INFORMATION SCIENCE & COMPUTING

International Book Series Number 2

Advanced Research in Artificial Intelligence

Supplement to
International Journal "Information Technologies and Knowledge" Volume 2 / 2008

ITHEA
SOFIA, 2008

Krassimir Markov, Krassimira Ivanova, Ilia Mitov (ed.)

Advanced Research in Artificial Intelligence

International Book Series "INFORMATION SCIENCE & COMPUTING", Number 2

Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 2 / 2008

Institute of Information Theories and Applications FOI ITHEA

Sofia, Bulgaria, 2008

This issue contains a collection of papers in the the field of Neural Networks, Genetic Algorithms and Natural Language Processing. Papers are selected from the International Conferences of the Joint International Events of Informatics "ITA 2008", Varna, Bulgaria.

International Book Series "INFORMATION SCIENCE & COMPUTING", Number 2
Supplement to the International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 2, 2008

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

Publisher: Institute of Information Theories and Applications FOI ITHEA, Sofia, 1000, P.O.B. 775, Bulgaria.
Издател: Институт по информационни теории и приложения ФОИ ИТЕА, София, 1000, п.к. 775, България
www.ithea.org, www.foibg.com, e-mail: info@foibg.com

General Sponsor: **Consortium FOI Bulgaria** (www.foibg.com).

Printed in Bulgaria

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)** **ISSN 1313-0501 (CD/DVD)**

PREFACE

The scope of the International Book Series “Information Science and Computing” (**IBS ISC**) covers the area of Informatics and Computer Science. It is aimed to support growing collaboration between scientists from all over the world. IBS ISC is official publisher of the works of the members of the ITHEA International Scientific Society.

The official languages of the IBS ISC are English and Russian.

IBS ISC welcomes scientific papers and books connected with any information theory or its application.

IBS ISC rules for preparing the manuscripts are compulsory. The rules for the papers and books for IBS ISC are given on www.foibg.com/ibsisc. The camera-ready copy of the papers and books should be received by e-mail: info@foibg.com.

Responsibility for papers and books published in IBS ISC belongs to authors.

The Number 2 of the IBS ISC contains collection of papers from the field of Neural Networks, Genetic Algorithms and Natural Language Processing. Papers are peer reviewed and are selected from the several International Conferences, which were part of the Joint International Events of Informatics “ITA 2008”, Varna, Bulgaria.

ITA 2008 has been organized by

Institute of Information Theories and Applications FOI ITHEA

in collaboration with:

- ITHEA International Scientific Society
- International Journal “Information Theories and Applications”
- International Journal "Information Technologies and Knowledge"
- Association of Developers and Users of Intelligent Systems (Ukraine)
- Association for Development of the Information Society (Bulgaria)
- V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
- Institute of Mathematics and Informatics, BAS (Bulgaria)
- Institute of Information Technologies, BAS (Bulgaria)
- Institute of Mathematics of SD RAN (Russia)
- Taras Shevchenko National University of Kiev (Ukraine)
- Universidad Politecnica de Madrid (Spain)
- BenGurion University (Israel)
- Rzeszow University of Technology (Poland)
- University of Calgary (Canada)
- University of Hasselt (Belgium)
- Kharkiv National University of Radio Electronics (Ukraine)
- Astrakhan State Technical University (Russia)
- Varna Free University “Chernorizets Hrabar” (Bulgaria)
- National Laboratory of Computer Virology, BAS (Bulgaria)
- Uzhgorod National University (Ukraine)
- Sofia University “Saint Kliment Ohridski” (Bulgaria)
- Technical University – Sofia (Bulgaria)
- New Bulgarian University (Bulgaria)

The main ITA 2008 events were:

KDS	XIVth International Conference "Knowledge - Dialogue – Solution"
i.Tech	Sixth International Conference "Information Research and Applications"
MeL	Third International Conference "Modern (e-) Learning"
ISK	Second International Scientific Conference "Informatics in the Scientific Knowledge"
INFOS	International Conference "Intelligent Information and Engineering Systems"
GIT	Sixth International Workshop on General Information Theory
CS	Third International Workshop "Cyber Security"
eM&BI	Second International Workshop "e-Management & Business Intelligence"
IMU ICT	International Seminar "Information Models' Utility in Information and Communication Technologies"
ISSI	Second International Summer School on Informatics

More information about ITA 2008 International Conferences is given at the www.foibg.com.

The great success of ITHEA International Journals, International Book Series and International Conferences belongs to the whole of the ITHEA International Scientific Society.

We express our thanks to all authors, editors and collaborators who had developed and supported the International Book Series "Information Science and Computing".

General Sponsor of IBS ISC is the **Consortium FOI Bulgaria** (www.foibg.com).

Sofia, June 2008 *Kr. Markov, Kr. Ivanova, I. Mitov*

TABLE OF CONTENTS

Papers in English

<u>Papers in Russian</u>

INDEX OF AUTHORS

EXTENDED NETWORKS OF EVOLUTIONARY PROCESSORS

Luis Fernando de Mingo, Nuria Gómez Blas, Francisco Gisbert, Miguel A. Peña

***Abstract**: This paper presents an extended behavior of networks of evolutionary processors. Usually, such nets are able to solve NP-complete problems working with symbolic information. Information can evolve applying rules and can be communicated though the net provided some constraints are verified. These nets are based on biological behavior of membrane systems, but transformed into a suitable computational model. Only symbolic information is communicated. This paper proposes to communicate evolution rules as well as symbolic information. This idea arises from the DNA structure in living cells, such DNA codes information and operations and it can be sent to other cells. Extended nets could be considered as a superset of networks of evolutionary processors since permitting and forbidden constraints can be written in order to deny rules communication.*

***Keywords**: Networks of Evolutionary Processors, Membrane Systems, and Natural Computation.*

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

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Natural sciences, and especially biology, represent a rich source of modeling paradigms. Well-defined areas of artificial intelligence (genetic algorithms, neural networks), mathematics, and theoretical computer science (L systems, DNA computing) are massively influenced by the behavior of various biological entities and phenomena. In the last decades or so, new emerging fields of so-called "natural computing" [1,2,3] identify new (unconventional) computational paradigms in different forms. There are attempts to define and investigate new mathematical or theoretical models inspired by nature [8], as well as investigations into defining programming paradigms that implement computational approaches suggested by biochemical phenomena. Especially since Adleman's experiment [4] these investigations received a new perspective. One hopes that global system-level behavior may be translated into interactions of a myriad of components with simple behavior and limited computing and communication capabilities that are able to express and solve, via various optimizations, complex problems otherwise hard to approach.

The origin of networks of evolutionary processors is a basic architecture for parallel and distributed symbolic processing, related to the Connection Machine [7] as well as the Logic Flow paradigm [5], which consists of several processors, each of them being placed in a node of a virtual complete graph, which are able to handle data associated with the respective node. All the nodes send simultaneously their data and the receiving nodes handle also simultaneously all the arriving messages, according to some strategies, see, e.g., [6,7].

Networks of evolutionary processors (NEP) [9,11] are language-generating device, if we look at the strings collected in the output node. We can also look at them as doing some computation. If we consider these networks with nodes having filters defined by random context conditions, which seems to be closer to the recent possibilities of biological implementation, then using these simple mechanisms we can solve NP-complete problems in linear time. Such solutions are presented for the *Bounded Post Correspondence Problem* in [10] for the *3-Colorability Problem* in [9] and for the {\it Common Algorithmic Problem} in [12] As a further step, in [12] the so-called hybrid networks of evolutionary processors are considered. Here deletion node or insertion node has its own working mode (performs the operation at any position, in the left-hand end or in the right-hand end of the word) and different nodes are allowed to use different ways of filtering. Thus, the same network may have nodes where the deletion operation can be performed at arbitrary position and nodes where the deletion can be done only at right-end of the word.

In this paper, we present some results regarding a network of evolutionary processor based on an extended behavior from the biological point of view. This is a preliminary work in which rules pass through the net, they are not associated to a fix processor.

Networks of Evolutionary Processors

Connectionist models consists of several processors which are located in the nodes of a virtual graph and are able to perform operations in that node, according to some predefined rules. Information is passed through connections in order to obtain a collaborative solution to a given problem. All processors work in a parallel way and they only perform simple operations.

A network of evolutionary processors is a tuple

$$\Gamma = (V, U, G, \mathcal{N}, \alpha, x_I, x_O)$$

- V, U are the net alphabet and input alphabet respectively.
- G is an undirected graph in which each node is a processor. Processors have a set of objects/strings and a set of evolution rules.
- N is a mapping that associates each processor with a set of filters.
- α is a mapping that defines the behavior of filters (weak or strong conditions).
- x_I, x_O are the input and output processors.

Objects in processors can evolve and communicated to other connected processors. That is, rules can be applied (evolution) or objects can pass filter conditions (communication). These two steps could be sequential (evolution and then communication) or parallel (evolution and communication at same time).

Evolution

A given string *x* can evolve provided there is some rule to apply it. Taking into account that there are an arbitrary large number of copies of string *x* in processor, several rules can be applied in parallel to different copies in one evolutionary step.

Therefore, the set of objects in a processor *i* after an evolution step, denoted by L'_i, are those before the evolution (L_i) adding objects obtained after rules in *i* are applied. That is,

$$L'_i = L_i \cup r_k(x),\ x \in A_i, r_k \in R_i$$

Communication

A given string/object *x* can pass filters in processor *i*, iff the following constraint is satisfied:

$$\varphi_i^{(\beta)}(x; \mathcal{N}(i)), \beta = \{s, w\}$$

Where, $\mathcal{N}(i)$ is the filter set associated to a given processor *i*. This set can contain just input and output filters, or forbidden context filters as well. That is, $\mathcal{N}(i) = \{PI, PO\}, \text{or } \mathcal{N}(i) = \{FI, PI, FO, PO\}$. Constraints are defined as follows:

- Constraints conditions with permitting filters (*PI, PO*); where *P* is either input filter or output filter, it depends if the string is sent out or received with strong conditions (s) or weak condition (w):

$$\varphi_i^{(s)}(x; \mathcal{N}(i)) \equiv P \subseteq alph(x)$$
$$\varphi_i^{(w)}(x; \mathcal{N}(i)) \equiv P \cap alph(x) \neq \emptyset$$

- Constraints conditions with permitting filters (*PI, PO*) and forbidden context (*FI, FO*); where *P, F* is either input filter or output filter, it depends if the string is sent out or received with strong conditions (s) or weak conditions (w):

$$\varphi_i^{(s)}(x;\mathcal{N}(i)) \equiv P \subseteq alph(x) \wedge F \cap alph(x) = \emptyset$$

$$\varphi_i^{(w)}(x;\mathcal{N}(i)) \equiv P \cap alph(x) \neq \emptyset \wedge F \cap alph(x) = \emptyset$$

Therefore, the set of objects in a processor *i* after a communication step, denoted by *L'ᵢ*, are those before the communication (L_i) removing objects sent out and adding objects from other processors connected to *i*. That is,

$$L_i' = L_i - \{w|\varphi^k(x;\mathcal{N}(N_i))\} \bigcup_{\{N_i,N_j\}\in E} \{x|\varphi^k(x;\mathcal{N}(N_j)) \wedge \varphi^k(x;\mathcal{N}(N_i))\}$$

The most important result of such networks of evolutionary processors is that they can solve NP-complete problems in linear time and linear resources.

Extended Networks of Evolutionary Processors

The communication step is only applied to objects in traditional nets. An extended version is proposed in order to be able to send rules from one processor to other ones. This property provides a more realistic behavior since operations are not fixed in processors, they can pass through the net in the same way objects do.

A rule can pass filter conditions provided:

$$\varphi_i^{(\beta)}(r_j : x \to y;\mathcal{N}(i)) \equiv \varphi_i^{(\beta)}(x;\mathcal{N}(i)) \wedge \varphi_i^{(\beta)}(y;\mathcal{N}(i))$$

That is, given rule r_j belonging to processor *i* can be sent out if both antecedent and consequent strings can pass filters in processor *i*, and can be received by other processors if the rule pass their input filters (weak or strong conditions).

A network of evolutionary processors can be transformed into an equivalent extended network of evolutionary processors just choosing the right filters. For example, antecent belonging to rules can be added to forbidden filter in order to avoid rules communication.

Following theorems regarding computational power of non-extended networks of evolutionary processors can be also applied to extended networks of evolutionary processors.

Theorem 1. A complete NEP of size 5 can generate each recursively enumerable language. [9]

Theorem 2. A star NEP of size 5 can generate each recursively enumerable language. [9]

Theorem 3. The bounded PCP can be solved by an NEP in size and time linearly bounded by the product of K and the length of the longest string of the two Post lists. [12]

Theorem 4. The families of regular and context-free languages are incomparable with the family of languages generated by simple NEPs. [10]

Theorem 5. The 3-colorability problem can be solved in O(m + n) time by a complete simple NEP of size 7m+2, where n is the number of vertices and m is the number of edges of the input graph. [10]

Conclusions and Future Work

This paper presents an extended behavior in networks of evolutionary processors. Now, rules can pass from one processor to another one provided filter constraints are satisfied. This mechanism allows operations and data to pass through the net, not only data like in networks of evolutionary processors. This idea tries to model DNA behavior in which information combines data and operations for living cells, the information is a whole, does not matther if it is data or operations. Rules can pass filters as well as objects do, according to filter specifications.

It is clear that this model is a superset of networks of evolutionary processor and therefore it can solve NP-complete problems in linear time and linear resources. Main advantage of such extended model is that the time to solve a problem is lower than non-extended models since rules can travel to transform objects at different processors.

There are some other possibilities when defining conditions of rules in order to pass filters,

$$\varphi_i^{(\beta)}(r_j : x \rightarrow y; \mathcal{N}(i)) \equiv \varphi_i^{(\beta)}(x \cup y; \mathcal{N}(i))$$
$$\varphi_i^{(\beta)}(r_j : x \rightarrow y; \mathcal{N}(i)) \equiv \varphi_i^{(\beta)}(x \cap y; \mathcal{N}(i))$$
$$\varphi_i^{(\beta)}(r_j : x \rightarrow y; \mathcal{N}(i)) \equiv \varphi_i^{(\beta)}(x \setminus y; \mathcal{N}(i))$$

A lot of open problems that can be taken into account to probe computational power of extended networks of evolutionary processors. First step will consist on solving same problems than non-extended net in order to compare time and resources.

Bibliography

[1] Zandron, C. (2002). A Model for Molecular Computing: Membrane Systems, Universita degli Studi di Milano, Italy.

[2] Paun, G. (2002). Membrane Computing. An Introduction. Springer-Verlag, Berlin.

[3] Ciobanu, G., Paun, G., and Perez-Jimenez, M. (2005). Applications of Membrane Computing. Springer-Verlag, Berlin.

[4] Adleman, L. (1994). Molecular computation of solutions to combinatorial problems. Science, 226:1021–1024.

[5] Errico, L. and Jesshope, C. (1994). Towards a new architecture for symbolic processing. Artificial Intelligence and Information Control Systems of Robots, pages 31–40.

[6] Fahlman, S., Hinton, G., and Seijnowski, T. (1983). Massively parallel architectures for ai: Massively parallel architectures for ai: Netl, thistle and boltzmann machines. In AAAI National Conference on Artificial Intelligence, pages 109–113.

[7] Hillis, W. (1985). The Connection Machine. MIT Press, Cambridge.

[8] Robinson, D. A. (1992). Implications of neural networks for how we think about brain function. Behavioral and Brain Sciences, 15(4): 644–655.

[9] J. Castellanos, C. Martın-Vide, V. Mitrana, and J. Sempere (2003). Networks of evolutionary processors. Acta Informatica, 39:517–529,

[10] J. Castellanos, C. Martın-Vide, V. Mitrana, and J. Sempere (2001). Solving np-complete problems with networks of evolutionary processors. Lecture Notes in Computer Science, 2084:621–628.

[11] E. C. Varju and V. Mitrana (2000). Evolutionary systems, a language generating device inpired by evolving communities of cells. Acta Informatica, 36:913–926,

[12] C. Martin-Vide, V. Mitrana, M. Perez-Jimenez, and F. S. Caparrini (2003). Hybrid networks of evolutionary processors. Lecture Notes in Computer Science, 2723:401–412.

