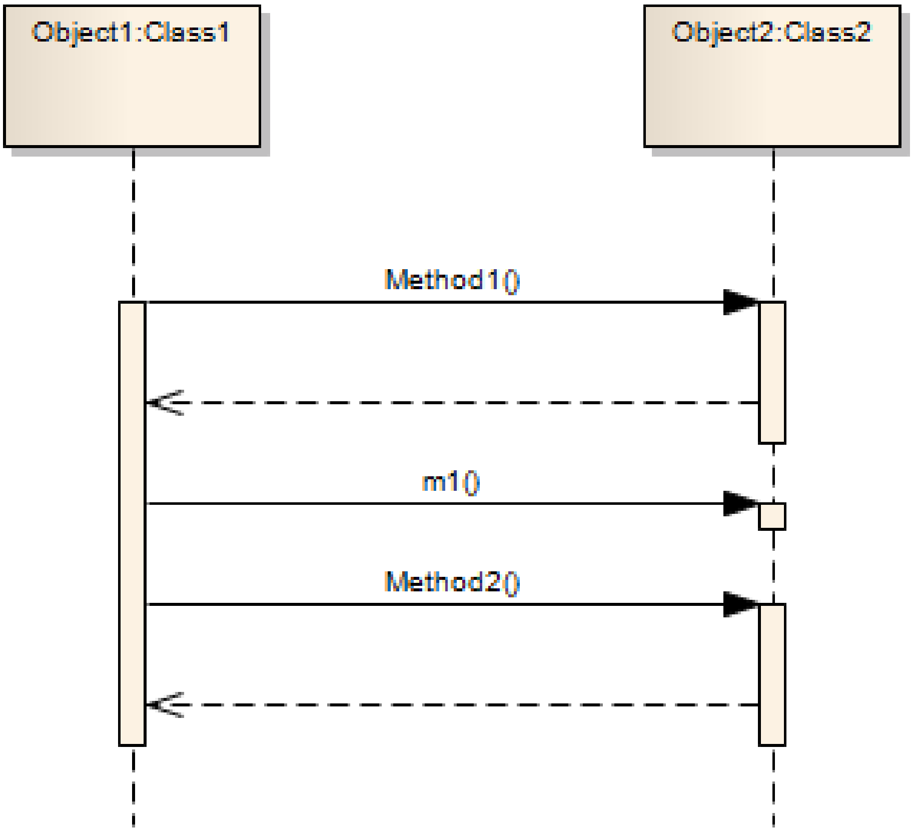
Basically this method of verification is a special form of verification of a life cycle. Every protocol builds a life cycle of verified objects of classes in combination with other classes.



Picture 4. Straightforward diagram of classes

Protocol for the class Class2
-Method1()
-Method2()

Picture 5. Protocol for the class Class2



Picture 6. Sequence diagram of interaction for the object of class Class2

In picture 4 we are presented an example of straightforward diagram of classes which consists of two classes each of which has two methods. In picture 5 protocol for the class Class2 is presented, and in picture 6 a diagram of the sequence of interaction for the object of class Class2 is presented. According to this diagram the protocol for class Class2 must contain Method1(), m1() and Method2(), but on the assumption of picture 5 m1() does not belong to protocol for class Class2, as a result this sequence diagram is not correct.

Method based on the construction of an abstract digital automat

Another approach to verification of the sequence diagrams is based on the introduction of the sequence diagrams as an abstract digital automaton.

Abstract theory of automatons, digressing from the structure of the automaton and taking no interest in the manner of its construction, studies the way of behavior of the automaton towards the external environment.

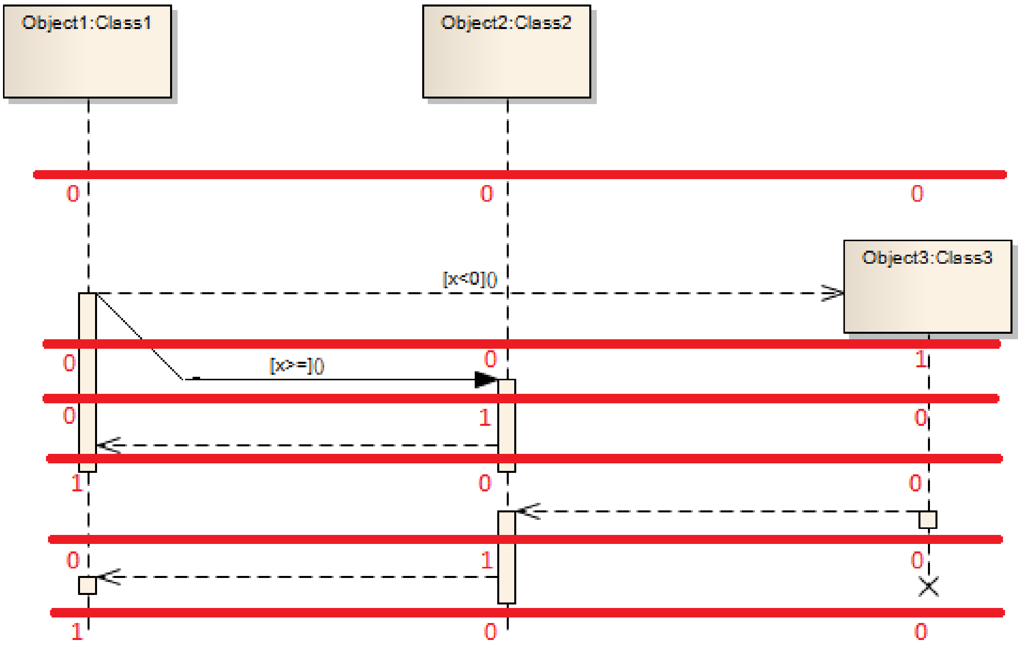
Abstract digital automaton A is defined by a total of five objects {X, S, Y, φ, λ}, where X = $\{x_i\}$, i$\in \overline{1,m}$ - is a set of input signals of automaton A (input alphabet of automaton A); S = $\{s_j\}$, j$\in \overline{1,n}$ - is a set of states of automaton A (states alphabet of automaton A); Y = $\{y_k\}$, k$\in \overline{1,l}$ - is a set of output signals of automaton A (output alphabet of automaton A); φ - is a function of transition of automaton A which indicates the display (X × S)$\rightarrow$ S, i.e. assigning to any pair of elements of a Cartesian product of sets (X × S) element of set S; λ - function of outputs of automaton A which either indicates display (X × S)$\rightarrow$ Y, or display S $\rightarrow$ Y [4].

In this case one automaton describes the sequence of interaction of the objects in one diagram of sequence, where the state means a total of states of all the objects present in the given diagram. Activeness or passiveness of one or another object can be defined (1 or 0) depending on its involvement in the interaction at a certain moment of time.

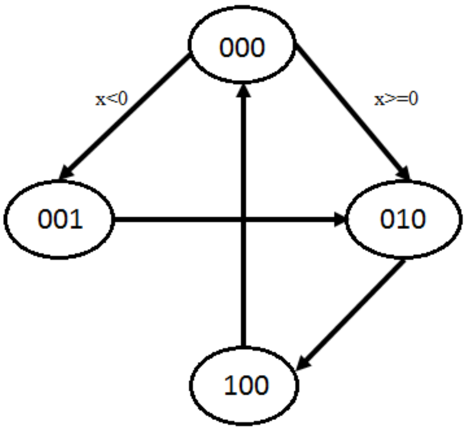
Thus, a set of states of automaton S provides a set of total states of every object shown in the diagram of sequences represented in the binary system. Both sets of input and output signals of automaton X and Y consequently compile a set of messages in the diagram of sequence which make the object either active or passive.