Authors' Information

Luis Fernando de Mingo López – *Dept. Organización y Estructura de la Información, Escuela Universitaria de Informática, Universidad Politécnica de Madrid, Crta. De Valencia km. 7, 28031 Madrid, Spain;*
e-mail: lfmingo@eui.upm.es

Nuria Gómez Blas – *Dept. Organización y Estructura de la Información, Escuela Universitaria de Informática, Universidad Politécnica de Madrid, Crta. De Valencia km. 7, 28031 Madrid, Spain;*
e-mail: ngomez@dalum.eui.upm.es

Francisco Gisbert – *Dept. Lenguajes, Sistemas Informáticos e Ingeniería del Software, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Madrid, Spain; e-mail: fgisbert@fi.upm.es*

Miguel Angel Peña – *Dept. Lenguajes, Sistemas Informáticos e Ingeniería del Software, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Madrid, Spain;*
e-mail: fgisbert@fi.upm.es

THE CASCADE ORTHOGONAL NEURAL NETWORK

Yevgeniy Bodyanskiy, Artem Dolotov, Iryna Pliss, Yevgen Viktorov

***Abstract**: in the paper new non-conventional growing neural network is proposed. It coincides with the Cascade-Correlation Learning Architecture structurally, but uses ortho-neurons as basic structure units, which can be adjusted using linear tuning procedures. As compared with conventional approximating neural networks proposed approach allows significantly to reduce time required for weight coefficients adjustment and the training dataset size.*

***Keywords**: orthogonal activation functions, ortho-synapse, ortho-neuron, cascade orthogonal neural network.*

***ACM Classification Keywords**: I.2.6 Learning – Connectionism and neural nets*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Nowadays artificial neural networks (ANNs) are widely applied for solving large class of problems related to processing information given as time-series or numerical data arrays generated by nonstationary stochastic or chaotic systems. The most attractive properties of the ANNs are their approximating possibilities and learning capabilities.

Traditionally by learning we understand a process of the net's synaptic weights adjustment accordingly to selected optimization procedure of accepted learning criterion [1, 2]. Quality of received result can be improved not only by adjusting weight coefficients but also by adjusting of the neural network architecture (the number of nodes). There are two basic approaches of the neural network architecture adjustment: 1) 'constructive approach' [3 - 5] – starts with simple architecture and gradually adds new nodes during learning; 2) 'destructive approach' [6 - 8] – starts with an initially redundant network and simplifies it throughout learning process.

Obviously, constructive approach needs less computational resources and within the bounds of this approach cascade neural networks (CNNs) [9 - 11] can be marked out. The most efficient representative of CNNs is the Cascade-Correlation Learning Architecture (CasCorLA) [9]. This network begins with the simplest architecture which consists of a single neuron. Throughout learning procedure new neurons are added to the network, producing multilayered structure. It is important that during each learning epoch only one neuron of the last cascade is adjusted. All pre-existing neurons process information with "frozen" weights. CasCorLA authors, S.E. Fahlman and C. Lebiere, point out high speed of learning procedure and good approximation properties of this network. But it should be observed that elementary Rosenblatt perceptrons with hyperbolic tangent activation functions are used in this architecture as nodes. Thus output signal of each neuron is nonlinearly depended from its weight coefficients. Therefore it is necessary to use gradient learning methods such as delta-rule or its modifications, and speed of operation optimization becomes impossible. In connection with the above it seems to be reasonable to synthesize cascade architecture based on elementary nodes with linear dependence of output signal from synaptic weights. It allows to increase speed of synaptic weight adjustment and to reduce minimally required size of training set.

Ortho-neuron

Within the variety of the functional structures, used for approximation of nonlinear dependences, orthogonal polynomials [12, 13] deserve a special attention. They possess quite attractive properties, which make it possible to reduce computational complexity and increase precision of received results. At the present time we can observe more and more realizations of the orthogonal polynomials theory in the field of neural networks [14 - 21], which demonstrate impressive effectiveness.

Elementary one-dimensional system described in “input-output” space of some unknown functional dependence $y(x)$ can be expressed by the following sum:

$$\hat{y} = \hat{f}(x) = w_0\varphi_0(x) + w_1\varphi_1(x) + ... + w_h\varphi_h(x) = \sum_{j=0}^{h} w_j\varphi_j(x), \tag{1}$$

where x and $y(x)$ are input and output variables of the estimated process correspondingly, $\varphi_j(x)$ – orthogonal polynomial of the *j*-th order (*j* = 0, 1, 2,..., *h*), which possesses the orthogonality property, *j*, *q*, – nonnegative integer numbers, *k* = 1, 2,…,*N* – current discrete time or the ordinal number of an element in the sampling.

Equation (1) can be realized by the elementary scheme shown at the Fig. 1 and called us the ortho-synapse [22].

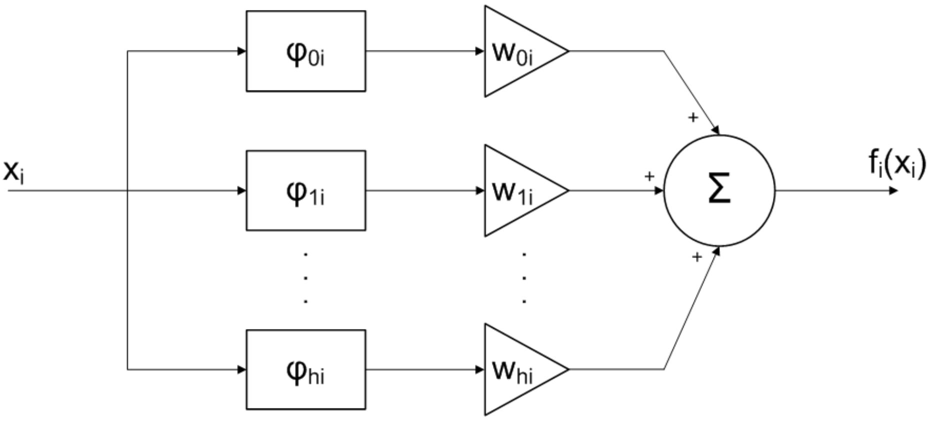


Figure 1. The ortho-synapse – OS$_i$

At the Fig. 1 x_i is the *i*-th (*i* = 1, 2,..., *n*) component of the multidimensional input signal $x = (x_1, x_2, ..., x_n)^T$, w_{ji} (*j* = 1, 2,…,*h*) – synaptic weights which should be determined, $f_i(x_i)$ – output signal of the ortho-synapse, which can be expressed in the form

$$f_i(x_i) = \sum_{j=0}^{h} w_{ji}\varphi_{ji}(x_i). \tag{2}$$

Different systems of orthogonal polynomials (Chebyshev, Hermite, Laguerre, Legendre, etc) can be used as the activation functions of ortho-synapse. Particular system of functions can be chosen accordingly to the specificity of the solved problem. If the input data is normalized on the hypercube [-1, 1]^n, the system of Legendre polynomials orthogonal on the interval [-1, 1] with weight $\gamma(x) = 1$ can be used:

$$\varphi_{ji}^{L}(x_i) = 2^{-j}\sum_{p=0}^{[j/2]}(-1)^p \frac{(2j-2p)!}{p!(j-p)!(j-2p)!}x_i^{j-2p}, \tag{3}$$

where [•] – is the integer part of a number.

Also to simplify calculations we can exploit recurrent formula

$$\varphi_{j+1,i}^{L}(x_i) = \frac{2j+1}{j+1}x_i P_j(x_i) - \frac{j}{j+1}P_{j-1}(x_i). \tag{4}$$

System of Legendre polynomials is the best suited for the case when we exactly know interval of data changes before network construction. This is quite common situation as well as an opposite one. For the other case the following system of Hermite orthogonal polynomials can be used:

$$H_l(u) = l!\sum_{p=1}^{[l/2]}(-1^p)\frac{(2u)^{l-2p}}{p!(l-2p)!}. \tag{5}$$

This system is orthogonal on $(-\infty,+\infty)$ with weight function $h(u) = e^{-u^2}$ and gives us a possibility to decrease influence of the data lying far from origin.

Also it can be readily seen that ortho-synapse has the same architecture like a nonlinear synapse of the neo-fuzzy-neuron [23 - 25], but provides smooth polynomial approximation, based on orthogonal polynomials, instead of piecewise-linear approximation.

We use ortho-synapse as a structural block for the architecture called us ortho-neuron and shown at the Fig. 2.

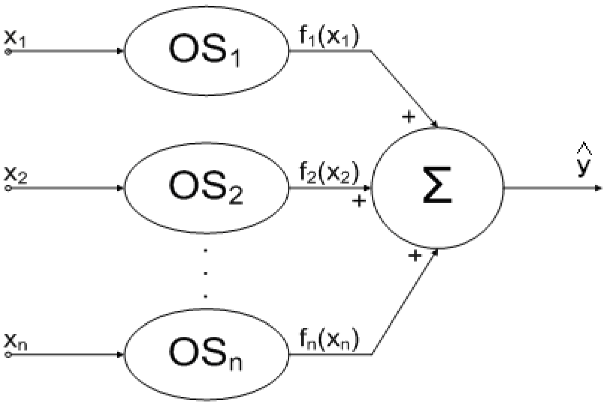


Figure 2. Ortho-neuron – ON

Ortho-neuron which has the same architecture like a neo-fuzzy-neuron realizes the mapping

$$\hat{y} = \sum_{i-1}^{n} f_i(x_i) = \sum_{i=1}^{n}\sum_{j=0}^{h} w_{ji}\varphi_{ji}(x_i), \tag{6}$$

and provides high precision of approximation and extrapolation of significantly nonlinear nonstationary signals and processes [16, 17, 19 - 22]. But in what follows ortho-neuron will be used as the elementary node of the architecture called us the Cascade Orthogonal Neural Network (CONN).

The Cascade Orthogonal Neural Network Architecture

The CONN architecture is shown at the Fig. 3 and mapping, which it realizes, have the following form:

- first cascade neuron :

$$\hat{y}_1 == \sum_{i=1}^{n}\sum_{j=0}^{h} w_{ji}^{[1]}\varphi_{ji}(x_i); \tag{7}$$

- second cascade neuron

$$\hat{y}_2 == \sum_{i=1}^{n}\sum_{j=0}^{h} w_{ji}^{[2]}\varphi_{ji}(x_i) + \sum_{j=0}^{h} w_{j,n+1}^{[2]}\varphi_{j,n+1}(\hat{y}_1); \tag{8}$$

- third cascade neuron

$$\hat{y}_2 == \sum_{i=1}^{n}\sum_{j=0}^{h} w_{ji}^{[3]}\varphi_{ji}(x_i) + \sum_{j=0}^{h} w_{j,n+1}^{[3]}\varphi_{j,n+1}(\hat{y}_1) + \sum_{j=0}^{h} w_{j,n+2}^{[3]}\varphi_{j,n+2}(\hat{y}_2); \tag{9}$$

- *m*-th cascade neuron

$$\hat{y}_m == \sum_{i=1}^{n}\sum_{j=0}^{h} w_{ji}^{[m]}\varphi_{ji}(x_i) + \sum_{l=n+1}^{n+m-1}\sum_{j=0}^{h} w_{jl}^{[m]}\varphi_{jl}^{[m]}(\hat{y}_{l-n}). \tag{10}$$

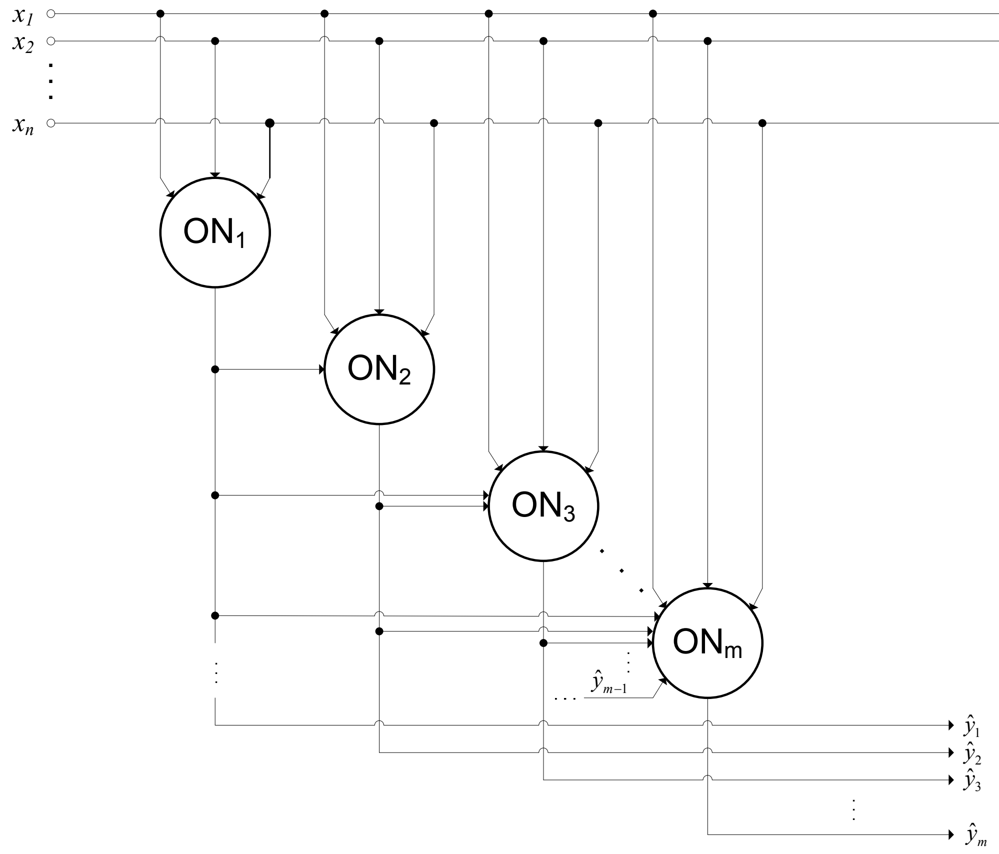


Figure 3. The Cascade Orthogonal Neural Network

Thus the CONN contains $(h+1)(n+\sum_{l=n+1}^{n+m-1} l)$ adjustable parameters and it is important that all of them are linearly included in the definition (10).

Let us define vector $(h+1)(n+m-1)\times 1$ of orthogonal polynomials of the *m*-th ortho-neuron $\varphi^{[m]} = (\varphi_{01}(x_1), \varphi_{11}(x_1), ..., \varphi_{h1}(x_1), \varphi_{02}(x_2), ..., \varphi_{h2}(x_2), ..., \varphi_{ji}(x_i), ..., \varphi_{hn}(x_n), \varphi_{0,n+1}(\hat{y}_1), ..., \varphi_{h,n+1}(\hat{y}_1), ..., \varphi_{h,n+m-1}(\hat{y}_{m-1}))^T$ and corresponding vector of synaptic weights $w^{[m]} = (w_{01}^{[m]}, w_{11}^{[m]}, ..., w_{h1}^{[m]}, w_{02}^{[m]}, ..., w_{h2}^{[m]}, ..., w_{ji}^{[m]}, ..., w_{hn}^{[m]}, w_{0,n+1}^{[m]}, ..., w_{h,n+1}^{[m]}, ..., w_{h,n+m-1}^{[m]})^T$ which has the same dimensionality. Then we can represent expression (10) in the vector notation:

$$\hat{y}_m = w^{[m]T}\varphi^{[m]}. \tag{11}$$

The Cascade Orthogonal Neural Network Learning

The Cascade Orthogonal Neural Network learning is performed in the batch mode using entire training set $x(1), y(1); x(2), y(2);...; x(k), y(k);...; x(N), y(N)$. At the beginning a set of orthogonal functions values $\varphi^{[1]}(1), \varphi^{[1]}(2),..., \varphi^{[1]}(N)$ is calculated for each training sample, so we obtain a sequence of vectors $(h+1)n \times 1$. Then using direct minimization of the learning criterion

$$E_N^{[1]} = \frac{1}{2}\sum_{k=1}^{N} e_1(k)^2 = \frac{1}{2}\sum_{k=1}^{N}(y(k) - \hat{y}_1(k))^2, \tag{12}$$

vector of synaptic weights can be evaluated

$$w^{[1]}(N) = \left(\sum_{k=1}^{N} \varphi^{[1]}(k)\varphi^{[1]T}(k)\right)^{+} \sum_{k=1}^{N} \varphi^{[1]}(k)y(k) = P^{[1]}(N)\sum_{k=1}^{N}\varphi^{[1]}(k)y(k). \tag{13}$$

If dimension of this vector is sufficiently large it is suitable to use procedure (13) in the form of recursive least squares method with sequential training samples processing:

$$\begin{cases} w^{[1]}(k+1) = w^{[1]}(k) + \dfrac{P^{[1]}(k)(y(k+1) - w^{[1]T}(k)\varphi^{[1]}(k+1))}{1+\varphi^{[1]T}(k+1)P^{[1]}(k)\varphi^{[1]}(k+1)}\varphi^{[1]}(k+1), \\ P^{[1]}(k+1) = P^{[1]}(k) - \dfrac{P^{[1]}(k)\varphi^{[1]}(k+1)\varphi^{[1]T}(k+1)P^{[1]}(k)}{1+\varphi^{[1]T}(k+1)P^{[1]}(k)\varphi^{[1]}(k+1)} \end{cases}. \tag{14}$$

It is necessary to notice that using procedures (13), (14) for adjusting weight coefficients essentially reduces learning time in comparison with gradient algorithms underlying delta-rule. Also orthogonality of activation functions ensures numerical stability during matrixes inversion.