In picture 7 we can see an example of a diagram of the sequence of interaction of three objects with the display of their states in different time intervals.

On graphical method of assigning, as any abstract automaton, the given automaton is represented as a direct graph where the states of the automaton are shown as the points of the graph and the transitions between states as arcs between the corresponding points. Transitions can be both conditional and unconditional. At the moments of time when the objects in the sequence diagram are not yet created or already deleted, they are considered as not active on default (0). Thus, the abstract automaton for the sequence diagram from picture 7 is presented in picture 8.



Picture 7. Sequence diagram of interaction of the objects with the display of their states



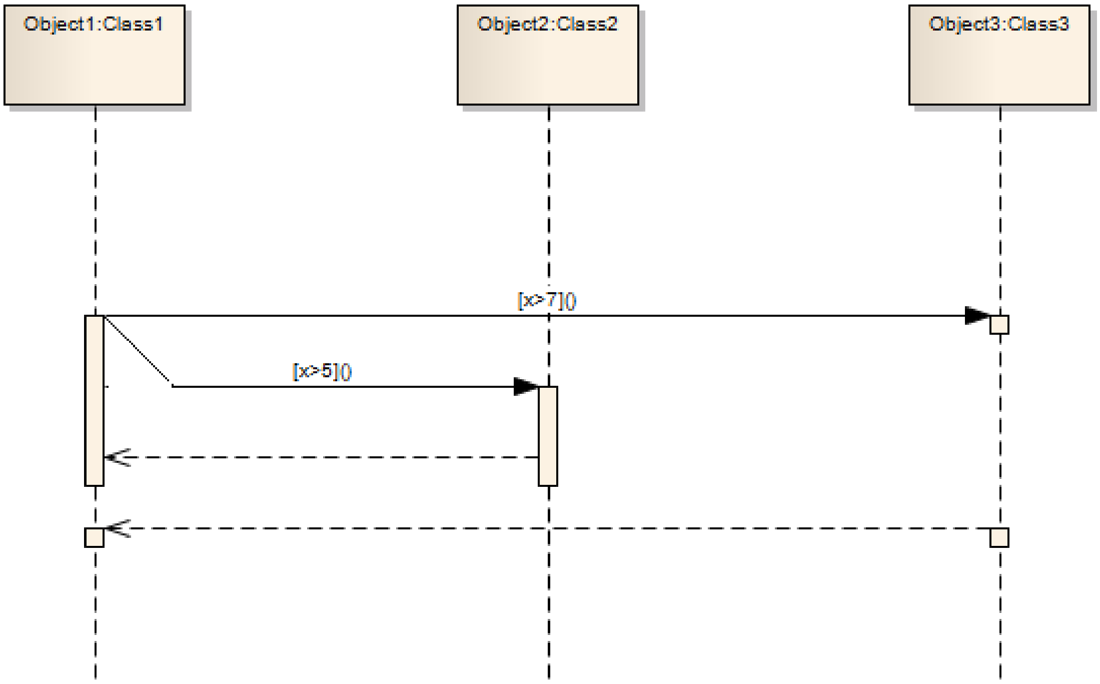
Picture 8. Abstract automaton for the sequence diagram from picture 7

As this approach is based on the introduction of the sequence diagram as digital automaton so the requirements preferred are the same as to any digital automaton:

- At every moment of time the object, which is introduced as a set of states of every object of the sequence diagram, can be in one and only in one of its states. Herewith, the object can be in a certain state as long as it is necessary if no event objects happen that is until new messages come;
- The number of states of an object must be definitely finite (in UML language consideration is given only to finite state automatons) and all of them must be specified in explicit manner;
- An automaton cannot contain isolated states and transitions. This requirement means that for every state, except for the source state, previous state must be defined. Every transition must connect two states of the automaton. Transition from the state into itself is permitted. This kind of transition is also called “self loop”;

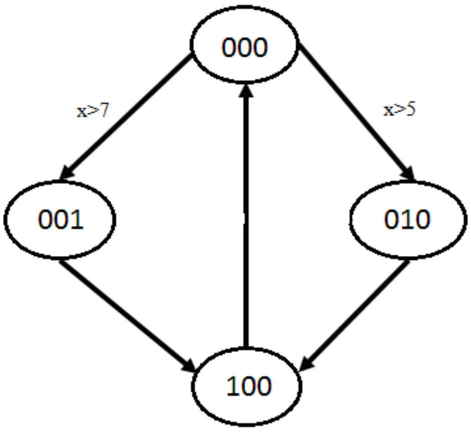
– An automaton cannot contain conflicting transitions that is such kind of transition from the same state when the object can simultaneously transfer into two or more sequent states (except the case of simultaneous sub-automatons). In UML language exclusion of collisions is possible based on typing so called guard conditions (conditional state transitions).

Thus, in particular, conflicting transition can become the branch with ill-conditioned requirement. In picture 9 we can see the sequence diagram and in picture 10 – abstract digital automaton to this diagram. This automaton has conflicting transition: for example, if x=10 then from the state (000) the automaton transfers simultaneously into two states (010) and (001) which contradicts the requirements to construction of the digital automaton. Thus, this sequence diagram is not correct.



Picture 9. The sequence diagram with the branch

The sequence diagram from picture 7 does not have conflicting transitions as the requirement of the branch x<0 and x>=0 do not contradict each other.



Picture 10. Abstract automaton for the sequence diagram from picture 9

Conclusion

The article describes several different approaches to verification of the sequence diagrams. All of them allow us to estimate the correctness of the sequence diagrams from a variety of angles. Thus, the method which is based on the construction of the driver allows us to expose only the most distinctive mistakes in the diagram. Method of protocols allows estimating the correctness of the sequence diagrams formally, observing those messages in it which do not correspond the methods in the diagram of classes. Suggested method is based on the introduction of the sequence diagram as abstract digital automaton allows exposing mistakes in the selection and other conflicting messages. Therefore, the most effective is the complex usage of the given methods

Bibliography

1. Leonenkov A. UML tutorial. Effective instrument of modelling of information systems / Leonenkov A. – Spb.: BHV – St. Petersburg, 2001. – 304p.
2. McGregor J. Testing Object-Based Software: practical guide. / J. McGregor, D. Sykes; translated from English – K.: LLC “TID”DC””, 2002. – 432p.
3. Lytvynov V.V. Formal verification of the class diagrams / V.V. Lytvynov, I.V. Bohdan// Academic periodical “Mathematical machines and systems” № 2. – 2013. – P. 41-47.
4. Samofalov K.G. Applicable theory of digital automatons / K.G. Samofalov, A.M. Romankevych< V.N. Valuiskyi, Y.S. Kanevskyi, M.M. Pynevych – K.: Head publishing house of the publishing association “ Vyscha shkola”, 1987. – 374p.