After first cascade learning completion, synaptic weights of the neuron ON$_1$ become 'frozen' and second cascade of network consisting from a single neuron ON$_2$ is generated. It has one additional input for the output signal of the first cascade. Then procedures (13), (14) again applied for adjusting vector of weight coefficients $w^{[2]}$, which dimensionality is $(h+1)(n+1)\times 1$.

The neural network growing process (increasing quantity of cascades) continues until we obtain required precision of the solved problem's solution, and for the adjusting weight coefficients of the last (*m*-th) cascade following expression are used:

$$w^{[m]}(N) = \left(\sum_{k=1}^{N} \varphi^{[m]}(k)\varphi^{[m]T}(k)\right)^{+} \sum_{k=1}^{N} \varphi^{[m]}(k)y(k) = P^{[m]}(N)\sum_{k=1}^{N}\varphi^{[m]}(k)y(k) \tag{15}$$

or

$$\begin{cases} w^{[m]}(k+1) = w^{[m]}(k) + \dfrac{P^{[m]}(k)(y(k+1) - w^{[m]T}(k)\varphi^{[m]}(k+1))}{1+\varphi^{[m]T}(k+1)P^{[m]}(k)\varphi^{[m]}(k+1)}\varphi^{[m]}(k+1), \\ P^{[m]}(k+1) = P^{[m]}(k) - \dfrac{P^{[m]}(k)\varphi^{[m]}(k+1)\varphi^{[m]T}(k+1)P^{[m]}(k)}{1+\varphi^{[m]T}(k+1)P^{[m]}(k)\varphi^{[m]}(k+1)} \end{cases}, \tag{16}$$

where vectors $w^{[m]}$ and $\varphi^{[m]}$ have dimensionality $(h+1)(n+m-1)\times 1$.

The main disadvantage of CasCorLA is the necessity of the batch mode learning usage, when all training set should be given priori. CONN can be trained in on-line mode, because of algorithm (16) possesses maximal possible squared rate of convergence. In this case at the first step architecture consisting of *m* cascades is generated. Each cascade trains using proper algorithm. Since outputs of the previous ortho-neurons become additional inputs for the *m*-th cascade, algorithm (16) realizes recurrent method of the prediction error [26], well known in the theory of adaptive identification. Changing cascades quantity during learning process also can be easily performed.

Simulation Results

We have applied proposed Cascade Orthogonal Neural Network to solve 'breast cancer in Wisconsin' benchmark classification problem.

Dataset containing 699 points have been used for this purpose (ftp://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/cancer1/datacum). 16 points had parameters with missed values so they have been eliminated from the dataset and remaining 683 points have been separated on training set – 478 points (70%) and test set – 205 points (30%).

Each point has 9-dimensional feature vector and 1 class parameter which should be determined and identifies either benign or malignant tumor have current examined patient. Features values have been normalized on interval [-1; 1].

There are 3 optional parameters must be specified to start CONN constructive algorithm: 1) type of the orthogonal polynomials system in each ortho-synapse; 2) quantity of orthogonal polynomials in each ortho-synapse; 3) maximal number of cascades. Since input data have been normalized on interval [-1; 1], we choose systems of 1 type Chebyshev orthogonal polynomials as activation functions for each ortho-synapse to avoid unlimited weight values growth. Previous experiments have shown that optimal ortho-synapse dimensionality is 3-4 polynomials per input, so these values have been chosen for experiment. For avoiding generalization loss maximal number of cascades has been limited by 10.

For comparison the same classification problem has been solved using the Cascade-Correlation Learning Architecture and Multilayered Perceptron. The CasCorLA had 8 cascades and each of them utilized gradient minimization for adjusting weight coefficients. MLP had 9x15x1 architecture and tuned with Levenberg-Marquard minimization procedure during only 20 epochs. Increasing number of epochs (in this case) results in generalization loss. Obtained results of classifications can be found in table 1.

When output signal be found within the range [0.3; 0.7] it is lesser probability that classification were correct. We quantify and marked out such classified samples as points outside the 'belief zone'.

Table 1 – Classification results for different architectures

ANN Architecture	Accuracy on training set / Points outside the 'belief zone'	Accuracy on testing set / Points outside the 'belief zone'
CONN	99,8% / 1	98% / 4
CasCorLA	95% / 46	99% / 15
MLP	99,2% / 4	98,5% / 3

We can see that CONN shows quite good results of classification, comparable with MLP's, and much better than CasCorLA's ones. Therewith CONN's learning procedure takes considerably lesser time and computational recources than backpropogation or Levengerg-Marquardt minimization Also using CONN architecture makes possible to avoid two significant disadvantages of CasCorLA and MLP: unrepeatability of obtained results and necessity to use first- or second-order derivative procedures.

Conclusion

The Cascade Orthogonal Neural Network is proposed. It differs from its prototype, Cascade-Correlation Learning Architecture, in increased speed of operation, numerical stability and real-time processing possibility. Theoretical justification and experiment results confirm the efficiency of developed approach.

Bibliography

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

[2] Haykin S. Neural Networks. A Comprehensive Foundation. Upper Saddle River, N.J.: Prentice Hall, Inc., 1999.

[3] Platt J. A resource allocating network for function interpolation. Neural Computation, 3, 1991. P.213-225.

[4] Nag A., Ghosh J. Flexible resource allocating network for noisy data. In: Proc. SPIE Conf. on Applications and Science of Computational Intelligence, SPIE Proc. 1998. P.551-559.

[5] Yingwei L., Sundararajan N., Saratchandran P. Performance evaluation of a sequential minimal radial basis function (RBF) neural network learning algorithm. IEEE Trans. on Neural Networks, 9, 1998. P.308-318.

[6] Cun Y.L., Denker J.S., Solla S.A. Optimal Brain Damage. In: Advances in Neural Information Processing Systems, 2, 1990. P.598-605.

[7] Hassibi B. Stork D.G. Second-order derivatives for network pruning: Optimal brain surgeon. In: Advances in Neural Information Processing Systems. Ed. Hanson et al. 1993. P.164-171.

[8] Prechelt L. Connection pruning with static and adaptive pruning schedules. Neurocomputing, 16, 1997. P.49-61.

[9] Fahlman S.E., Lebiere C. The cascade-correlation learning architecture. Advances in Neural Information Processing Systems. Ed. D.S. Touretzky. San Mateo, CA: Morgan Kaufman, 1990. P.524-532.

[10] Schalkoff R.J. Artificial Neural Networks. N.Y.: The McGraw-Hill Comp., Inc., 1997..

[11] Avedjan E.D., Barkan G.V., Levin I.K. Cascade Neural Networks. Avtomatika i Telemekhanika, 3, 1999. P.38-55.

[12] Bateman H., Erdelyi A. Higher Transcendental Functions. Vol.2. N.Y.: McGraw-Hill Comp., Inc., 1953.

[13] Graupe D. Identification of Systems. Huntington, N.Y.: Robert E. Kreiger Publishing Comp., Inc., 1976.

[14] Scott I., Mulgrew B. Orthonormal function neural network for nonlinear system modeling. In: Proceedings of the International Conference on Neural Networks (ICNN-96), June 1996.

[15] Patra J.C., Kot A.C. Nonlinear dynamic system identification using Chebyshev functional link artificial neural network. IEEE Trans. on System, Man and Cybernetics. Part B, 32, 2002. P.505-511.

[16] Bodyanskiy Ye., Kolodyazhniy V., Slipchenko O. Artificial neural network with orthogonal activation functions for dynamic system identification. Synergies between Information Processing and Automation. Ed. O. Sawodny and P. Scharff – Aachen: Shaker Verlag, 2004. P.24-30.

[17] Bodyanskiy Ye., Kolodyazhniy V., Slipchenko O. Structural and synaptic adaptation in the artificial neural networks with orthogonal activation functions. Sci. Proc. of Riga Technical University. Comp. Sci., Inf. Technology and Management Sci, 20, 2004. P.69-76.

[18] Liying M., Khorasani K. Constructive feedforward neural network using Hermite polynomial activation functions. IEEE Trans. on Neural Networks, 4, 2005. P.821-833.

[19] Bodyanskiy Ye., Pliss I., Slipchenko O. Growing neural networks based on orthogonal activation functions. Proc. XII-th Int. Conf. “Knowledge – Dialog – Solution”. Varna, 2006. P84-89.

[20] Bodyanskiy Ye., Slipchenko O. Ontogenic neural networks using orthogonal activation functions. Prace naukowe Akademii Ekonomicznej we Wroclawiu, 21, 2006. P.13-20.

[21] Bodyanskiy Ye., Pliss I., Slipchenko O. Growing neural network using nonconventional activation functions. Int. J. Information Theories & Applications, 14, 2007. P.275-281.

[22] Bodyanskiy Ye., Viktorov Ye., Slipchenko O. Orthosynapse, ortho-neurons, and neuropredictor based on them. Systemi obrobki informacii. Issue 4(62), 2007. P.139-143.

[23] Yamakawa T., Uchino E., Miki T., Kusanagi H. A neo fuzzy neuron and its applications to system identification and prediction of the system behavior. Proc. 2-nd Int.Conf. on Fuzzy Logic and Neural Networks “LIZUKA-92”. Lizuka, Japan, 1992. P.477-483.

[24] Uchino E., Yamakawa T. Soft computing based signal prediction, restoration and filtering. Intelligent Hybrid Systems: Fuzzy Logic, Neural Networks and Genetic Algorithms. Ed. Da Ruan. Boston: Kluwer Academic Publisher, 1997. P.331-349.

[25] Miki T., Yamakawa T. Analog implementation of neo-fuzzy neuron and its on-board learning. Computational Intelligence and Applications. Ed. N.E. Mastorakis. Piraeus: WSES Press, 1999.P.144-149.

[26] Ljung L. System Identification: Theory for the User. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1987.

Authors' Information

***Yevgeniy Bodyanskiy** – Doctor of Technical Sciences, Professor of Artificial Intelligence Department and Scientific Head of the Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, Lenina av. 14, Kharkiv, 61166, Ukraine, Tel +380577021890, e-mail: bodya@kture.kharkov.ua*

***Artem Dolotov** – Ph.D. Student, Kharkiv National University of Radio Electronics, Lenin Av., 14, Kharkiv, 61166, Ukraine, Tel +380508361789, e-mail: artem.dolotov@gmail.com*

***Iryna Pliss** – Candidate of Technical Sciences (equivalent Ph.D.), Senior Researcher, Leading Researcher of the Control Systems Research Laboratory, Kharkiv National University of Radio Electronic, Lenina av. 14, Kharkiv, 61166, Ukraine, Tel +380577021890, e-mail: pliss@kture.kharkov.ua*

***Yevgen Viktorov** - Ph.D. Student, Kharkiv National University of Radio Electronics, Lenin Av., 14, Kharkiv, 61166, Ukraine, Tel +380681613429, e-mail: yevgen.viktorov@gmail.com*

OFFLINE HANDWRITING RECOGNITION USING GENETIC ALGORITHM

Shashank Mathur, Vaibhav Aggarwal, Himanshu Joshi, Anil Ahlawat

***Abstract**: In this paper, a new method for offline handwriting recognition is presented. A robust algorithm for handwriting segmentation has been described here with the help of which individual characters can be segmented from a word selected from a paragraph of handwritten text image which is given as input to the module. Then each of the segmented characters are converted into column vectors of 625 values that are later fed into the advanced neural network setup that has been designed in the form of text files. The networks has been designed with quadruple layered neural network with 625 input and 26 output neurons each corresponding to a character from a-z, the outputs of all the four networks is fed into the genetic algorithm which has been developed using the concepts of correlation, with the help of this the overall network is optimized with the help of genetic algorithm thus providing us with recognized outputs with great efficiency of 71%.*

***Keywords**: Handwriting Recognition, Segementation, Artificial Neural Networks, Genetic Algorithm.*

***ACM Classification Keywords**: I.2 Artificial Intelligence, I.4 Image processing and Computer Vision, I.5 Pattern Recognition.*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Over the years, computerization has taken over large number of operations that had been taken care of manually, one such example is of offline cursive handwriting recognition, which is the ability of a computer to receive and interpret intelligible handwriting input present in the form of scanned images [I]. The various methods for character recognition have already been published[II] but the method presented here is advanced than those methods since cursive handwriting can be recognized with the help of a combination of artificial neural networks and genetic algorithm, this becomes the primary advantage of the method over other existing methods. The methodology here has been developed with four multilayer artificial neural networks with Levenberg-Marquardt back propagation algorithm along with genetic algorithm unlike few published methods that use a multilayer feed-forward neural network[III] thus providing an efficient output as compared to the previously published works.

The recent spurt in the advancement in handwriting recognition has provided publications one of which discussed here is recognition of text written in 'Oriya', a traditional south-eastern Indian language [IV] but do not involve any combinations of artificial neural networks and optimization techniques such as genetic algorithms which lead to lower efficiency in recognition as compared to the ones that our approach here presents. Several areas have seen application of neural networks such as the Processing of Verbs and Nouns [V], Face Detection [VI] and Real-time Face Detection [VII]. The application of genetic algorithms in various areas like initial population generation methods [VIII], mooring pattern optimization [IX], substitution ciphers [X] and designing of reverse logistic networks [XI] has proved its advancement over its predecessors. The handwriting recognition model described here works at three stages, segmentation of the handwritten text, recognition of segmented characters with the help of artificial neural networks and lastly selecting the best solution from the four artificial neural network outputs with the help of genetic algorithm.

The cursive handwriting recognition is carried out with the help of artificial neural networks, which is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionistic approach to computation and which has the capability of being adaptive and thus can change its structure based on the information provided to it. The artificial neural networks made in this case contain 625 input neurons, 3 hidden layers and 26 output neurons. The recognition model involves the usage of artificial neural networks which form the basis of recognition after the network has been exhaustively trained to recognize different types of handwritings. This is achieved by training the four artificial neural networks using a highly efficient supervised learning algorithm, Levenberg-Marquardt back propagation algorithm [XII,XIII] unlike other published methods like fast learning method for neural networks based on sensitivity analysis [XIV].

In general, a single neural network is used for recognition purposes.

The method described does not have any limitations on the type of font or text that is being taken in as input [XV], any handwritten information can be converted into editable textual information. Plethora of the character recognition methods have been published previously [XVI] but they do not incorporate advanced optimization techniques such as the ones provided by the genetic algorithms which have provided high efficiency in the method described here. Only a few handwritten character recognition papers for applications like form registration [XVII] have been published till date with optimization with the help of genetic algorithms but we advance by application of both artificial neural networks and genetic algorithm to cursive handwriting.

The purpose of this paper is to present a new methodology for cursive handwriting recognition using artificial neural networks and genetic algorithm. Handwriting segmentation is carried out with the help of a novel algorithm which is capable of extracting handwriting words from the handwritten text given as input in the form of an image and carries out segmentation of the selected word to generate vectors for individual characters of a word. These vectors are given as an input to four neural networks. The four neural networks generate four sets of final outputs, each out of which the genetic algorithm chooses the fittest set of values to provide the user with a highly efficient handwriting recognition model. The usage of Levenberg-Marquardt algorithm along with genetic algorithm assures highly efficient results.

Handwriting Segmentation Algorithm

The handwritten document is scanned and taken as an input to obtain individual characters which are written in a text file and are later read back and are passed as inputs to the four Artificial Neural Networks. In the algorithm, the scanned gray scale image is read into an image matrix which is converted into a monochromatic image matrix which pixel values of 0 for black points and 255 for white points. Then row wise searching is started from the point (0,0) to find out the first black point. This is the assumed top point of the first word of the handwritten text that has been inputted. This point is referred to the "Upper Point". After the upper point is found, all the black pixels that are connected to this pixel are given a value of 999.

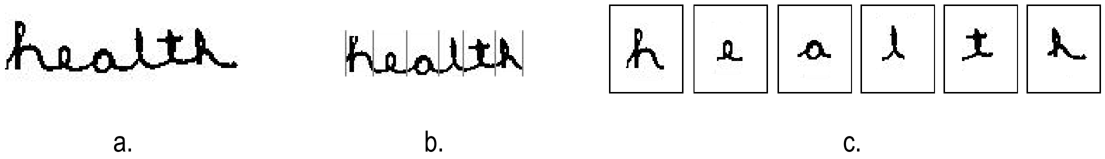


Fig.1(a-c). Different stages of handwriting segmentation.

Once this step is complete, then all the characters linked to that word have a value of 999 in the matrix under consideration. After finding all the connected points, a row wise search starts from the bottom to the top to find the first 999 value. This value corresponds to the "Lowest Point". After this point is obtained, the area between the top and bottom point is searched on the left to check if any word has been missed on the left. In case another word is present on the left, then the new top point is obtained and the bottom point is found again else the procedure continues. After this is carried out, the left and right points of the word are found out by column wise searches. After the four points, the Upper, Lower, Right and Left point are found out, the word can be extracted and stored in a different matrix. This step is followed by marking of intersection points between various characters in the cursive handwriting. The word is searched for the number of cuts in a column wise manner. All the cases in which number of cuts is one and has an edge on either left or its right are marked and then all the successive markings are averaged to provide one optimum point through which a cut is marked with gray value of 0.5. After all the cuts are marked, in a loop all the characters are written into a file which is later read by the neural network to recognize the character.

Architecture of Artificial Neural Network used

The four artificial neural networks used consist of an input layer, three hidden layers and an output layer for each of the individual networks. The input layers takes the input from the image segmentation algorithm thus has 625 input neurons. The number of hidden layer neurons is as shown in the Table 1. The output layer consists of 26 neurons; this is due to the fact that there are 26 characters to be identified. Thus each output neuron corresponds to every character. Four artificial neural networks have been employed for the character recognition. The properties of each of the four artificial neural networks are designed using different parameters as shown in Table 1. The outputs of these four artificial neural networks are fed into the genetic algorithm with chooses the fittest and the best solution and provides us with the recognized alphabet. Thus providing us with an overall high efficiency for the offline handwriting recognition using artificial neural networks and genetic algorithm.

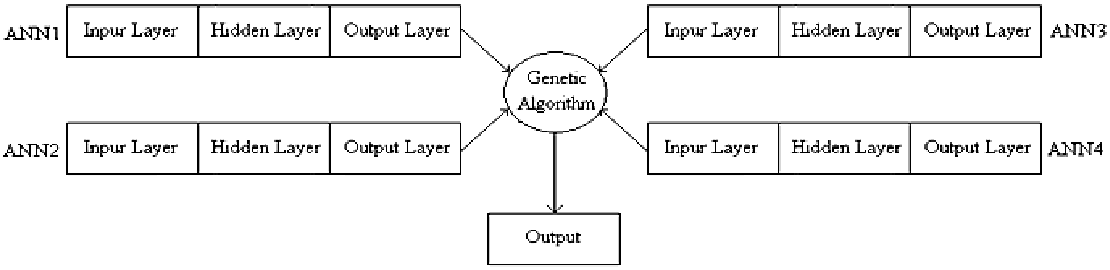


Fig.2. Architecture of neural networks used.

Table 1. List of parameters used for the four artificial neural networks.

Name of Parameter		**Artificial Neural Network 1**	**Artificial Neural Network 2**	**Artificial Neural Network 3**	**Artificial Neural Network 4**
Number of neurons	**Input Layer**	625	625	625	625
	Hidden Layer1	8	8	6	8
	Hidden Layer2	16	14	16	12
	Hidden Layer3	8	8	6	8
	Output Layer	26	26	26	26
Training function(Algorithm)		trainlm (Levenberg-Marquardt algorithm)			
Number of Epochs		5000	4000	3000	2500
Performance Function		Mse (Mean square error)		Sse (Sum squared error)	
Training Goal		0.01	0.05		0.012
Memory Reduction Parameter		50			
Transfer Function		Purelin (Pure Linear)			

Segmented handwritten character recognition method

A column vector is generated by the Image Segmentation consisting of 625 elements, which is saved in a ".txt" file. This ".txt" file is read and the elements are fed as an input to each of the four neural networks. These neural networks read all the weights and biases values that were saved in another files during the training process. Corresponding to the input, output is generated at the output layer. A "1" is set at the index of the characters that has been recognized. There can be more than one character that could be recognized based on the noise in the input of the neural network.

Methodology used for segemented handwriting recognition

The following method is applied for handwriting recognition:

1. Provide initial inputs of sample handwritten letters to train the four artificial neural networks with different parameters (Table 1.).
2. Start the process of training the networks with different sets of letters.
3. Store the weight matrices and bias values obtained after training as files.
4. Read the file containing the input matrix.
5. Feed this as the input to all the four neural networks.
6. Send the outputs of the neural networks to the Genetic Algorithm.

As we have developed four artificial neural networks we need to select the best and fittest solution from the set of outputs obtained. This is carried out with the help of Genetic Algorithm which is applied such that it accepts the outputs of the four artificial neural networks. There are cases when the neural networks recognize more than one character. For that case we have created four neural networks, which have different training parameters and also trained differently from each other. We pass the input column matrix in all the four neural networks and get the output. The outputs are sent to the Genetic Algorithm and hence making the initial population for the Algorithm. We then calculate the number of "1's" at the output for each neural network.

The one with the minimum number of "1's" is selected. The characters corresponding to the index of those ones are shortlisted. These shortlisted indexes are sent to the fitness function where the correlation coefficients of the indexes are calculated. We specify the threshold value. Below this value of the correlation coefficients the indexes are discarded. The correlation coefficients above this value are selected and this forms the new generation of the indexes. The steps are repeated again and again for different training sets. These results are taken and the character which has the maximum correlation with the input set is selected and shown as the output. With the combination of four artificial neural networks and their optimization using genetic algorithm, the efficiency of the offline handwriting recognition model increases to a great extent delivering accurate results of recognition.

The following method is applied for applying genetic algorithm to the outputs of the four artificial neural networks:

1. **Initialization**: Select the output of the neural network with the indexes comprising of "1's". This corresponds to the initial population for the Genetic Algorithm.
2. **Selection**: Select the indexes from the neural network that has minimum number of "1's".
3. **Fitness function**: Compute the correlation coefficients of the selected indexes.
4. **Mutation & Crossover**: For the correlation coefficients less than the threshold value 0.50 repeat the step of the fitness function for a different training set. Discard the indexes that have coefficient values less than 0.3.
5. **Evaluation**: Select the index which has the maximum correlation coefficient with the input matrix.
6. Output the selected character.

Results and Discussions

In order to check the accuracy of the individual modules of the approach discussed here, handwriting samples were collected from various people and segmentation was carried out after which the four neural networks were trained and various characters were subjected to the neural networks and genetic algorithm to see how well the recognition process is carried out for the various different handwritten scanned input characters.

The testing phase was to check the efficiency of recognition and the recognition rates of the individual four neural networks. Once the network weights and biases are initialized, the network is ready for training. The network can be trained for function approximation (nonlinear regression), pattern association, or pattern classification. The training process requires a set of examples of proper network behavior — network inputs p and target outputs t.

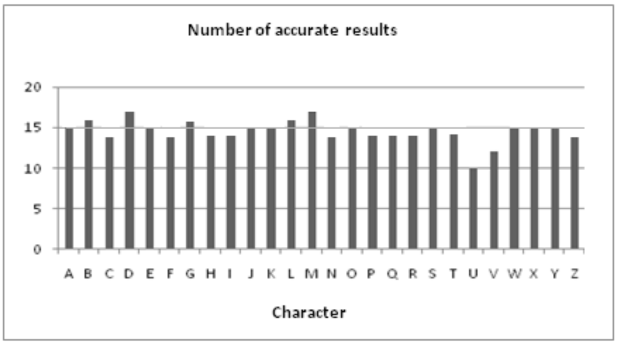


Fig.4. handwriting segmentation testing.

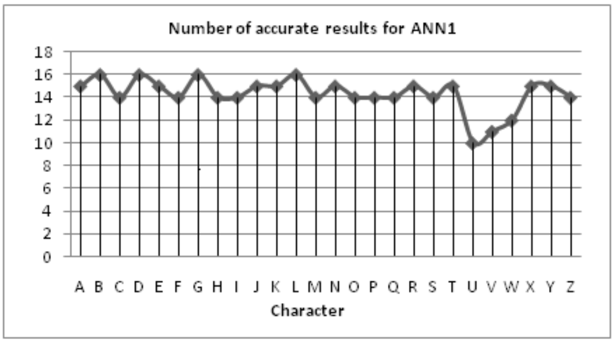


Fig.5. Graph obtained for testing of artificial neural network 1.

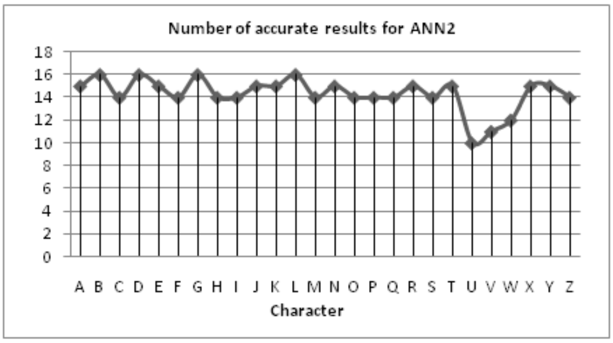


Fig.6. Graph obtained for testing of artificial neural network 2.

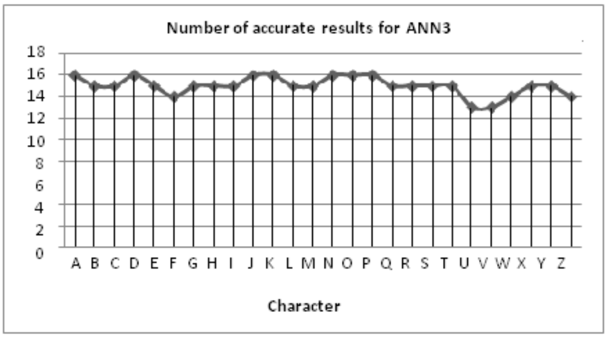


Fig.7. Graph obtained for testing of artificial neural network 3.

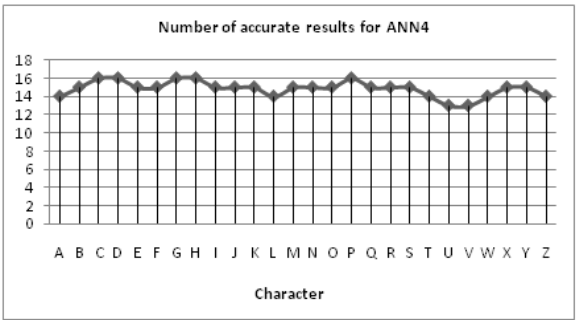


Fig.8. Graph obtained for testing of artificial neural network 4.

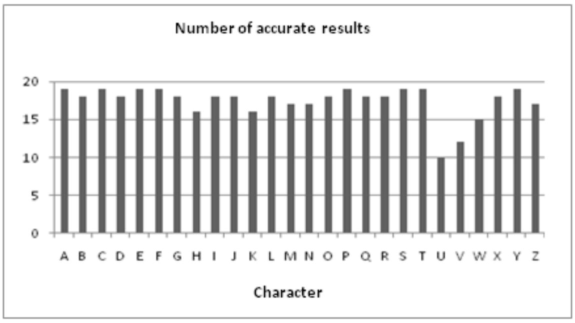


Fig.9. Artificial neural network and genetic algorithm testing.

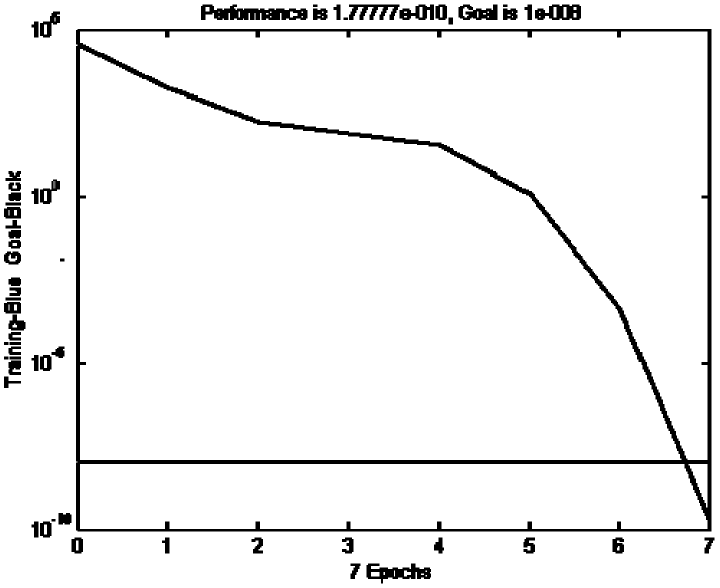


Fig. 10. Error graph

During training the weights and biases of the network are iteratively adjusted to minimize the network performance function. The gradient is determined using a technique called back propagation, which involves performing computations backward through the network. When the networks are trained the weights and the biases of each neural networks are saved into ".txt" files. The following error graph was obtained after training the neural network with trainlm. The algorithms have been specially designed for handwriting recognition using the various mathematical logics for image segmentation and neural networks for recognition.

The three above mentioned algorithms for handwriting segmentation, neural networks and optimization using genetic algorithms were programmed and trained with the help of 260 handwritten character samples, 10 samples per letter on a Pentium 4 (3.4 GHz), 2GB RAM and MATLAB 7.0 and tested individually and then the integration of the three algorithms after exhaustive testing has provided us with a highly efficient handwriting recognition with the help of which a human handwritten text can be converted into a textual format on a computer thus providing the user flexibility to edit the text. The algorithms were tested with 200 handwritten samples out of which 142 samples were correctly recognized providing us with an overall efficiency of 71.0%.

Conclusion

The paper has presented a new method of handwriting recognition using a unique and robust combination of artificial neural networks and genetic algorithms. On programming and testing the modules a very high efficiency has been noted. The handwriting recognition is indeed a tough task which can be easily done with the help of the methodology described here. A high efficiency reflects the accuracy of segmentation as well as the recognition using the neural network that has been optimized by genetic algorithms. The concept here has abridged handwriting recognition completely with artificial intelligence after the application of both artificial neural networks and genetic algorithms.

Bibliography

[I.] Plamondon Rejean, Srihari Sargur N. 2000, "On-line and Off-line Handwriting Recognition : A comprehensive survey", IEEE Transactions on pattern analysis and machine intelligence, Vol.22. No.1,pp.63-84.

[II.] Prema K. V., Reddy N V Subba 2002, "Two-tier architecture for unconstrained handwritten character recognition",Sadhna,Vol. 27, part 5, pp. 585-594.

[III.] Ngom Alioune, Stojmenovic´ Ivan, Milutinovic´Veljko 2001, "STRIP- A Strip Based Neural-Network Growth Algorithm for Learning Multiple-Valued Functions",IEEE Transactions on Neural Networks, Vol. 12, No.2, pp. 212-227.

[IV.] Tripathy N. and Pal U. 2006,"Handwriting segmentation of constrained Oriya text", Sadhna, Vol. 31, Part 6, pp. 755-769.

[V.] Cangleosi Angelo, Parisi Domenico 2004, The Processing of Verbs and Nouns in Neural Networks: Insights from Synthetic Brain Imaging, Brain and Language, 89(2), pp. 401-408.

[VI.] Rowley Henry A.,Baluja Shumeet, Kanade Takeo 1998, "Neural Network based Face Detection", IEEE transactions on Pattern Analysis and Machine Intelligence.

[VII.] Curran Kevin, Li Xeulong, Caughley Neil Mc 2005, The Use of Neural Networks in Real-time Face Detection, Journal of Computer Sciences, Vol. 1,pp. 47-62.

[VIII.] Hill Raymond R. 1999, "A Monte carlo study of genetic algorithm intial population generation methods", Proceedings of winter simulation conference,pp. 543-547.

[IX.] Carbono A. J. Juvinao, Menezes Ivan F. M. ,Martha Luiz Fernando 2005,"Mooring parttern optimization using Genetic Algorithms", 6th World Congress of Structural and Multidisciplinary optimization.

[X.] Verma A. K., Dave Mayank, Joshi R. C. 2007, "Genetic Algorithm and Tabu Search Attack on the Mono-Alphabetic Substitution Cipher in Adhoc Networks", 3(3), pp. 134-137.

[XI.] Schleiffer Ralf, Wollenweber Jens, Sebastian Hans-Juergen, Golm Florian Kapoustia Natasha 2004, Application of Genetic Algorithms for the Design of Large-Scale Reverse Logistic Networks in Europe's Automotive Industry, Proceedings of the 37th Hawaii International Conference on System Sciences.