Authors' Information

Irirna Bogdan *– Postgraduate, the assistant lecturer, Chernihiv National Technological University, 95, Shevchenko street, Chernihiv-27, Ukraine, 14027; e-mail: irakirienko@gamil.com*

Major Fields of Scientific Research: design of object-oriented software, management software projects and their risks, expert systems

Vitaliy Lytvynov *– Dr. Sc. Prof. Chernihiv National Technological University, 95, Shevchenko street, Chernihiv-27, Ukraine, 14027; e-mail: vlitvin@ukrsoft.ua*

Major Fields of Scientific Research: modeling of complicated systems, computer-aided management systems, decision support systems

CHRONOLOGICAL MODELLING OF THE WEST – EUROPEAN INFORMATION ABOUT THE MEDIEVAL MAPS OF THE OTTOMAN WORLD 16TH-18TH CENTURIES

Jordan Tabov, Galina Panayotova

***Abstract**: We present results of a study of the information about the medieval maps and travellers' geographic descriptions of Turkey. It is based on Ian Manners' monograph "European Cartographers and the Ottoman World 1500-1750" (University of Chicago, 2007). The obtained chronological distributions show some peculiarities; we discuss one of them.*

***Keywords**: chronological distribution of information, ottoman world 1500-1750, maps, Little Ice Age*

***ACM Classification Keywords**: I.6 SIMULATION AND MODELLING, I.6.3 Applications*

Introduction

The monograph "European Cartographers and the Ottoman World 1500-1750" and Maps from the Collections of O. J. Sopranos show the level of the mapmakers and the knowledge concerning the Turkish territory between the 15th and 18th centuries. It opens with the intellectual and geographical discoveries of the period that undermined the medieval view of the cosmos and illustrates how mapmakers sought to produce and map a new geography of the world. The maps depict a number of selected spots, characterized by their surface being constantly changed, the way they were at the time they were drawn (or the way the cartographer saw them) and they can also be regarded as universal means of communication. The old maps provide a lot of information on how the world was apprehended in the past.

The goal of the investigation is to find a comparative assessment of the number of the published maps in different periods of time. Here "a comparative assessment" means that we are not interested in the exact number of items. We are investigating the changes of this number in order to build a pattern of the chronologically distributed information as regarding the medieval maps of Turkey and to compare it with another kind of information about Turkey of the same period. The casually picked–up maps in the Sopranos collection are considered to be "representative" and give us the grounds to justify the hypothesis, that the chronological distribution of all maps of Turkey of the period between the 15th and 18th centuries, would be characterized by similarly the same peculiarities.

Used Data and methods

This study is based on bibliographic data published in Jan Manners' book "European cartographers and the Ottoman World 1500-1750" [Manners, 2007] and maps from the collection of O. J. Sopranos.

The book contains a list of maps which includes 59 exhibits (List of figures).

From page 21 to page 57 there are included the following sections: "Mapping and discovery during the Renaissance", The "rediscovery" of Ptolemy, "Asia Propria" – and sixteenth century Ptolemaist Atlases and Isma'il Abu al Fida [Manners, 2007]. There are 57 exhibits enumerated. The information from pages 57 – 67 comprises the section “The Mediterranean Traditions of Carting” and it includes 18 exhibits. From page 61 to page 81 there are depicted maps of cities which existed during this period (Mapping the city). Most numerous are the maps of Istanbul (Constantinople) – 10 of them. There are maps of Babylon, Damascus, Alexandria, Medina, Mecca, Jerusalem and still others. The section of maps drawn by explorers in these lands (Through the eyes of travelers) includes the information from page 81 to page 95 as well as 25 exhibits.

The basis of the idea for the study in this paper of a set of objects is called chronological distribution of information in historical texts: in the article [Tabov, 2003] J. Tabov has proposed methodology for its construction. Its variations have been applied to the design of specific "historical allocations" (abbreviated as XP) for coin finds [Tabov et al, 2003] of old manuscripts [Tabov et al, 2004] which have reached to us as well as other written sources: [Hristova and Dobreva, 2004], [Tabov & Panayotova, 2010].

Chronological distributions

The chronological distribution of the old maps illustrates the intensive use of maps in the 16th-18th centuries.

The command ”search” includes maps from List of figures from the respective intervals of time. There are 59 exhibits corresponding to periods of 25 years.

The information is presented in the following table, where in L(x) there is denoted the distribution of these 59 maps over periods of 25 years.

	To 1500	1500-1525	1525-1550	1550-1575	1575-1600	1600-1625	1625-1650	1650-1675	1675-1700	1700-1725	1725-1750	After 1750
L(x)	3	1	6	15	2	2	0	6	8	6	3	7
½[L(x)+L(x+1)	-	2	3,5	10,5	8,5	2	1	3	7	7	4,5	5

Table 1.

The data can be considered as a time series and there can be investigated the trends in specific variables over time. Using the method of "averaging” the peaks and troughs of any seasonal influences are smoothed by the process. This follows from the substitution of the initial levels of the row with the arithmetic average in the selected time interval (third row of Table 1).

According to the third row of Table 1 we construct a graph (in Excel) of the chronological distribution of the maps listed in “List of figures” of the monograph of Manners [Manners, 2007].

The exhibits included in the present list (from a chronological point of view and complied with the year of printing) represent a random sample of published during the period of 1500-1750 maps of the Ottoman Empire. Table 1 and Diagram 1 may be regarded as a reference to chronological distribution of all published maps of the Empire.

The graph shows the following anomaly: during the period from 1600 to 1650 - in the course of 50 years there were published very few maps, a fact that requires special studies, analyses and explanations. In the period around 1540 - 1580 there was a maximum of all graphs. An increase of the number of published maps is observed after the year of 1650.

Similar anomalies are observed in other publications related to the history of the Ottoman world during this period. Significant decrease in the amount of information has been established by other studies: [Kiel, 2005], [Tabov & Panayotova, 2010] and etc.

The book [Kiel, 2005] offered an explanation that regards this anomaly as a result of climatic changes - the so-called "Little Ice Age", which in recent years has been actively discussed in the scientific literature. We will draw attention to two of the most important summarizing researches on this topic – those of Loehle - 2007 and Huston McCulloch from the year of 2008.

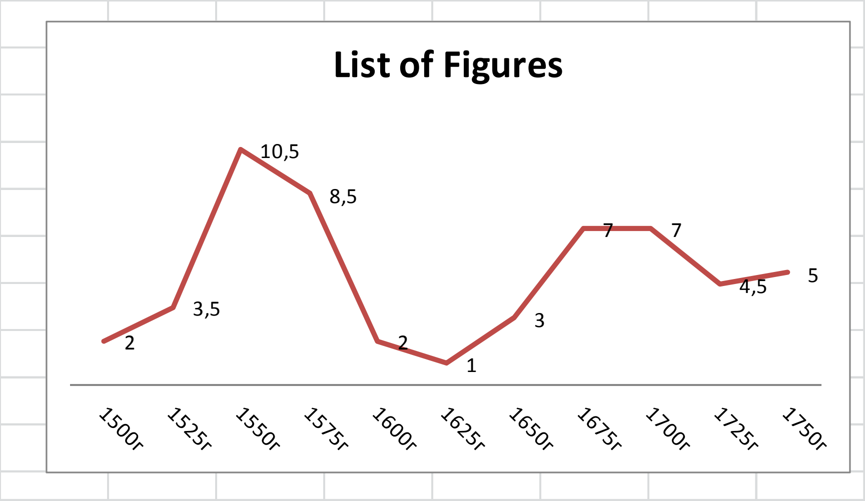


Diagram 1. Chronological distribution of the maps in “List of figures”

In Diagram 2 there is shown a graph of the variation of the average annual temperature from 1400 to 1800 in [Loehle and Huston McCulloch, 2008]. The three lines indicate the highest, middle and lowest deviation from the average (conditionally) annual temperature for the period.