[XII.] Mishra Deepak, Yadav Abhishek, Ray Sudipta, Kalra Prem K. 2005, "Levenberg-Marquardt Learning Algorithm for Integrate-and-Fire Neuron Model", Neural Information Processing - Letters and Reviews, Vol.9, No.2, pp. 41-21.

[XIII.] Wilson Edward 1994, "Backpropagation Learning for Systems with Discrete-Valued Function, Proceedings of the World Congress on Neural Networks.

[XIV.] Alata Mohanad, Al-Shabi Mohammad 2006, "Text detection and character recognition using fuzzy image processing", Journal of Elec. Engineering, Vol.75, pp. 258-267.

[XV.] Singh Abnikant, Singh Markandey, Ratan Ritwaj, Kumar S. 2005, "Optical Character Recognition for printed Tamil text", Journal of Zhejiang University, 6A(11), pp. 1297-1305.

[XVI.] Deodhare Dipti, Suri NNR Ranga, Amit R. 2005, "Preprocessing and Image Enhancement Algorithms for a Form-based Intelligent Character Recognition System",International Journal of Computer Science and Application, Vol.2 ,No.2, pp. 131-144.

[XVII.] Guijarro-Berdi ˜nas Bertha, Fontenla-Romero Oscar, Alonso-Betanzos Amparo 2006, A very fast Learning Method for Neural Networks Based on Sensitivity Analysis, Journal of Machine Learning Research, Vol. 7, pp. 1159-1182.

Authors' Information

***Shashank Narain Mathur** – Member of ITHEA International Scientific Society, Bachelors of Technology (Computer Science Engineering), Amity School of Engineering and Technology affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India. Email: shashanknarainmathur@yahoo.com*

***Vaibhav Aggarwal** – Bachelors of Technology (Computer Science Engineering), Amity School of Engineering and Technology affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India.
Email: vai_aggarwal@hotmail.com*

***Himanshu Joshi** – Bachelors of Technology (Computer Science Engineering), Amity School of Engineering and Technology affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India.*

***Anil Kumar Ahlawat** – Member of ITHEA International Scientific Society, Department of Computer Science Engineering, Amity School of Engineering and Technology affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India. Email: a_anil2000@yahoo.com*

SIMULTANEOUS CONTROL OF CHAOTIC SYSTEMS USING RBF NETWORKS

Angel Castellanos, Rafael Gonzalo, Ana Martinez

Abstract: *Chaos control is a concept that recently acquiring more attention among the research community, concerning the fields of engineering, physics, chemistry, biology and mathematic. This paper presents a method to simultaneous control of deterministic chaos in several nonlinear dynamical systems. A radial basis function networks (RBFNs) has been used to control chaotic trajectories in the equilibrium points. Such neural network improves results, avoiding those problems that appear in other control methods, being also efficient dealing with a relatively small random dynamical noise.*

Keywords: *Neural Network, Radial basis function, Backpropagation, Chaotic Dynamic Systems, Control Feedback Methods.*

ACM Classification Keywords: *F.1.1 Models of Computation: Self-modifying machines (neural networks); F.1.2 Modes of Computation: Alternation and nondeterminism; G.1.7 Ordinary Differential Equations: Chaotic systems;*

Conference: *The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Nowadays, with the electronics communications growth, signals processing has become a technology of multiple facets. It has passed from implementation of tuned circuits to digital processors of signals. The base of the industry continues being the design and realization of filters to carry out noise elimination con carrier signals of information. Chaos is a special feature of parametric nonlinear dynamical systems. It is usually difficult to accurately predict its future behavior. The chaotic phenomena take place everywhere, so much in natural systems as in mechanisms built by the man. Previous works have been mainly focused in describing and characterizing the chaotic behavior in situations where there is not any intervention. A family of artificial neural networks has gotten good results on the prediction and control of the nonlinear plants [Haykin, S].

Chaotic systems are characterized by their sensitive dependence to small perturbations. An abundance of theorical and experimental research has been developed to capitalize on this fact and utilize it to control chaotic systems by applying very small, appropriately timed perturbations.

The control of chaotic signals is one of the most relevant research areas that have appeared in last years, taking the attention of computer scientists [Hübler A. W.]. Recently they have being proposed ideas and techniques to transform chaotic orbits into desired periodic orbits, using temporarily programmed controls [Chen G. & Dong X]. In 1990, Ott, Gebogi Yorke [Ott E., Grebogi C. & Yorke J. A.] developed methods to control two nonlinear system stabilizing one of the no stable periodic orbits embedded in the chaotic attractor.

This paper shows how mixed chaotic signals can be controlled using a radial basis function neural networks (RBFNs) as filter, in order to separate and to control at the same time.

The radial basis function networks provides a more effective control, it can be applied to the system at any point, even being too far from the desired state, avoiding long transient times. The control can be applied if there are only a few data of the systems, and it will remain stable much more time even with small random dynamical noise.

A Radial Basis Function Networks (RBFNs) Model.

Radial basis function emerged as a variant of artificial neural network in late 80’s. Radial basis function (RBF) neural networks provide a powerful alternative to multilayer perceptron (MLP) neural networks to approximate or to classify a pattern set.

RBFs differ from MLPs in that the overall input-output map is constructed from local contributions of Gaussian axons, require fewer training samples and train faster than MLP. The most widely used method to estimate centers and widths consist on using an unsupervised technique called the k-nearest neighbor rule (see figure 1). The centers of the clusters give the centers of the RBFs and the distance between the clusters provides the width of the Gaussians. Computation of the centers, used in the kernels function of the RBF neural network, is being the main focus to study in order to achieve more efficient algorithms in the learning process of the pattern set. The choice of adequate centers implies a high performance, concerning the learning times, convergence and generalization. The activation function for RBFs network is given by $\phi_i = \phi\left(\frac{\|X(n) - C_i\|}{d_i}\right) \; for \; i = 1,2...,m$ where $C_i = (c_{i1},...c_{ip})$ are the center of the function radial, d_i is standard deviation. The Gaussian function $\phi(r) = e^{\left(\frac{-x^2}{2}\right)}$ is the most useful in these cases [Moody, J. and Darken C].

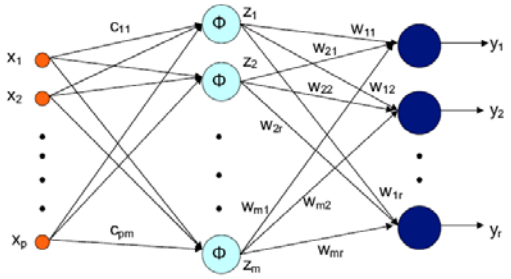


Figure 1.- Radial Basis Function Neural Network.

Simultaneous control of chaotic systems.

A discrete dynamical function is going to be controlled, all the trajectories are focused towards the stable point $x_{n+1} = f(x_n)$ where $x_n \in \Re^2$, $f : \Re^2 \rightarrow \Re^2$.

Several systems are employed: The Lozi, Ikeda and Tinkerbell.

Lozi system [Chen G. & Dong X.] is described by the following equations:

$$\begin{cases} x_{k+1} = l_1(x_k, y_k, p) = -p|x_k| + y_k + 1 \\ \quad y_{K+1} = l_2(x_k, y_k, q) = qx_k \end{cases}$$ where p and q are two real parameters

The values of the parameters that are taken to study the Lozi system are $p = 1$, $q = 0.997$. The stable points of the systems are given by :

$$Q^+ = (q_1, q_2) = \left(\frac{1}{p-(q-1)}, \frac{q}{p-(q-1)}\right)$$

$$Q^- = (q_3, q_4) = \left(\frac{-1}{p-(q-1)}, \frac{-q}{p-(q-1)}\right)$$

The attractor of Lozi system is the one that figure 2 shows.

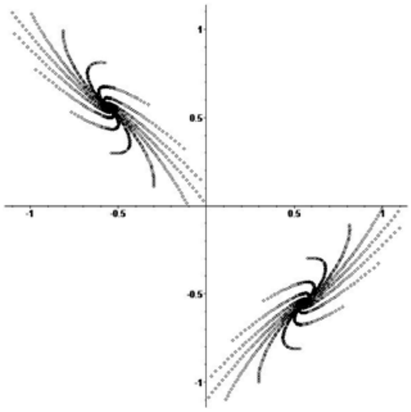


Figure 2

Ikeda system [Casdagli, M.] is described by the following equations:

$$\begin{cases} x_{k+1} = 1 + \mu(x_k \cos z - y \sin z) \\ y_{k+1} = \mu(x_k \sin z - y \cos z) \end{cases}$$, with

$$z = 0.4 - \frac{6}{1 + x^2 + y^2}, \mu = 0.7$$

The attractor of Ikeda function is the one that figure 3 shows.

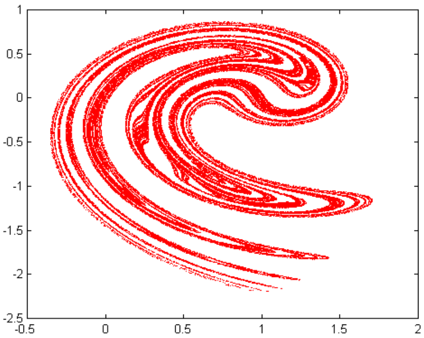


Figure 3 .

Applying the Newton methods with double precision, it can be found that the point $P = (0.60144697, 0.18998107)$ is the equilibrium point.

Thinkerbell system is described by:

$$\begin{cases} x_{k+1} = x_k{}^2 - y_k{}^2 + C_1 x_k + C_2 y_k \\ y_{k+1} = 2x_k y_k + C_3 x_k + C_4 y_k \\ C_1 = 0.9, C_2 = -0.6013 \\ C_3 = 2.0, C_4 = 0.4 \end{cases}$$

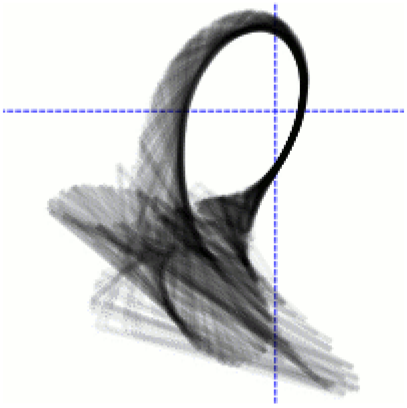


Figure 4.

The attractor of Thinkerbel function is the one that figure 4 shows.

The equilibrium point is $P = (0,0)$.

A radial basis function networks (RBFNs) has been used to control chaotic trajectories in the equilibrium points. The neural network employed as the main controller is a radial basis function networks.

The radial basis function network employed as the main controller consisting of three layers of neurons (input layer, hidden layer and output layer). The input layer has two neurons, one for each of the variables of the function ($f : \Re^2 \to \Re^2$) of the systems. In the hidden layer the configuration in the learning phase were seven neurons. And in the output layer again two neurons, one for each coordinate of equilibrium point of chaotic functions. We added noise too in the input patterns, in each case. We use a basis radial function with the competitive rule concienceful, the metric Euclidean, the function transfer tanhaxon and the learning rule momentum.

Learning Procedure

1. Input patterns. The input patterns are obtained of chaotic functions (Lozi, Ikeda and Thinkerbell) taking initial points $L_0 = (0.3, -1)$, $I_0 = (-0.9, 0.8)$ and $T_0 = (-0.3, 0.4)$. The time series of Lozi, Ikeda and Thinkerbell are calculated obtaining the collection of training patterns. The patterns set are obtained from mixed the tree time series previous. Also included a set of patterns with added noise.
2. Output patterns. The output patterns are the equilibrium points where the function must be controlled.
3. Hidden neurons. Several simulations have been performed in order to know how the number of hidden neurons affects the mean square error, in find the corresponding stable point. The best results obtained are with seven hidden neurons.

4. Number of input patterns. The variation of error along the number of input patterns has been studied, among them files with 500, 1500 and 3000 for each system. The figure 5 shows error for 1500 patterns and figure 6 shows error for 3000 patterns of the input file.
5. To finish the learning phase of the network, another input pattern is obtained starting with other initial points for each chaotic function, finding the time series and training the neural network again.

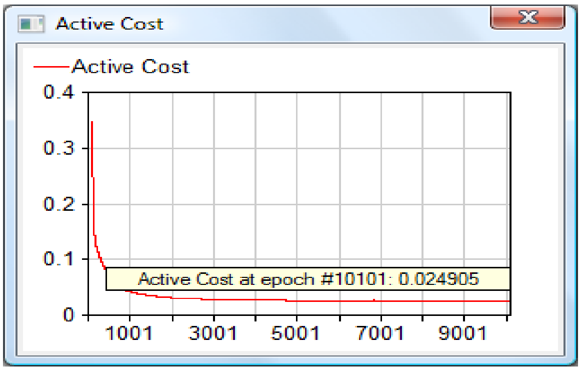


Figure 5

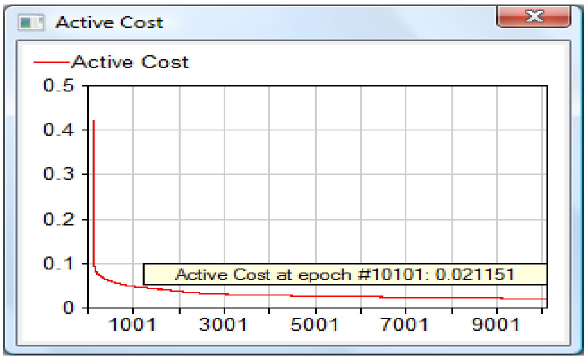
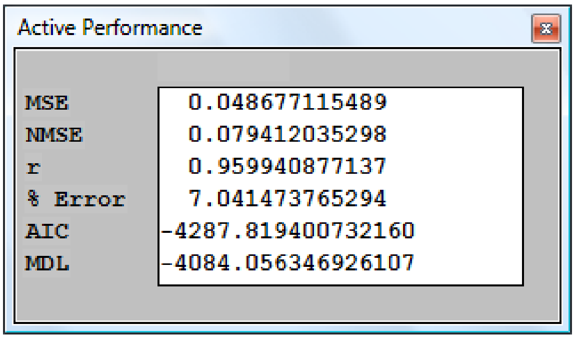


Figure 6

Achieving the control

Once the training phase is ended, it is necessary to check if the neural network is able to separate and control the function in the stable point in each case. Then choose several points for each system chaotic, are far enough from the stables points. These points are the basis for the pattern generation of the chaotic function. Each pattern set are made up of 1500 patterns for training RBFNs.

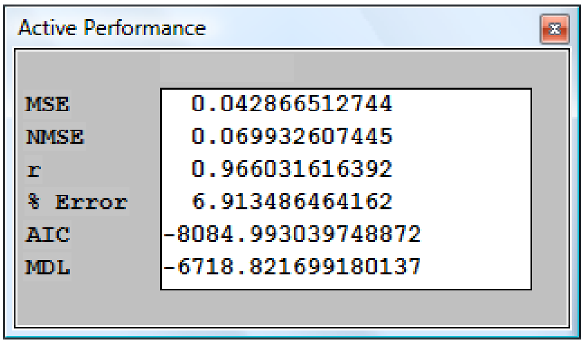
The network is also able to control the function when there exist only a few data and with some kind of noise. Next dynamical noise is added to the input, distributed on interval $[-0.01,0.01]$. The results after training are similar to the obtained error without noise. We obtained the table 1 and table 2.



Active Performance

MSE	0.048677115489
NMSE	0.079412035298
r	0.959940877137
% Error	7.041473765294
AIC	-4287.819400732160
MDL	-4084.056346926107

Table 1



Active Performance

MSE	0.042866512744
NMSE	0.069932607445
r	0.966031616392
% Error	6.913486464162
AIC	-8084.993039748872
MDL	-6718.821699180137

Table 2

Conclusion

In this paper we have demonstrated the feasibility for using radial basis function neural network to implement schemes for simultaneous control of deterministic chaos in several nonlinear dynamical systems. The simulation experiments have illustrates the proposed method is effective for chaotic systems.

An important advantage of this control technique is that the obtained controllers are very stable, presenting a good behavior even with a small random dynamical noise or with a few data.

Bibliography

[Anderson, James A. 1995] Anderson, James. A. An Introduction to Neural Networks Cambridge. MA: MIT Press (1995).

[Chen G. & Dong X, (1992)] On feedback control of chaotic dynamical systems. Int. J. of Bifurcations and chaos, 2, 407-411 (1992).

[Chen G. & Dong X.; (1993)] From Chaos to Order. Int. J. of Bifurcations and Chaos, 3, 1363-1409 (1993).

[Haykin, S. 1994] Neural Networks. A Comprehensive Foundation. Macmillan, New York (1994).

[Hübler A. W. 1989] Adaptative control of chaotic systems. Helvetica Physica 62,343-346 (1989).

[Moody, J. and Darken C. (1989)]. Fast learning in networks of locally-tuned processing units. Neural Computation, 1:281-294 (1989).

[Ott E., Grebogi C. & Yorke J. A., (1990)]. Controling Chaos. Phys. Rev. Lett, 64, 1196-1199 (1990)

Authors' Information

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

Ana Martinez*– Natural computing group. Universidad Politécnica de Madrid, Spain. a.martinez@upm.es*

Rafael Gonzalo*– Natural computing group. Universidad Politécnica de Madrid, Spain. rgonzalo@fi.upm.es*

STUDY OF THE APPLICATION OF NEURAL NETWORKS IN INTERNET TRAFFIC ENGINEERING

Nelson Piedra, Janneth Chicaiza, Jorge López, Jesús García

Abstract: *In this study, we showed various approachs implemented in Artificial Neural Networks for network resources management and Internet congestion control. Through a training process, Neural Networks can determine nonlinear relationships in a data set by associating the corresponding outputs to input patterns. Therefore, the application of these networks to Traffic Engineering can help achieve its general objective: "intelligent" agents or systems capable of adapting dataflow according to available resources. In this article, we analyze the opportunity and feasibility to apply Artificial Neural Networks to a number of tasks related to Traffic Engineering. In previous sections, we present the basics of each one of these disciplines, which are associated to Artificial Intelligence and Computer Networks respectively.*

Keywords*: Traffic Engineering, Artificial Neural Networks, Internet Traffic Engineering, Computer Networks.*

ACM Classification Keywords*: D.4.4 Communications Management*

Conference*: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

The rapid expansion of the Internet regarding services, applications, coverage and users, has changed its traditional approach. A few years ago, the Internet was only a restricted means for data flow. Today, due to the liberalization, flexibility and easy access to Internet, the demand for the requirements of the applications has increased: better quality of service, higher bandwidth, less delay, better transmission quality, amongst others. This involves the research and develpment of more inventive solutions in order to provide a better quality of service for users.

One of the key factors that providers and users must face is the congestion in the service, which what causes undesirable consequences for both parts: loss of money and dissatisfaction for both. A quick solution to this situation is the increase in the capacity of the resources offered by those services. However, this is not an acceptable alternative because the use of the service is not constant and/or static and the budget of resources is limited. Besides, the distribution of data traffic is a stochastic process; therefore, during some periods there are low levels of activity or there is not any activity at all; thus, this capacity is subused.

To improve the performance of networks, we apply the principles, concepts and technologies of Traffic Engineering (TE); consequently, congestion is reduced, and traffic and resources are properly managed. The *Internet Engineering Task Force* (IETF) RFC 3272 describes the supports of *Internet Traffic Engineering* (ITE).

Due to their capacity and characteristics, *Artificial Neural Networks* (ANN) are being applied in various fields in which traditional methods and techniques have not efficiently solved underlying problems.

ANNs appeared with the purpose of emulating some characteristics of human beings, specifically, the capacity for memorizing, relating ideas and perform actions.

Through a training process, ANNs can determine nonlinear relationships in a data set by associating the corresponding outputs to input patterns. Therefore, the application of these networks to Traffic Engineering can help achieve its global objective: "intelligent" agents or systems capable of adapting dataflow according to available resources.

This document consists of three chapters. The first and second chapters deal with the fundamentals of Traffic Engineering and Artificial Neural Networks respectively. Later, in the third chapter, some experimental applications as well as the comparison of results of ANNs with other techniques for the implementation of specific traffic engineering functions are analyzed.

1. Traffic Engineering

The general objective of Traffic Engineering is to improve the performance of an operational network [Awduche et al., 2002]; consequently, reducing its congestion and increase the efficiency in using its resources [Delfino et al., 2006].

Trafic Engineering attempts to solve one of the main problems of IP networks: to adjust IP traffic flows to make a better use of bandwidth as well as send specific flows on specific paths too [Alcocer and García].

IETF has proposed several techniques to provide Quality of Service (QoS) on the Internet. Currently, IP networks have three significant characteristics: (1) they provide real-time services, (2) they have become mission critical, and (3) their operating environments are very dynamic [Awduche et al., 2002]. From this perspective, it is complex to model, analyze and solve problems related to maintenance, management and optimization of computers networks.

1.1. Concepts of Traffic Engineering

According to García (2002), Traffic Engineering can be defined as the process of controlling data flow through a network, that is, the process of optimizing the use of available resources from various flows and optimizing the global use of resources and benefits of the network [Xio et al., 1999] y [Xio et al., 2000] and [García et al., 2002]. Consequently, TE encompasses the application of technology and scientific principles to the measurement, characterization, modeling, and control of Internet traffic.

Traffic Engineering deals with planning, control and network optimization with the purpose of achieve its goal: to adapt traffic flow to the physical network resources so that there are no congested resources whereas other resources are subused.

In IETF RFC 3272, the principles of *Internet Traffic Engineering* (ITE) are described, including aspects such as context, model and taxonomy. Moreover, there is a historical review, contemporary TE techniques and recommendations as well as other fundamental aspects.

According to [Awduche et al., 1999] and [Awduche et al., 2002], ITE deals with the management of the capacity of network traffic distribution, considering aspects such as evaluation and performance optimization of operational IP networks.

1.2. Causes of network congestion

From what Delfino (2006) [Delfino et al., 2006], network congestion can be caused by:

- Insufficient network resources (for example, link bandwidth or buffer space).
- Inefficient use of resources due to static traffic assignment to certain routes.

The first problem can be solved by increasing the capacity of resources. For the second problem, Traffic Engineering adapts traffic flows to physical network resources; thus, trying to optimally balance the use of these resources, so that there are no subused resources or over-utilized resources that cause bottlenecks. Solving congestion problems at reasonable costs is one of the main objectives of ITE.

When utilizing resources economically and reliably, we must consider requirements and performance metrics: delay, jitter, packet loss and throughput [Awduche et al., 2002]. The application of TE concepts to operational networks helps to identify and structure goals and priorities in terms of enhancing the quality of service. The application of traffic engineering concepts also aids in the measurement and analysis of the achievement of these goals. As a general rule, traffic engineering concepts and mechanisms must be sufficiently specific and well defined to address organizational requirements, but simultaneously flexible and extensible to accommodate unforeseen future demands.

1.3. Traffic Engineering Tasks

In [Villén-Altamirano], we can find the four major traffic engineering tasks and their recommendations:

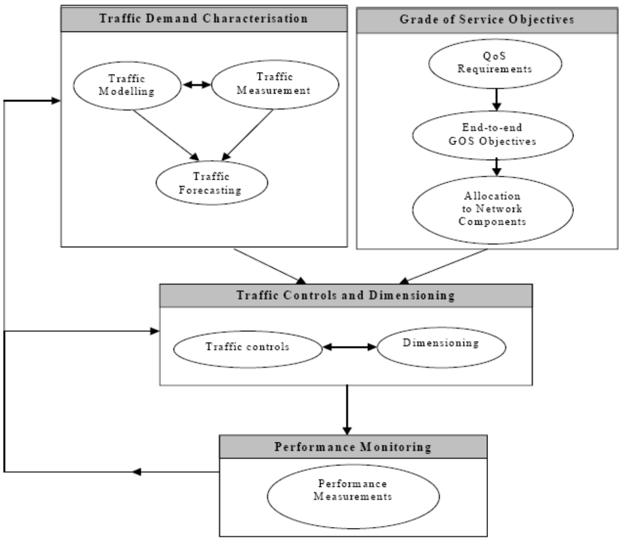

Traffic Demand Characterisation
Traffic Modelling
Traffic Measurement
Traffic Forecasting
Grade of Service Objectives
QoS Requirements
End-to-end GOS Objectives
Allocation to Network Components
Traffic Controls and Dimensioning
Traffic controls
Dimensioning
Performance Monitoring
Performance Measurements

Figure 1. Traffic Engineering Tasks [Villén-Altamirano]

For modeling the complex behavior of the network, **traffic models**, we use the *Traffic Characterization* task. Using these models traffic demand is characterized by a limited set of parameters (mean, variance, index of dispersion of counts, etc). Only those parameters that are relevant to determine the impact of traffic demand on network performance. **Traffic forecasting** is also required for planning and dimensioning purposes. This is necessary to forecast traffic demands for the time period foreseen. In order to validate these models, **traffic measures** are used.

GoS objectives are derived from Quality of Service (QoS) requirements. Grade of Service is defined as “a number of TE parameters to provide a measure of adequacy of plant under specified conditions; these GOS parameters may be expressed as probability of blocking, probability of delay, etc”.

TE must provide a design and operation of the network that guarantees the support of the traffic demand as well as the achievement of GoS objectives. Thus, **network dimensioning** (of the physical and logical network) assures that the network has enough resources to attend the traffic demand. Among the **traffic controls** we can distinguish: traffic routing, network traffic management controls, service protection methods, packet-level traffic controls, and signaling and intelligent network controls.

Although the *network performance monitoring* it can be correctly dimensioned. GoS monitoring is needed to detect errors or incorrect approximations in the dimensioning and to produce feedback for traffic characterization and network design.

1.4. Historical Review and Recent Developments

The first routing algorithms tried to minimize the use of network resources by choosing the shortest path, but this selection criterion can cause congestion in some network links whereas other links could be infra-utilized [García et al., 2002]. When applying TE concepts, some flows could go through other links with less traffic even if they are on a longer route (Figure 2).

Currently, MPLS (Multi Protocol Label Switching), is highly regarded as the proper technology to provide capacity for Traffic Engineering and QoS, -especially for backbone applications- [Sawant and Qaddour]. Among other aspects, MPLS offers: resources reservations, fault tolerance and resource optimization. The combination of

MPLS and DiffServ-TE (Differentiated Services for Traffic Engineering) has advantages to provide QoS while the utilization of network resources is optimized [Minei, 2004]. Among the characteristics of MPLS to provide TE, we have [Roca et al.]:

- Establishing explicit routes (physical path at LSP -Label Switched Path- level).
- Generating statistics regarding the use of LSPs. This information could be used for network planning and optimizing.
- Flexibility in network administration. Constraint-Based Routing can be applied so that routes for certain QoS or special services can be selected.

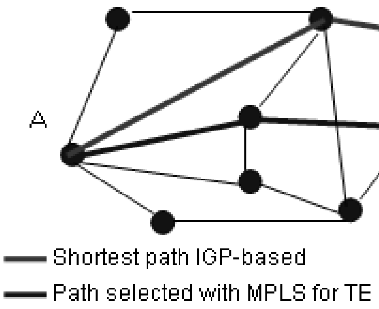


Figure 2. Routing of Packages by means of IGP and MPLS [Roca et al.]

Besides MPLS and DiffServ, other approaches have been proposed or implemented to offer TE. Some routing approaches, used a few years ago, are described in [Awduche et al., 2002].

- It is known that Internet evolved from ARPANET and adopted dynamic routing algorithms with distributed control to determine the routes that packets should take to reach their destination. This type of algorithms are adaptations of shorther path algorithms where costs are besed on link metrics. One of the weaknesses of using link metrics is that unbalanced loads in the network can occur. “In ARPANET, packets were forwarded to their destination along a path for which the total estimated transit time was the smallest”. This approach is known as Adaptive Routing, where routing decisions were based on the current state of the network in terms of delay and connectivity. One inconvenient of this approach is that it can cause congestion in different segments of the network; thus, resulting in network oscillation and instability.
- Type-of-Service (ToS) routing involves different routes going to the same destination with selection dependent upon the ToS field of an IP packet. A separate shortest path tree is computed for each ToS. Classical ToS-based routing is has been updated outdated and the ToS field has been replaced by a Diffserv field. The Diffserv model essentially deals with traffic management on a per hop basis.
- “SPF is modified slightly in ECMP (Equal Cost Multi-Path) so that if two or more equal cost shortest paths exist between two nodes, the traffic between the nodes is distributed among the multiple equal-cost paths”. Thus, it is possible that one of the paths will be more congested than the other.
- Nimrod is “a routing system developed to provide heterogeneous service specific routing in the Internet, while taking multiple constraints into account (RFC, 1992)”. Essentially, Nimrod is a link state routing protocol with mechanisms that allow restriction of the distribution of routing information. “Even though Nimrod did not enjoy deployment in the public Internet, a number of key concepts incorporated into the Nimrod architecture, such as explicit routing which allows selection of paths at originating nodes”.
- The overlay model using IP over ATM requires the management of two separate networks with different technologies (IP and ATM) resulting in increased operational complexity and cost. “The overlay model based on ATM or frame relay enables a network administrator or an automaton to employ traffic engineering concepts to perform path optimization by re-configuring or rearranging the virtual circuits so that a virtual

circuit on a congested or sub-optimal physical link can be re-routed to a less congested or more optimal one".

- In Constrained-Based Routing (CBR), the network administrator can select certain paths for special services with different quality levels (explicit delay guarantees, bandwidth, fluctuation, packet loss, etc). CBR can compute routes subject to the satisfaction of a set of constraints (bandwidth, administrative policies, etc), that is, this procedure considers parameters beyond the network topology in order to compute the most convenient route.

  "Path oriented technologies such as MPLS have made constraint-based routing feasible and attractive in public IP networks". CBR, MPLS and TE in IP networks are defined in RFC 2702.

- As said, MPLS is used to provide TE. Today, there is a wide variety of protocols used for the distribution of labels. MPLS architecture does not specify one of these protocols, but recommends their choice depending on the specific network requirements. The protocoles used can be grouped into two classes: explicit routing protocols and implicit routing protocols. Explicit routing is suitable to offer traffic engineering and allows the creation of tunnels. On the other hand, implicit routing allows establishing LSPs but does not offer traffic engineering characteristics [Sienra, 2003].

Among the most common routing protocols we have; the Constraint-based Routing Label Distribution Protocol (CR-LDP) and the Resource Reservation Protocol-Traffic Engineering (RSVP-TE). CR-LDP is an extension of the LDP, which is an implicit routing protocol, sets up a determined path in advance, that is, LSPs will be established with MPLS Quality of Service. CR-LDP is a solid-state protocol, in other words, after establisshing the connection, this connection remains "open" until it is closed. The operation of RSVP-TE is similar to that of CR-LDP, since it sets up a point-to-point LSP that guarantees an end-to-end service. The difference is that RSVP-TE requires periodic refreshment of the route to remain active (soft state). With these last protocols and the application of various traffic engineering strategies, it is possible to assign different quality of service levels in MPLS networks.

1.5. Recommendations for Internet Traffic Engineering

In [Villén-Altamirano] some recommendations for Traffic Engineering are proposed. They are classified according to their major tasks.

RFC3272 [Awduche et al., 2002] describes high-level functional and non-fuctional recommendations for ITE. Functional recommendations are necessary to achieve TE objectives and non-functional recommendations are related to quality attributes or state characteristics of a TE system.

Likewise, in [Feamster et al., 2003], there are some guidelines to provide traffic engineering between domains, more specifically; some approaches of BGP (Border Gateway Protocol) are discussed.This protocol by itself does not facilitate common TE tasks.

2. Artificial Neuronal Networks

The idea of Artificial Neuronal Networks (NNA) was conceived originally as a try for modeling the bio—physiology in the human brain; this is, to understand and explain how the brain works. The aim was to create a model capable to emulate the human process for reasoning. Most part of the starting works in neuronal networks was done by physiologists but not by engineers [TRECSoluciones, 1995].

Since Santiago Ramón y Cajal discovered the neuronal structure in the nervous system, many contributions have tried to "reproduce" or at less imitate in a "litte scale" the way the human brain works.; in this context, in 1943, Walter Pitts and Warren McCulloch, proposed a mathematical model of neuron which explains the way that those processing units work.

In 1949, the physiologist Donald Hebb pointed out in his book "*The Organization of Behavior*" the learning rule known as *Rule of Hebb*. His proposal had relation with synapses conductivity, or with neurons connections. Hebb showed that the repeated activation in a neuron for other through a established synapses, increases its conductivity and made it more alike to be active successively, inducing to the formation of a neuronal circuit strongly connected.

In the summer of 1951, Minsky and Edmons made the first neuronal networks machine which consisted basically of 300 empty tubes and an automatic pilot from a B-24. They called their creation “Sharc”; it was a network with 40 artificial neurons which imitated a rat’s brain.