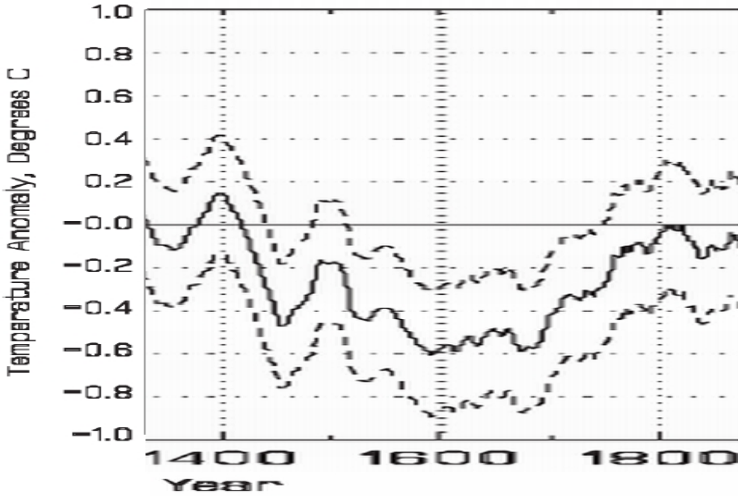


Diagram 2. Graph of the deviation from the average annual temperature from 1400 to 1800 in [Loehle and Huston McCulloch, 2008]

We will use simple graphic (for the period 1500-1700) of the average deviation from the mean annual temperature. The construction of this graph is based in Diagram 2 and shown in Diagram 3. Here the intervals are within 25 years.

The information is presented in Table 2 (second row), where *T*° denotes the average deviation of the mean annual temperature in each period.

	1500	1525	1550	1575	1600	1625	1650	1675	1700	1725	1750
T°	- 0,20	- 0,44	- 0,40	- 0,53	- 0,57	-0,52	- 0,50	- 0,60	- 0,40	- 0,30	- 0,25
½[L(x)+L(x+1)	2	3,5	10,5	8,5	2	1	3	7	7	4,5	5

Table 2.

In Diagram 3 there is shown a simple graph of the deviation from the average annual temperature for the period 1500-1700.

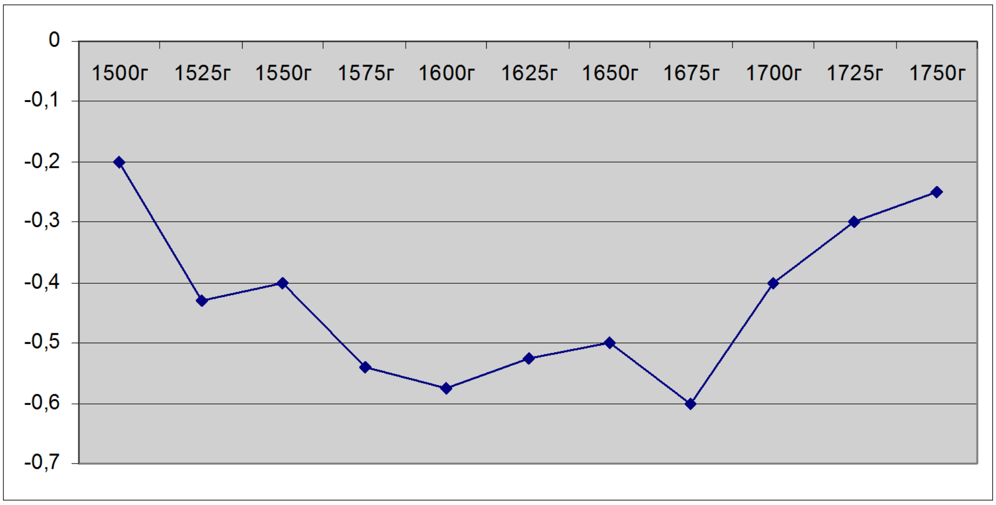


Diagram 3. Simple graphic of the average deviation from the average annual temperature for the period 1500-1700

The data of Table 2 enable us to check the hypothesis of a correlation between the chronological distribution of the maps in “List of figures” of the monograph of Manners [Manners, 2007] and the deviation from the average annual temperature for the period 1500-1700. Using standard software we calculate the coefficient of correlation which is

$$R = - 0{,}189740344$$

This result gives reason to conclude that there is no direct correlation between the values that are being compared, i.e. Kiel assumption is unfounded.

Concluding remarks

The above brief analysis of randomly chosen quantitative information from the exhibit European Cartographers and the Ottoman World 1500-1750; Maps from the Collections of O.J.Sopranos provide evidence in support of the following:

i) In chronological distribution of the number of maps that are the subject of our study, we can judge for the "intensity" of the maps published at different times. Comparing the periods and amplitudes of the growth and decrease we notice some features that require special studies, analyses and explanations. There are features such as - a small number of maps in the first half of the 17th century.

ii) There's been tested the hypothesis of the existence of correlation between chronological distribution of the maps in "List of figures" of the monograph of Manners [Manners, 2007] and the average deviation from the mean annual temperature for the period 1500-1700. There's also calculated a slight negative correlation. The result shows that climate change is not the cause of perceived anomalies.

Bibliography

[Hristova and Dobreva, 2004] S. Hristova, M. Dobreva. Some observations on the chronological distribution of mediaeval manuscripts and church items preserved in Bulgaria. In: Mathematics and Education in Mathematics. Proc. Of the Thirty Third Spring Conference of the Union of Bulgarian Mathematicians, Borovets, April 1-4, 2004, 214-217.

[Kiel, 2005]. Kiel, M. People and tawns in Bulgaria during the Ottoman rule. Collections of papers Amicitia, Sofia, 2005.

[Loehle and Huston McCulloch, 2008] Loehle, C. and J. Huston McCulloch. Correction to: A 2000-year global temperature reconstruction based on non-tree ring proxies. Energy & Environment Vol. 19 (2008), No. 1, 93-100.

[Loehle, 2007], Loehle, C. A 2000-year global temperature reconstruction based on non-tree ring proxies. Energy & Environment Vol. 18 (2007), No. 7+8, 1049-1058.

[Manners, 2007] Manners, I. European Cartographers and the Ottoman World 1500-1750. University of Chicago, 2007.

[Tabov et al, 2003] Tabov, J., Vasilev K. and Velchev A. A mathematical model of monetary circulation in Medieval Bulgaria. Storiadelmondo. 2003: http://www.storiadelmondo.com/14/tabov.monetary.pdf

[Tabov et al, 2004] Tabov, J., A. Velchev, M. Dobreva, K. Sotirova. Chronological distribution of the Bulgarian mediaeval manuscripts preserved in Bulgaria. In: Mathematics and Education in Mathematics. Proc. Of the Thirty Third Spring Conference of the Union of Bulgarian Mathematicians, Borovets, April 1-4, 2004, 257-261.

[Tabov&Panayotova, 2010] Tabov J., G.Panayotova. Model of the chronological distribution of information from the Ottoman archives 1580-1700 in seven articles of Mahiel Kiel. In: T.Atanasova (Editor). Collectanea of papers "Modeling and guidance of information processes".KTR, Sofia, 2010, 142-156.

[Tabov, 2003] Tabov, J. Chronological Distribution of Information in Historical Texts. Computers and the Humanities, 24 (2003), 235-240.

Authors' Information

Jordan Tabov *– Institute of Mathematics and Informatics of the Bulgarian Academy of Sciences, Acad. G.Bonchev Str. Block 8, 1113 Sofia, Bulgaria; e-mail: tabov@math.bas.bg*

Major Fields of Scientific Research: Applications of mathematics and informatics in the humanities, Didactics of mathematics and informatics

Galina Panayotova *– UniBIT – State University of Library Studies and Information Technology, Tsarigradsko shose bul. 119, 1784 Sofia, Bulgaria; e-mail: panayotovag@gmail.com*

Major Fields of Scientific Research: Hyperbolic systems of PDEs, Mathematical modeling and Applications of mathematics, Information Technology

CONSTRUCTION OF CLASS LEVEL DESCRIPTION FOR EFFICIENT RECOGNITION OF A COMPLEX OBJECT

Tatiana Kosovskaya

Abstract: *Many artificial intelligence problems are NP-complete ones. To increase the needed time of such a problem solving a method of extraction of sub-formulas characterizing the common features of objects under consideration is suggested. Repeated application of this procedure allows forming a level description of an object and of classes of objects. A model example of such a level description and the degree of steps number increasing is presented in the paper.*

Keywords: *artificial intelligence, pattern recognition, predicate calculus, complexity of an algorithm, level description of a class.*

ACM Classification Keywords: *I.2.4 Artificial Intelligence Knowledge Representation Formalisms and Methods – Predicate logic, I.5.1 Pattern Recognition Models – Deterministic, F.2.2 Non numerical Algorithms and Problems – Complexity of proof procedures.*

Introduction

Many artificial intelligence problems may be formalized by means of predicate calculus language [Kosovskaya, 2011]. Such a formalization (called a logic-objective approach to AI problems solving) allows to take into account not only properties of an object as a whole but properties of its parts and relations between them. It is proved in [Kosovskaya, 2007] that such a way formalized problems are NP-complete ones, and upper bounds of the number of their solving steps are proved for an exhaustive algorithm and for algorithms based on derivation in a predicate calculus.

A level description of recognized classes was introduced in [Kosovskaya, 2008]. It is based on the definition of auxiliary predicates in the terms of the initial ones. These predicates are determined as “frequently” occurred sub-formulas of the class description having “small complexity”. Conditions of the step number decreasing while solving a recognition problem with the use of a level description were proved. But such an extraction of the mentioned sub-formulas was leaved to a human will.

The notion of partial deduction was introduced in [Kosovskaya, 2009] to recognize an object with incomplete information. The use of partial deduction allows to state that the given information is enough to claim that the r-th part (with the extracting of this part) of the recognized object belongs to the pointed class with the certainty degree q.

The notion of partial deduction was offered in [Kosovskaya, 2012] for determination of a distance (as well as for the degree of similarity) between objects described in the frameworks of the logic-objective approach. The base of a distance determination is sub-formulas of object descriptions which differ one object from another.

Below it is offered to use the notion of partial deduction for the extraction of “frequently” occurred sub-formulas of a class description and construction a level description of this class with the use of a training set. These sub-formulas describe similar characteristics of objects from the same class.

Logic-objective approach to recognition problem setting

To recognize objects from the done set Ω every element of which is represented as a set of its elements $\omega = \{\omega_1, ..., \omega_t\}$, a logic-objective approach was described in [Kosovskaya, 2007]. Let the set of predicates $p_1, ..., p_n$ (every of which is defined on the elements of ω) characterizes properties of these elements and relations between them.

Logical description $S(\omega)$ of an object ω is a collection of all true formulas in the form $p_i(\tau)$ or $\neg p_i(\tau)$ (where τ is an ordered subset of ω) describing properties of ω elements and relations between them.

Let the set Ω is a union of classes $\Omega = \cup_{k=1}^{K} \Omega_k$. Logical description of the class Ω_k is such a formula $A_k(\mathbf{x})$ that if the formula $A_k(\omega)$ is true then $\omega \in \Omega_k$. The class description may be represented as a disjunction of elementary conjunctions of atomic formulas.

Here and below the notation $\mathbf{x}$ is used for an ordered list of the set x. To denote that all values for variables from the list $\mathbf{x}$ are distinct the notation $\exists \mathbf{x}_{\neq} A_k(\mathbf{x})$ is used.

The introduced descriptions allow solving many artificial intelligence problems [Kosovskaya, 2011]. These problems may be formulated as follows. **Identification problem:** to pick out all parts of the object ω which belong to the class Ω_k. **Classification problem:** to find all such class numbers k that $\omega \in \Omega_k$. **Analysis problem:** to find and classify all parts τ of the object ω. The solution of these problems is reduced to the proof of logic sequents $S(\omega) \Rightarrow \exists \mathbf{x}_{\neq}\ A_k(\mathbf{x})$, $S(\omega) \Rightarrow \vee_{k=1}^{K} A_k(\mathbf{x})$, $S(\omega) \Rightarrow \vee_{k=1}^{K} \exists \mathbf{x}_{\neq}\ A_k(\mathbf{x})$ respectively and determination of the values of $\mathbf{x}$ and k.

The proof of every of these sequents is based on the proof of the sequent

$$S(\omega) \Rightarrow \exists \mathbf{x}_{\neq}\ A(\mathbf{x}) \tag{1}$$

where $A(\mathbf{x})$ is an elementary conjunction.

It is proved in [Kosovskaya, 2010] that every of these problems is an NP-complete one. If the sign $\exists$ is changed by the sign ? then every of these problems is an NP-hard one.

Moreover, the number of steps of an algorithm solving the problem (1) (and the problem with the changing of the sign $\exists$ by the sign ?) is $O(t^m)$ (m is the number of arguments in $A(\mathbf{x})$) for an exhaustive algorithm, and $O(s^a)$ (s and a are the maximal and respectively the summary numbers of occurrences of the same predicate in the description $S(\omega)$ and in the formula $A(J)$ respectively) for logical derivation in the first order predicate calculus.

Level description of a class

Let $A_1(\mathbf{x}_1), ..., A_K(\mathbf{x}_K)$ be a set of class descriptions. Let's find all sub-formulas $P_i^1(\mathbf{y}_i^1)$ with the “small complexity” which “frequently” appear in the formulas $A_1(\mathbf{x}_1), ..., A_K(\mathbf{x}_K)$ and denote them by atomic formulas with new predicates p_i^1 having new first-level arguments y_i^1 for lists $\mathbf{y}_i^1$ of initial variables. Such a new predicate p_i^1 is called a first-level predicate. Write down a system of equivalences

$$p_i^1(y_i^1) \Leftrightarrow P_i^1(\mathbf{y}_i^1).$$

Let $A_k^1(\mathbf{x}_k^1)$ be a formula received from $A_k(\mathbf{x}_k)$ by means of a substitution of $p_i^1(y_i^1)$ instead of $P_i^1(\mathbf{y}_i^1)$. Here $\mathbf{x}_k^1$ is a list of all variables in $A_k^1(\mathbf{x}_k^1)$ including both some (may be all) initial variables of $A_k(\mathbf{x}_k)$ and first-level variables appeared in the formula $A_k^1(\mathbf{x}_k^1)$.