In 1957, Frank Rosenblatt presented the Perceptron, a neuronal network with supervised learning which learning rule was a modification to the Hebb’s proposal.

Almost one decade later, in 1969, Marvin Minsky and Seymour Paper wrote a book called “*Perceptrons*”, in which they probed the limitations of perceptrons in solving problems relatively easy; when they published the book, all the research about perceptrons were suspended and annulled.

In the 60´s other two supervised models were proposed, based in the Perceptron of Rosenblatt called Adaline and Madaline. In those cases, the adaptation of the weights was done taking into account the error, calculated as the difference between the wished output and the one given by the network, similar to the perceptron, nevertheless, the learning rule used is different.

The modern age for ANN surges with the backpropagation learning technique. In 1977, James Anderson developed a lineal model, called Lineal Associator, which consisted of some lineal integrators elements (neurons) which added their inputs. In 1982, John Hopfield showed a work on neuronal networks in the National Sciences Academy; which describes clearly and with mathematical rigor a network which was give his name, and is a variation from the Lineal Associator. Also, in this year, Fujitsu Enterprise started the development of thinking computers for application in robotic.

The 80´s decade was overpowering for spreading the ANN, as some non supervised and hybrid models and more developed kind of networks were proposed. Nowadays, many works show their successful application in different non lineal problems, which have not been modeled using traditional methods such as Statistics, Operations Research and others.

2.1. Structure and Functioning

A Biological Neuronal Network (brain) is constituted by a series of interconnected elements, called neurons, which operate in parallel. It has been estimated that in our brain there are around 100 thousand million neurons and more than 100 billion of connections (synapses).

Neurons, as the other cells in the body, work through electric impulses and chemical reactions. The electric impulses that a neuron uses to exchange information with other neurons in a network go through the axon which makes contact with the dendrites in the next neuron through the synapses. The intensity in the signal — synaptic weight- transmitted depends in the efficiency of the synaptic transmission. The signal transmitted to the neuron can be inhibitor or stimulator. The neuron shoots, or sends the impulse through its axon, if the stimulation exceeds its inhibition by a critic value — neuron threshold- [TRECSoluciones, 1995].

2.2. Elements of Artificial Neuron

Following, is presented the basic structure of the artificial neuron.

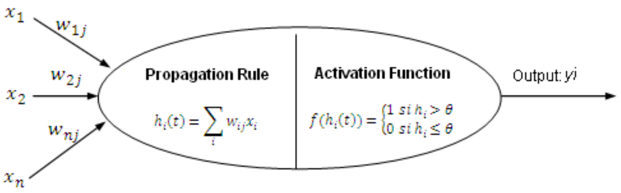

x_1
w_{1j}
w_{2j}
x_2
w_{nj}
x_n
Propagation Rule
$h_i(t) = \sum_i w_{ij} x_i$
Activation Function
$f(h_i(t)) = \begin{cases} 1 \ si \ h_i > \theta \\ 0 \ si \ h_i \leq \theta \end{cases}$
Output: yi

Figure 3. Elements of Artificial Neuron

- Xj, **neurons inputs**.
- Wij (**weights**) are coefficients which can be adapted inside the network. They determine the intensity in the input signal registered.
- **Propagation function**. Allowing obtaining, from the inputs and the weights the value of the post- synaptic potential of the neuron (*hi*). The most common function is the pondered addition of all the inputs (Figure 3). However, the propagation function can be more complex than just products addition.
- **Activation or Transference Function**. The result of the propagation function is transformed in the real output of the neuron through an algorithmic process known as activation function.

There are some activation functions to determine the neuron's output; for example, when the output value in the neuron is compared with a threshold value; if the addition is higher than the threshold value, the neuron will generate a signal; if the addition is lower than the threshold value, none signal will be generated; this function is called *heaviside*. It also can be used the lineal, sigmoid, hyperbolic tangent and others function. Particularly the sigmoid one works quite well and is normally the most common.

- Yj, **neuron output**.

A more complete artificial neuron model includes other elements such as: an output function, which is applied after the activation function is calculated; in most cases the identity function is used, therefore, it is not part of the basic figure presented.

2.3. Training of artificial neural networks

Every learning process, has two phases; the training one and the testing one; in both cases, we supply the ANN with a series of prototypes or cases, this knowledge is which allows the network to learn from the experience; in the case of **supervised models**, the network get its errors comparing the calculated value and the desired value. When there is a difference between those two values, the learning rules are applied to modify the weights in the ANN, until minimize the global error or any other cost function. On the other hand, in **unsupervised models** (or self-organized), the desired output is not known; in this case the network must organize itself to find common characteristics in the training data.

An additional element that must be established in the training phase is the learning rate. The learning rate in the ANN depends of different controllable factors which must be taken into account. Obviously, a low value in the learning rate means more time for training in order to produce a well trained ANN. With higher learning values, the network could not be able to discriminate in the same way that a system that learns slower does. Generally, additional factors -apart of time- must be considered when discussing the training off-line:

- Network Complexity: size, paradigm, architecture.
- Type of learning algorithm
- Error allowed in the final network

If changing any of those factors the training time can increase to an elevated value or obtain an unacceptable error.

2.4. ANN Architecture and Topology

The ANN topology is determined for the neurons organization and their disposition in the network. One layer is a inter-connected neurons set, most of connections happen between neurons in adjacent layers.

Therefore, the collection of parameters that define an ANN architecture are: number of layers, generally one input layer and one output layer and 0 or more intermediate (hidden); the number of neurons by layer, one or more; and the connectivity grade between the neurons, which is the number of connections between the neurons in different layers or between neurons in the same layer. In the Figure 4, it is described the architecture of a more used network called Feedward [Pizarro].

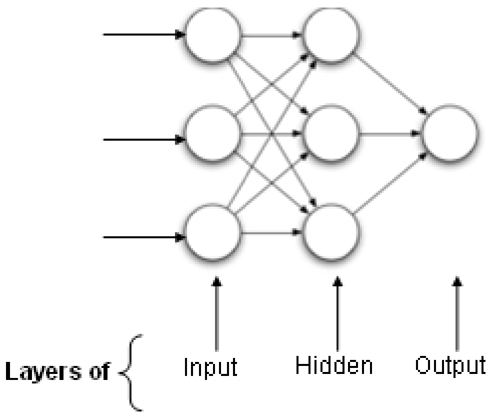


Figure 4. Multilayer Perceptron Estructure

2.5. Evaluation of the Neuronal Network to be used

The model of ANN to be used, can be selected according to:

- The number of layers, the ANN can be Monolayer —one input layer and one output layer- or Multilayer, generalization of the last one, which are added intermediate layers (hidden) between the input and the output.
- The connection type, the ANN can be: Feedforward, if the signal propagation is produced in just one way, therefore, they do not have a memory. And Recurrent if they keep feedback links between neurons in different layers, neurons in the same layer or in the same neuron.
- The connection grade, They can be: Totally connected, in the case where all the neurons in a layer are connected with the neurons in the next layer (feedforward networks) or with the neurons in the last layer (recurrent networks); and Partially connected networks, in the case when there is not total connection between neurons from different layers [Soria].
- The learning paradigm, networks can be supervised or unsupervised (or hybrid), which basic functions were described before.

Between the main neuronal models which combine the networks types mentioned before, there are:

- Perceptron, is a supervised network, monolayer, feedforward and is the base for the most of the a architecture of the ANN which interconnect between their selves.
- Backpropagation, as the perceptron, the backpropagation network uses supervised learning; however, this one is multilayer. The importance of this network is its generalization capacity or produce satisfactory outputs for inputs that the system has never seen before during its training phase.
- Self-organized maps, they constitute a practice of unsupervised learning and competitive; it considers that the influence that a neuron exercises on the others is a function of the distance between them. They can be applied to cover two basic functionalities; as classificatory or to represent multidimensional data in less dimension spaces (normally one or two dimensions), preserving the topology from the input.

Once presented the fundaments and models of the most important neuronal networks, following will be presented some successful applications in the Traffic Engineering field.

3. Applications of Neural Networks in Internet Traffic Engineering

Below we mention some characteristics of Artificial Neural Networks that can be crucial when applying them in areas such as Internet Traffic Engineering:

- ANNs, through a training process, are capable of determine nonlinear relationships in a data set by associating the corresponding output or outputs to input patterns. Consequently, many ANN models are used for determining forecasts based on a data source. This characteristic can be used for making predictions. For example, to determine the available bandwidth, detect traffic congestion patterns, forecast the use of resources (for instance, links) and even to establish or improve routing algorithms and, in general, to apply it to the tasks related to TE.

- The types of learning available for some models are batch learning (off-line) and on-line learning. They can be used for forecasting and classification depending on the data available and the available processing capacity. On-line learning is usually used in those problems in which there are a lot of training patterns. With these capacities trace files generated by some devices and network applications could be processed (in real-time or off-line); thus, facilitating TE tasks such as traffic modeling, control optimization and network dimensioning.
- Supervised Models such as the Multilayer Perceptron through the backpropagation algorithm or Adaline; or unsupervised models such as Kohonen Maps (due to their capacity for memorizing patterns) can be applied to extract or eliminate noise in signals.
- A neural network considers changes in the environment and can adapt itself to these changes, that is, once the network has been trained and tested, it will be capable of establishing the learned relationships on a new data set.
- An ANN-based approach can learn specific models from each network system and provide acceptable solutions of the underlying real systems.

Now, we will mention some characteristics of the tasks to be performed by Traffic Engineering (associated to the processes in Figure 1). Later, some projects of ANN applications in this area will be described.

- Measurement and network performance forecasting. The use of shared network resources and bandwidth are dynamic [Eswaradass et al., 2005]. Therefore, a bandwidth forecast is a very complicated task for being approached with traditional methods such as Statistics.
- Network systems modeling is a complicated tasks that can be solved trough neural networks (network traffic is nonlinear and very difficult to model and predict). In addition, traffic statistics of various applications show that each type of traffic presents a different traffic patter. By using a neural network, we can characterize the heterogeneous nature of changes in network traffic [Eswaradass et al., 2005].
- Network planning. Since a neural network is capable of establishing patterns that model traffic nature, it will also be capable of establishing mechanisms for network planning by providing guides to adapt traffic flow to physical network resources (so that there are no congested resources where as others are underutilized, this is a Traffic Engineering objective).

3.1. Bandwidth Forecasting

There are some methodologies and tools for estimating bandwidth capacity and availability respectively (some of them are mentioned in the Eswaradass', Sun and Wu job). However, they do not provide complete metrics; for instance, they do not predict bandwidth. Due to the heterogeneous and dynamic nature of network traffic, there are a few available works to predict network performance in terms of available bandwidth and lantency [Eswaradass et al., 2005].

In [Eswaradass et al., 2005], an available bandwidth forecasting method is proposed, this approach is based on Artificial Neural Networks. The prediction must consider various network applications (TCP, UDP, ICPM and others). This system has been tested on traditional trace files and compared to a system known as NWS (Network Weather Service, a model that is widely used for prediction). The experimental results showed that the neural networks approach always provides a better prediction (more precision based on the minimum global error) on NWS systems.

Predictions have been made by making an ANN for each type of network traffic, integrating partial results to obtain global predictions. Besides, noise and performance predictions are categorized after noise reduction.

In Table I, some details about the model are shown, according to [Eswaradass et al., 2005]. Although, it is not specified in the document, we can conclude that the network used (due to the description of the solution) is the Multilayer Perceptron, to which the real bandwidth value has been provided and the adjustment of its is based on the network error calculations.

Table I
Description of the Neural Network Model

Configuration Parameters		
Parameter	Description	Value
Learning rate	Determines the network learning rate	0.01
Number of epochs	Indicates the number of times a data set is trained	700
Network Architecture		
Layer type	Description	
Input layer	Depends upon the number of selected parameters: timestamp, average packet rate and average bit rate (in this case 3).	
Hidden layer	3 hidden layers and 3 perceptrons in each layer. The nonlinear sigmoid function is used as an activating function.	
Output layer	Available bandwidth/minute	
Training Patterns. As input data for the training process, trace files generated in the University of Auckland have been used. These historic files have been previously pre-processed and contain the record of time and network traffic — of different types: TCP, ICMP and UDP-. Each trace log contains incoming packet (timestamp, packet length, source and destination IP addresses). According to [Eswaradass et al., 2005], the number of packets in each second and the number of bits in each second are sufficient to produce estimates of the consumed bandwidth over time.		
Cost function. The metric used for evaluation is the relative prediction error, err.*err = P-redictedValue-ActualValue ActualValue.* PredictedValue is the bandwidth predicted for the next n seconds and ActualValue is the bandwidth measured for the next n seconds. Mean error, which is calculated by averaging all of the relative errors, is used as the cost function to be minimized.		
Simulation Software. For the simulation of the network model WEKA has been used, which is a free software package that offers a collection of various algorithms for solving data mining, including ANN.		

1) Implementation in ANNs: Predictions have been made by making an ANN for each type of network traffic, integrating partial results to obtain global predictions. Besides, noise and performance predictions are categorized after noise reduction.

In Table I, some details about the model are shown, according to [Eswaradass et al., 2005]. Although, it is not specified in the document, we can conclude that the network used (due to the description of the solution) is the Multilayer Perceptron, to which the real bandwidth value has been provided and the adjustment of its is based on the network error calculations.

2) Discussion on the Problems and Strengths of the ANN-based approach: Below we discuss some problems, strengths and future works of the neural networks-based approach.

- With more parameters and input data, the accuracy of the results is better. However, the increase of parameters and input data will increase prediction time and network training [Eswaradass et al., 2005]. Therefore, trace files must be analyzed in order to identify small data sets and input parameters.
- The selection of parameters can be done with the technique known as analysis of main components, which is implemented through a unsupervised network model, that is, all the components of an input pattern —or many parameters of trace files- could be provided for the unsupervised ANN; finally, we will get only more important parameters for forecasting.
- One problem to be solved is the selection of a proper training set. According to [Eswaradass et al., 2005], "the prediction performance with an ANN is not satisfactory for short-term trace files, which contain data for a

couple of hours or less than 1 day", or files containing traffic data more than 3 weeks. On the contrary, "network traffic data in 7-10 days is enough for neural network training".

- The construction cost of an ANN in general "is greater than those prediction systems that use linear prediction models".
- The ANN-based prediction mechanism is viable and practical. It can be used as an only prediction component or can be incorporated into the NWS for a better network prediction.
- This approach uses batch learning, that is, considers historic trace files as training patters. The next step is to provide run-time prediction. For this, the on-line processing algorithm should be used.

Table II summarizes the experimental results of the prediction performed with the data recorded at the University of Auckland uplink. AUCKLAND II is a collection of 1-day trace between December 1999 and June 2000. AUCKLAND IV is a single trace that contains data of the traffic reported between February and April 2001 (6 1/2 week trace). As seen, ANN-based prediction is the most accurate in all cases. The most reduced error percentage (5% for daily traces) occurs when a separated prediction for each type of traffic is done. Consequently, if the trace file would contain traffic data of a single application, the prediction could be even more accurate.

Table II

Global error reduction percentage for NWS and ANN

	Original Prediction*	Before noise reduction**	After noise reduction***
AUCKLAND IV	1.39%	2.33%	3.14%
AUCKLAND II	2.49%	3.68%	5%

* Prediction performed considering the various traffic flows as a whole

** Prediction performed after separating the various types of traffic

***Results after removing ICMP and UDP, only keeping TCP, which is the dominant constituent of the network traffic (95%).

3.2. Classification of Internet Traffic

The classification of Internet traffic can be used for differentiating services or for applying network security schemes. The traditional classification is usually done using the packet header field of 'port number", the layer 4 header (TCP/UDP). However, the use of this number could be unreliable in the classification of Internet traffic given the nature and characteristics of this network: free. Therefore, it is not mandatory that these applications use specific port numbers [Li et al., 2000] in [Trussell et al., 2005].

In [Trussell et al., 2005], a classification and estimation method of traffic intensity in an application is proposed. This method is based on the size distribution of packets registered in a switch (or router) during a short period time, identifying flows with significant quantity of time-sensitive data, such as voice over IP or real-time video. A switch (or router) can give preference to these flows, thus, being a mechanism to increase the Quality of Service (QoS).

As said, packet size distribution, as part of the characteristics of an application, is used as an indicator of application type. The distribution data can be obtained from the IP packet (layer 3) in order to avoid accessing the TCP header, which takes additional time and computation [Trussell et al., 2005].