A set $S^1(\omega)$ of all atomic formulas of the type $p_i^1(\omega_{ij}^1)$ for which the formula $P_i^1(\boldsymbol{\tau}_{ij}^1)$ (for some $\tau_{ij}^1 \subset \omega$) is valid is called a first-level object description. Such a way extracted lists of ω elements $\omega_{ij}^1 = \tau_{ij}^1$ are called first-level objects.

Repeat the above described procedure with all formulas $A_k^1(\mathbf{x}_k^1)$. After L repetitions L-level descriptions in the following form will be received [Kosovskaya, 2008].

$$A_k^L(\mathbf{x}_k^L)$$

$$p_1^1(y_1^1) \Leftrightarrow P_1^1(\mathbf{y}_1^1)$$

$$\cdots$$

$$p_{n1}^1(y_{n1}^1) \Leftrightarrow P_{n1}^1(\mathbf{y}_{n1}^1)$$

$$\cdots$$

$$p_i^l(y_i^l) \Leftrightarrow P_i^l(\mathbf{y}_i^l)$$

$$\cdots$$

$$p_{nL}^L(y_{nL}^L) \Leftrightarrow P_{nL}^L(\mathbf{y}_{nL}^L).$$

Such an L-level description may be used for efficiency of an algorithm solving a problem formalized in the form of logical sequent (1).

Let's describe an algorithm solving the problem in the form (1) with the use of a level description of a class.

- First, for every i check $S(\omega) \Rightarrow \exists \mathbf{y}_i^1{}_{\neq}\ P_i^1(\mathbf{y}_i^1)$ and find all values of true first-level predicate arguments. Add these first-level true atomic formulas to the object description and form $S^1(\omega)$. If an l-level ($l= 1, \ldots, L-1$) object description $S^l(\omega)$ is formed then for every i check $S^l(\omega) \Rightarrow \exists \mathbf{y}_i^l{}_{\neq}\ P_i^{l+1}(\mathbf{y}_i^{l+1})$ and find all values for true $(l+1)$-level predicate arguments.
- Second, add these $(l+1)$-level true atomic formulas to the object description $S^l(\omega)$ and receive $S^{l+1}(\omega)$.
- Then substitute $p_i^l(y_i^l)$ instead of $P_i^l(\mathbf{y}_i^l)$ into $A_k^l(\mathbf{y}_k^l)$.
- Repeat the previous steps for $l = 1, \ldots, L$.
- At last check $S^L(\omega) \Rightarrow \exists \mathbf{y}_k^L{}_{\neq}\ A_k^L(\mathbf{y}_k^L)$.

To decrease the number of steps of an exhaustive algorithm (for every t greater than some t_0) with the use of 2-level description it is sufficient that

$$n_1 t^{r}+t^{s1+n1}<t^{m}, \tag{2}$$

where r is a maximal number of arguments in the formulas $p_i^1(y_i^1) \Leftrightarrow P_i^1(\mathbf{y}_i^1)$, n_1 is the number of first-level predicates, s^1 is the number of atomic formulas in $S^1(\omega)$, m is the number of variables in the initial class description [Kosovskaya, 2008].

Analogous condition for decreasing the number of steps of a logical algorithm solving the problem (1) is

$$\Sigma_{k=1\ldots K} s^{ak} - \Sigma_{j=1\ldots n1} s^{\rho j} \geq \Sigma_{\kappa=1\ldots K} (s^1)^{ak1}, \tag{3}$$

where a_k and a_k^1 are the numbers of atomic formulas in $A_k(\mathbf{x}_k)$ and $A_k^1(\mathbf{x}_k^1)$ respectively, s and s^1 are the maximal numbers of atomic formulas with the same predicate in $S(\omega)$ and $S^1(\omega)$ respectively, ρ_j is the number of atomic formula in $P_j^1(\mathbf{y}_j^1)$ [Kosovskaya, 2008].

Partial deduction

The notion of partial deduction was introduced by the author in [Kosovskaya, 2009] to recognize objects with incomplete information. During the process of partial deduction instead of the proof of (1) we search such a maximal sub-formula $A'(\mathbf{x}')$ of the formula $A(\mathbf{x})$ that $S(\omega) \Rightarrow \exists \mathbf{x}'_{\neq} A'(\mathbf{x}')$ and there is no information that $A(\mathbf{x})$ is not satisfiable on ω.

Let a and a' be the numbers of atomic formulas in $A(\mathbf{x})$ and $A'(\mathbf{x}')$ respectively, m and m' be the numbers of objective variables in $A(\mathbf{x})$ and $A'(\mathbf{x}')$ respectively. Then partial deduction means that the object ω contains an r-th part ($r = m'/m$) of an object satisfying the description $A(\mathbf{x})$ with the certainty $q = a'/a$.

More precisely, the formula $S(\omega) \Rightarrow \exists \mathbf{x}_{\neq} A(\mathbf{x})$ is partially (q, r) - deduccible if there exists a maximal sub-formula $A'(\mathbf{x}')$ of the formula $A(\mathbf{x})$ such that $S(\omega) \Rightarrow \exists \mathbf{x}'_{\neq} A'(\mathbf{x}')$ is deducible and τ is the string of values for the list of variables $\mathbf{x}'$, but the formula $S(\omega) \Rightarrow \exists \mathbf{x}_{\neq} [DA'(\mathbf{x})]^{x'}_{\tau}$ is not deductible. Here $[DA'(\mathbf{x})]^{x'}_{\tau}$ is obtained from $A(\mathbf{x})$ by deleting from it all conjunctive members of $A'(\mathbf{x}')$, substituting values of τ instead of the respective variables of $\mathbf{x}'$ and taking the negation of the received formula.

Class description based on the training set

Given a training set $\Omega^0 = \cup_{k=1}^K \Omega_k^0$ let's make such a class description that every object from Ω^0 would be successfully classified. Every object $\omega = \{\omega_1, \ldots, \omega_t\}$ from Ω^0 is represented by its description $S(\omega)$. If one replaces in $S(\omega)$ every constant ω_j by a variable x_j ($j = 1, \ldots, t$) and substitute the sign & between the received atomic formulas then such an elementary conjunction $A(\mathbf{x})$ would be valid for every object with the same description.

A disjunction upon all objects from Ω_k^0 of all such a way received elementary conjunctions may be regarded as a description of the class Ω_k^0. Moreover, if for a display screen image the indexes of neighboring pixels are changed, for example, by x and $x + 1$ then every image differing from the one in the training set only by its localization on the display screen will be correctly classified.

The object that does not satisfy any of the received class description may be classified according to the metric described in [Kosovskaya, 2012].

Formation of a level description for one class

Let the class description $A_k(\mathbf{x})$ is a disjunction of elementary conjunctions $A_{k,1}(\mathbf{x}_{k,1}), \ldots, A_{k,J}(\mathbf{x}_{k,J})$. For every *i* and *j* (*I* < *j*) check whether $A_{k,i}(\mathbf{x}_{k,i}) \Rightarrow \exists \mathbf{x}_{k,j\,\neq}\ A_{k,j}(\mathbf{x}_{k,j})$. Using the notion of partial deduction we may receive such a maximal sub-formula $Q^1_{i,j}(\mathbf{x}_{i,j})$ of the formula $A_{k,j}(\mathbf{x}_{k,j})$ that $A_{k,i}(\mathbf{x}_{k,i}) \Rightarrow \exists \mathbf{x}_{i,j\,\neq}\ Q^1_{i,j}(\mathbf{x}_{i,j})$. But $Q^1_{i,j}(\mathbf{x}_{i,j})$ is also a maximal sub-formula of $A_{k,i}(\mathbf{x}_{k,i})$ (up to the names of variables) such that $A_{k,i}(\mathbf{x}_{k,i}) \Rightarrow \exists \mathbf{x}_{i,j\,\neq}\ Q^1_{i,j}(\mathbf{x}_{i,j})$ because the both formulas $A_{k,i}(\mathbf{x}_{k,i})$ and $A_{k,i}(\mathbf{x}_{k,i})$ are elementary conjunctions.

So such a way received formula $Q^1_{i,j}(\mathbf{x}_{i,j})$ is a common sub-formula of $A_{k,i}(\mathbf{x}_{k,i})$ and $A_{k,i}(\mathbf{x}_{k,i})$ (up to the names of variables).

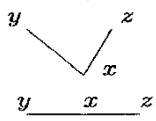
A common sub-formula $Q^l_{i1\ldots il,j1\ldots jl}(\mathbf{x}_{i1\ldots il,j1\ldots jl})$ of the formulas $Q^{l-1}_{i1\ldots il}(\mathbf{x}_{i1\ldots il})$ and $Q^{l-1}_{j1\ldots jl}(\mathbf{x}_{j1\ldots jl})$ (up to the names of variables) may be received in the similar way.

Note that the length of $Q^l_{i1\ldots il,j1\ldots jl}(\mathbf{x}_{i1\ldots il,j1\ldots jl})$ decreases while increasing the value of *l*. That is why the process would stop. One can fix such a number *r* (*r* > *1*) that if the length of $Q^l_{i1\ldots il,j1\ldots jl}(\mathbf{x}_{i1\ldots il,j1\ldots jl})$ is less than *r* then it is not involved into the further search of sub-formulas.

Choose sub-formulas $Q^l_{i1\ldots il,j1\ldots jl}(\mathbf{x}_{i1\ldots il,j1\ldots jl})$ satisfying a condition (2) or (3) in dependence of what algorithm would be used for the proof of (1). All these sub-formulas are denoted by $P_i^1(\mathbf{y}_i^1)$ ($i = 1, \ldots n1$) and form the set of first-level predicates.

The *(l + 1)*-level predicates are formed from $Q^l_{i1\ldots il,j1\ldots jl}(\mathbf{x}_{i1\ldots il,j1\ldots jl})$ which sub-formulas are included into the set of *l*-level predicates taking into account a condition (2) or (3).

Example of sub-formulas extracting



Given two predicates $V(x,y,z) \Leftrightarrow$ "$\angle yxz < \pi$" and $L(x,y,z) \Leftrightarrow$ "*x belongs the segment (y,z)*" describe the class of "boxes" according to the training set represented on the Figure 1 and extract common sub-formulas in order to built a level description.

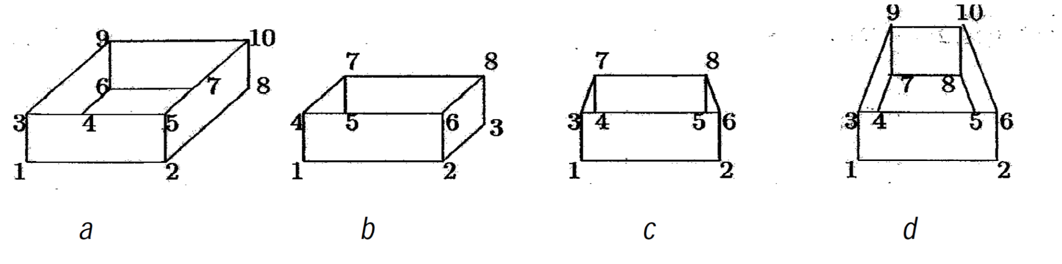


Figure 1. Standard different foreshortened contour images

These standard images allow forming a description (up to mirror image) of almost all boxes. Such a description is a disjunction of 4 elementary conjunctions containing respectively 10, 8, 10, 8 variables and 30, 22, 26, 32 atomic formulas. For example, the elementary conjunction corresponding to the image *b* is $V(x_1,x_4,x_2)$ & $V(x_2,x_1,x_6)$ & $V(x_2,x_6,x_3)$ & $V(x_2,x_1,x_3)$ & $V(x_3,x_2,x_8)$ & $V(x_4,x_5,x_1)$ & $V(x_4,x_6,x_1)$ & $V(x_4,x_7,x_5)$ & $V(x_5,x_4,x_7)$ & $V(x_5,x_7,x_6)$ & $V(x_6,x_2,x_5)$

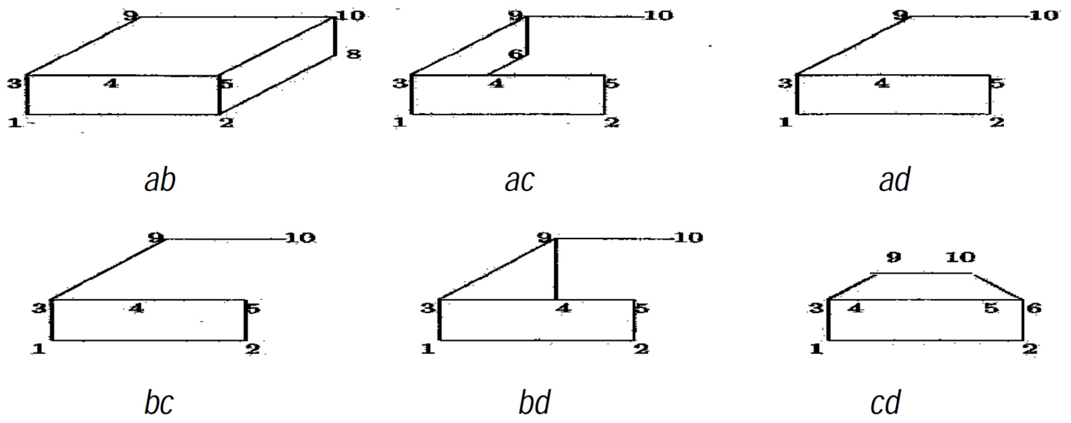
& $V(x_6,x_2,x_4)$ & $V(x_6,x_5,x_8)$ & $V(x_6,x_4,x_8)$ & $V(x_6,x_8,x_2)$ & $V(x_7,x_5,x_4)$ & $V(x_7,x_8,x_5)$ & $V(x_7,x_8,x_4$ & $V(x_8,x_3,x_6)$ & $V(x_8,x_6,x_7)$ & $V(x_8,x_3,x_7)$ & $T(x_5,x_4,x_6)$.

Given a “box” inside a complex contour image containing t nodes it would be recognized in $O(t^{10})$ steps by an exhaustive algorithm and in $O(s^{29})$ steps by a logical algorithm (here s is the maximal number of occurrences of the same predicate in the description $S(\omega)$).

Pair wise partial deduction of these elementary conjunctions allows extracting common sub-formulas corresponding to the images represented on Figure 2.

These sub-formulas contain respectively 8, 8, 7, 7, 7, 8 variables and 18, 15, 11, 11, 15, 16 atomic formulas.

The following extraction by means of pairwise partial deduction between common sub-formulas corresponding images *ab, ac, ad, bc, bd, cd* gives a sub-formula corresponding to the image represented on Figure 3.



ab *ac* *ad*

bc *bd* *cd*

Figure 2. Images corresponding to extraction of common sub-formulas

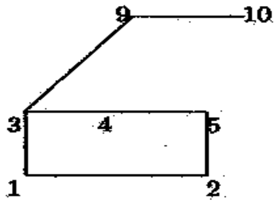


Figure 3. Image corresponding to the second extraction of common sub-formulas

Elementary conjunction $P^1(x_1,x_2,x_3,x_4,x_5,x_9,x_{10}) = V(x_1,x_3,x_2)$ & $V(x_2,x_1,x_5)$ & $V(x_3,x_4,x_1)$ & $V(x_3,x_5,x_1)$ & $V(x_3,x_9,x_4)$ & $V(x_3,x_9,x_5)$ & $V(x_3,x_9,x_1)$ & $V(x_5,x_2,x_4)$ & $V(x_5,x_2,x_3)$ & $V(x_9,x_{10},x_3)$ & $T(x_4,x_3,x_5)$ corresponding to this image defines a first-level predicate $p^1(x^1)$. The first-level variable x^1 is a variable for a list of 7 initial variables.

Elementary conjunctions $P_1^2(\mathbf{y}_1^1)$, $P_2^2(\mathbf{y}_1^1)$, $P_3^2(\mathbf{y}_1^1)$, $P_4^2(\mathbf{y}_1^1)$ corresponding to the images *ab, ac, bd, cd* and written with the use of the predicate $p^1(x^1)$ define second-level predicates $p_1^2(x_1^2)$, $p_2^2(x_2^2)$, $p_3^2(x_3^2)$, $p_4^2(x_4^2)$.

For example, a sub-formula corresponding to the image *ab* is $P_1^2(\mathbf{y}_1^1) = p^1(x^1)$ & $V(x_2,x_5,x_8)$ & $V(x_2,x_1,x_8)$ & $V(x_5,x_4,x_{10})$ & $V(x_5,x_3,x_{10})$ & $V(x_8,x_2,x_{10})$ & $V(x_{10},x_8,x_5)$ & $V(x_{10},x_5,x_9)$ & $V(x_{10},x_8,x_9)$, where $\mathbf{y}_1^1$ is a list of variables $x^1,x_1,x_2,x_4,x_5,x_8,x_9,x_{10}$ and x^1 is a variable for a list of initial variables $x_1,x_2,x_3,x_4,x_5,x_9,x_{10}$.

Given a "box" inside a complex contour image containing *t* nodes the proof the sequence from S(ω) of elementary conjunction $P^1(x_1,x_2,x_3,x_4,x_5,x_9,x_{10})$ defining the first-level predicate $p^1(x^1)$ and the denotation of variables $x_1,x_2,x_3,x_4,x_5,x_9,x_{10}$ would be done in O(t^7) steps by an exhaustive algorithm and in O(s^{11}) steps by a logical algorithm.

Elementary conjunctions $P_1^2(\mathbf{y}_1^1)$, $P_2^2(\mathbf{y}_1^1)$, $P_3^2(\mathbf{y}_1^1)$, $P_4^2(\mathbf{y}_1^1)$ contain respectively only 1, 1, 0, 1 "new" variables and 7, 4, 4, 5 "new" atomic formulas. The proof of the sequence from S^1(ω) of these elementary conjunctions defining the second-level predicates $p_1^2(x_1^2)$, $p_2^2(x_2^2)$, $p_3^2(x_3^2)$, $p_4^2(x_4^2)$ and the denotation of the "new" variables would be done in O(t) steps by an exhaustive algorithm and in O(s^7) steps by a logical algorithm.

Elementary conjunctions obtained from the class description by means of second-level predicates instead of the corresponding sub-formulas contain respectively 2, 0, 2, 2 "new" variables and 7, 4, 11, 16 "new" atomic formulas. The proof of the sequence from S^2(ω) of these elementary conjunctions and the denotation of the "new" variables would be done in O(t^2) steps by an exhaustive algorithm and in O(s^{16}) steps by a logical algorithm.

As O(t^7) + O(t) + O(t^2) = O(t^7) < O(t^{10}) and O(s^{11}) + O(s^7) + O(s^{16}) = O(s^{16}) < O(s^{29}) then both an exhaustive algorithm and a logical algorithm using the built level description of the class of "boxes" make the less number of steps then the same ones using the initial description.

Conclusion

In the frameworks of logic-objective approach, objects and classes of an AI problem are described in the terms of properties of the object parts and relations between them. Such an approach allows taking into account characteristics that are common for many objects from the same class. It is very important because while checking the belonging of an object to a class, some generalized characteristics of an object have the main significance. The most of the objects of a class must have these generalized characteristics.

A level description of recognized classes was offered to decrease the computational complexity of a recognition problem solving. It is not proved now that the proposed manner of sub-formulas extraction provides such a decreasing. But it is illustrated above by an example (and may be illustrated by several other examples) that the computational complexity of an analysis problem decreases and a generalized characteristic of a class is formed.

Acknowledgement

The paper is published with financial support of RFBR grant 14-08-01276.

Bibliography

[Kosovskaya, 2007] T.M. Kosovskaya. Proofs of the number of steps bounds for solving of some pattern recognition problems with logical description. In: Vestnik of St.Petersburg University. Ser. 1. 2007. No. 4. P. 82 – 90. (In Russian)

[Kosovskaya, 2008] T.M. Kosovskaya. Level descriptions of classes for decreasing step number of pattern recognition problem solving described by predicate calculus formulas. In: Vestnik of St.Petersburg University. Ser. 10. 2008. No. 1. P. 64 – 72. (In Russian)

[Kosovskaya, 2009] T.M. Kosovskaya. Partial deduction of a predicate formula as an instrument for recognition of an object with incomplete description. In: Vestnik of St.Petersburg University. Ser. 10. 2009. No. 1. P. 77 – 87. (In Russian)

[Kosovskaya, 2010] T.M. Kosovskaya. Some artificial intelligence problems permitting formalization by means of predicate calculus language and upper bounds of their solution steps // Proceedings of SPIIRAS. 2010, No 14.

[Kosovskaya, 2011] Kosovskaya T. Discrete Artificial Intelligence Problems and Number of Steps of their Solution // In: International Journal "Information Theories and Applications", Vol. 18, Number 1, 2011. P. 93 – 99.

[Kosovskaya, 2012] Kosovskaya T. Distance between objects described bt predicate formulas // In: International Book Series. Informational Science and Computing. Book 25. Mathematics of Distances and Applications (Michel Deza, Michel Petitjean, Krasimir Markov(eds)), ITHEA-Publisher, Sofia, Bulgaria, 2012. P. 153 – 159.

Authors' Information

Tatiana Kosovskaya *– Dr., Professor of St. Petersburg State University, University av., 28, Stary Petergof, St. Petersburg; Senior researcher of St. Petersburg Institute in Informatics and Automation of Russian Academy of Science, 14 line, 39, St.Petersburg, 199178, Russia; Professor of St.Petersburg State Marine Technical University, Lotsmanskaya ul., 3, St.Petersburg, 190008, Russia , 198504, Russia, e-mail: kosovtm@gmail.com*

Major Fields of Scientific Research: Logical approach to artificial intelligence problems, theory of complexity of algorithms.

TABLE OF CONTENTS