1) Comparing MMSE, POCS and ANNs: In [Trussell et al.,2005], three methods for estimation of the traffic are compared: CLLSQ (Constrained Least SQuares), POCS (Projections Onto Convex Sets) and Neural Networks. According to this document, methods that use ANNs performed best in the tests. The detection of several significant classes can be done reliably. Below we describe some details of the project:

Table III

Details of the Project [Trussell et al., 2005]

Training Data. The data for the ANN training are collected from the North Carolina University backbone network using a tool for analysis of network traffic, named TCPDUMP. The data was collected continuously for four hours. The recorded parameters are: source port number, destination port number, packet size. “The applications were identified using the source and destination port numbers depending on the port assignments by IANA” (Internet Assigned Numbers Authority).
Histogram Generation. “In order to reduce the dimensionality of the data, the Ethernet packet sizes range from 60-1514 bytes were considered (some of them divided into a manageable number of bins)”.
Clustering. “To verify the conjecture that applications could be reliably characterized by their histograms”, the histogram collection using several clustering methods was analyzed, “which all resulted in natural groupings of the histograms of applications”.
Estimation an Detection. “The total distribution of packet sizes at a particular network node is the mixture of the distribution of the individual applications. Therefore, we can model the total network traffic as the linear combination of major applications”.

“The architecture used for the neural networks was a simple single hidden layer with a single output neuron”, the activation function for hidden layer is the log-sigmoid. For estimation, the output neuron used a linear function; while for the detection case, the output neuron used a log-sigmoid function”. According to [Trussell et al., 2005], in the case of estimation, it was determined that using six neurons in the hidden layer is appropriate to model the problem “In the case of detection, it was found that two hidden neurons were sufficient to give good results”. In this document, no method is indicated to establish the number of hidden neurons that should be used in every layer. Consequently, a simulation using software tools and the “trial and error” are required to determine this data.

The result of estimation performance is given in Table IV. The RMS (Root Mean Square) error obtained by the ANN is inferior than the other methods. “This result is obtained by training on one set of 24 samples and testing on the other set . If the estimation was limited to the percentage of a single application, all methods improve” and, as in the previous case, the ANN performs best.

Table IV

RMS error [Trussell et al., 2005]

Application	Average	Error RMS		
		CLLSQ	POCS	ANN
RTP	0.0119	0.0010	0.0029	0.0004
Napster	0.0111	0.0016	0.0013	0.0001
eDonkey	0.0097	0.0052	0.0010	0.0002

2) Estimation of the presence of a single application: To estimate the probability of a specific application being present in the traffic flow a neural network was used. “Since the original data contains most applications in each data set, to test detection, we created artificial data sets, based on actual data files”. According to [Trussell et al., 2005], the method obtains a very high accuracy of detecting the presence of specific applications, even at low percentages. In some applications, there is a lower detection rate due to the fact that these applications have statistical properties that are similar to other applications (in this case eDonkey).

An strengths of the ANN approach is that “will allow the reduction of the size of the histograms and a corresponding decrease in computation time† Very small weight on a particular bin of input vectors for all neurons indicates that this bin is not needed for estimation or detection”.

3.3. Overload Control in Computer Networks

Neural Networks can also be used for controlling overload in Computer Networks. In [Wu and Michael], a supervised network model capable of learning control actions based on historical records. The result, according to this, is a control system that is simple, robust and near-optimal.

Guaranteeing good performances of overload control systems is essential. Therefore, control actions are required to protect network resources from excessive loads. These actions must be based on mechanisms that regulate new arriving requests.

According to [Wu and Michael], there are two kinds of control strategies, namely, local or centralized; according to the amount of information the control decisions are based.

As known, “traffic is stochastic and the mapping from traffic to optimal decisions is complex. To solve this problem, ANN can be used, “bearing in mind its ability of learning unknown functions from a large number of examples and its implementation in real time once being trained”. The first step is to “generate examples for the training the network. The second step is to train a group of neurons based on these data. After training, the neurons cooperate to infer the control decisions based on locally available information”.

1) Requirements for Implementing Overload Control: A network device is “overloaded if its work load averaged over a period exceeds a predefined threshold. Overload control can be implemented by gating new calls. The gate values, i.e., the fraction of admitted calls, are updated periodically. An effective control is to find out the optimal gate values for each period”. In [Wu], five requirements are described and than an ideal control algorithm should satisfy.

2) Solution using Neural Networks: The network inputs are parameters about requests to a network device and output corresponding control decisions accordding to maximum value allowed. The input-output mapping is reached through learning process using examples generated by CCM (Centralized Control Method). “It is difficult to train the neural networks properly using examples generated for a large range of traffic intensity, but on the other hand, training them at a fixed traffic intensity makes them inflexible to changes”; thus, losing generalization performance. Hence, for each network device, a group of neural networks was built; “each member being a single layer perceptron trained using examples generated at particular background traffic intensity”.

In [Wu and Michael], the training of a member of the group of neural networks is explained. This training is similar to that of a back-propagation network in output signals and the calculation of the mean square error. “Each hidden unit is trained at a particular traffic intensity”.

Wu’s approach compares the CCM, LCM and ANN methods. To obtain results, they performed simulations on part of the Hong Kong metropolitan network. Call attempts (call arrival rates between different nodes) “were generated according to the Poisson process, and accepted with probability given by the corresponding gate values”.

Finally, the results prove that ANN “has a throughput higher than CCM, moreover, decreases the time for making decisions (about 10% of the CPU time of CCM); thus, NNM can be implemented in real time”.

3.4. Fault Diagnosis

In Computer Networks, the proper management of error messages can facilitate fault diagnosis in a system. For example, when occur a network breakdown, a lot of error messages are generated, making it difficult to differentiate the primary sources and secondary consequences of a problem. Thus, it is desirable to have an efficient and reliable error message classifier [Wu and Michael].

Several learning machine-based algorithms can be used for classification tasks, such as, decision rules, nearest neighbor-based, tress, and more; nevertheless, they do not support a high level of “noisy and ambiguous features inherent in many diagnosis tasks”. Thus, ANN can approximate highly nonlinear functions with a high precision.

In [Wu and Michael] we can see that the hybrid classifier is composed of an input layer, a hidden layer that contains R nodes representing classification rule vectors and a perceptron output layer. The approach is based on a competitive network model called winner-take-all.

In this case, el training set was a collection of error messages generated from a telephone exchange computer. The training set consists of 442 samples and the test set of 112 samples. As mentioned in previous cases the ANN-based approach yields better results than the other options analyzed.

Conclusion

The Traffic Engineering in order to reach its aim of improving the performance of an operational networkwork, minimizing the congestion of resources and the effective use of them, must take into account the different requirements and metrics of performance, mechanisms and politics that improve the integrity and reliability of the network [Awduche et al., 2002] covering aspects like: characterization of the traffic demand, planning, control and optimization of the network.

Nowadays, publications, studies, applications and efforts related to NNA are considerable, despite its complexity. There are different simulation tools that can facilitate its comprehension and results verification. According to [Werbos, 1998] y [Brio and Sanz, 2001], can be considered that the application of neuronal networks have reached their maturity.

The application of the ANN in traffic engineering is quite promising. As Del Brio [Brio and Sanz, 2001] points out, the characteristics which make that a specific situation to be an ideal candidate for NNA application are the following, which are massively present in all the traffic engineering problems.

- There is not a method which describes the problem completely; therefore, modelling it becomes a complex task.
- To have an important amount of data, which will serve as examples or patterns for the learning of the network; the data related to the problem is imprecise or include noise; the problem is high dimensionality.
- In changing working conditions, The NNA can adapt their selves perfectly due to its adapting capacity (re-training).

There are different proposes that have shown a potential application of the Artificial Neuronal Networks in the Communication Networks field; in this work, applications in specific tasks of the Traffic Engineering have been shown, such as: prediction, control, monitoring and resources performance. Have been seen some approximations for prediction of bandwidth prediction [Eswaradass et al., 2005] overcharge control [Wu and Michael], traffic classification [Trussell et al., 2005] and diagnosis of error messages [Wu and Michael].

Due to the own characteristics of network traffic, the application of the methods and conventional statistics techniques is not appropriate to provide optimal predictions; On the other hand, the experimental results provided by NNA models demonstrate that those tools offer best predictions - minimal error — in contrast with other systems.

For data prediction tasks, in general, different models can be used: deterministic, statistical, probabilistic, and based on machine learning; each model has its own strengths and weaknesses. Real problems can be disarranged in different modules, each one implemented with different techniques; it implies, depending of the problem characteristic and requirements, the best technique can be selected or use hybrid models to obtain better results.

The described works for NNA application in Traffic Engineering have in common a pre-processing phase, in which, the incoming data are treated, depurated and selected, before being processed by the neurons of the NNA; this phase can be the most extensive and determine extensively the success in the realization of other parts of the project, helping to control risks, to reach a maximum performance and avoid mistaken conclusions.

Bibliography

[Alcocer and García] Alcocer, F. and García, J. Curso de Teleeducación sobre MPLS.

[Awduche et al., 1999] Awduche, D., Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and McManus, J. Requirements for traffic engineering over MPLS. Technical report, IETF. 1999.

[Awduche et al., 2002] Awduche, D., Chiu, A., Elwalid, A., and Widjaja, I. Overview and principles of Internet Traffic Engineering. Technical Report, IETF. 2002.

[Brio and Sanz, 2001] Brio, M. D. and Sanz, A. 2001. Redes Neuronales y Sistemas Difusos. 2da. edition.

[Delfino et al., 2006] Delfino, A., Rivero, S., and SanMartín, M. Ingeniería de tráfico en Redes MPLS. Technical report, Congreso Regional de Telecomunicaciones. 2006.

[Eswaradass et al., 2005] Eswaradass, A., Sun, X., and Wu, M. A neural network based predictive mechanism for available bandwidth. The ACM Digital Library. 2005

[Feamster et al., 2003] Feamster, N., Borkenhagen, J., and Rexford, J. 2003. Guidelines for Internet Traffic Engineering. ACM SIGCOMM Computer Communication Review.

[García et al., 2002] García, J., Raya, J., and Raya, V. Alta velocidad y calidad de servicio en Redes IP. 2002.

[Li et al., 2000] Li, F., Seddigh, N., Nandy, B., and Malute, D. 2000. An empirical study of today's Internet Traffic for Differentiated Services IP QoS.

[Minei, 2004] Minei, I. MPLS Diffserv-aware Traffic Engineering. Technical report, Juniper Networks Inc. 2004.

[Pizarro] Pizarro, F. El paradigma de las Redes Neuronales Artificiales. Technical report, Departamento de Informática Tributaria de España.

[Roca et al.] Roca, T., Chica, P., and Muñoz, M. Ingeniería de Redes. trabajo sobre mpls.

[Sawant and Qaddour] Sawant, A. and Qaddour, J. Mpls diffserv: a combined approach. Illinois State University.

[Sienra, 2003] Sienra, L. Ofreciendo Calidad de Servicio mediante MPLS. Centro de Investigación e Innovación en Telecomunicaciones (CINIT). 2003.

[Soria]. [TRECSoluciones, 1995] Soria, E. Redes neuronales artificiales. TRECSoluciones 1995. Redes neuronales artificiales.

[Trussell et al. (2005)] Trussell, H., Nilsson, A., Patel, P., and Wang, Y. Characterization, Estimation and Detection of Network Application Traffic. North Carolina State University Raleigh. 2005.

[Villén-Altamirano] Villén-Altamirano, M. Overview of itu recommendations on traffic engineering. Department of Computer Science of University of Cyprus.

[Werbos, 1998] Werbos, P. Neural Networks combating Fragmentation. IEEE Spectrum Magazine. 1998.

[Wu and Michael], Wu, S. and Michael, K. Neural networks: Techniques and applications in Telecommunications Systems. The Honk Kong University of Science and Technology.

[Xio et al., 2000] Xio, X., Hannan, A., Bailey, B., and Ni, L. Traffic Engineering with MPLS in the Internet. IEEE Network Magazine. 2000.

[Xio et al., 1999] Xio, X., Irpan, T., Hannan, A., Tsay, R., and Ni, L. Traffic Enginnering with MPLS, America's Network Magazine. 1999.

Authors' Information

Nelson Piedra *- Universidad Técnica Particular de Loja, Escuela de Ciencias de la Computación - UTPL, Ecuador, nopiedra@utpl.edu.ec*

Janneth Chicaiza *- UTPL - Unidad de Proyectos y Sistemas Informáticos, Ecuador, jachicaiza@utpl.edu.ec*

Jorge López *- UTPL - Unidad de Proyectos y Sistemas Informáticos, Ecuador, jalopez2@utpl.edu.ec*

Jesús García Tomás *- Universidad Politécnica de Madrid - UPM, Facultad de Informática, España, jgarcia@fi.upm.es*

INTELLIGENT COMPUTATIONS FOR FLOOD MONITORING

Nataliia Kussul, Andrii Shelestov, Serhiy Skakun

***Abstract**: Floods represent the most devastating natural hazards in the world, affecting more people and causing more property damage than any other natural phenomena. One of the important problems associated with flood monitoring is flood extent extraction from satellite imagery, since it is impractical to acquire the flood area through field observations. This paper presents a method to flood extent extraction from synthetic-aperture radar (SAR) images that is based on intelligent computations. In particular, we apply artificial neural networks, self-organizing Kohonen's maps (SOMs), for SAR image segmentation and classification. We tested our approach to process data from three different satellite sensors: ERS-2/SAR (during flooding on Tisza river, Ukraine and Hungary, 2001), ENVISAT/ASAR WSM (Wide Swath Mode) and RADARSAT-1 (during flooding on Huaihe river, China, 2007). Obtained results showed the efficiency of our approach.*

***Keywords**: flood extent extraction, neural networks, data fusion, SAR images.*

***ACM Classification Keywords**: I.4.6 Segmentation - Pixel classification*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Increasing numbers of natural disasters have demonstrated to the mankind the paramount importance of the natural hazards topic for the protection of the environment and the citizens. Climate change is likely to increase the intensity of rainstorms, river floods, and other extreme weather events. The dramatic floods of Central and Eastern Europe in summer 2002 and spring 2001 and 2006 emphasize the extreme in climatic variations. Floods are among the most devastating natural hazards in the world, affecting more people and causing more property damage than any other natural phenomena (Wood 2001). That is why; the problems of flood monitoring and flood risk assessment are among priority tasks in national satellite monitoring systems and international system of systems GEOSS (GEO Work Plan, 2007-2009).

Efficient monitoring and prediction of floods and risk management is impossible without the use of Earth Observation (EO) data from space. Satellite observations enable acquisition of data for large and hard-to-reach territories, as well as continuous measurements. One of the important problems associated with flood monitoring is flood extent extraction, since it is impractical to acquire the flood area through field observations. Flood extent can be used for hydraulic models to reconstruct what happened during the flood and determine what caused the water to go where it did, for damage assessment and risk management, and can benefit to rescuers during flooding (Corbley 1999).

The use of optical imagery for flood monitoring is limited by severe weather conditions, in particular presence of clouds. In turn, SAR (synthetic aperture radar) measurements from space are independent of daytime and weather conditions and can provide valuable information to monitoring of flood events. This is mainly due to the fact that smooth water surface provides no return to antenna in microwave spectrum and appears black in SAR imagery (Elachi 1988; Rees 2001).

As a rule, flood extent extraction procedure consists of the following steps: image calibration, geocoding, orthorectification using digital elevation model (DEM) and shadowing effects removal, filtration, thematic processing, testing and results verification. This paper proposes to use artificial neural networks (NN), in particular self-organizing Kohonen's maps (SOMs) (Haykin 1999; Kohonen 1995), for SAR image segmentation and classification. SOMs provide effective software tool for the visualization of high-dimensional data, automatically discover of statistically salient features of pattern vectors in data set, and can find clusters in training data pattern space which can be used to classify new patterns (Kohonen 1995). We applied our approach to the processing of data acquired from three different satellites: ERS-2/SAR (during flooding on Tisza river, Ukraine and Hungary, 2001), ENVISAT/ASAR WSM (Wide Swath Mode) and RADARSAT-1 (during flooding on Huaihe river, China, 2007).

Existing approaches to flood extent extraction

To this end different methods were proposed to flood extent extraction from satellite imagery. In European Space Agency (ESA) multi-temporal technique is used to flood extent extraction from SAR images (ESA Earth Watch, http://earth.esa.int/ew/floods). This technique uses SAR images of the same area taken on different dates (one image is acquired during flooding and the second one in “normal” conditions). The resulting multi-temporal image clearly reveals change in the Earth’s surface by the presence of colour in the image. This method has been implemented in ESA’s Grid Processing on Demand (G-POD, http://eogrid.esrin.esa.int). In (Cunjian et al. 2001), threshold segmentation algorithm is applied to flood extent extraction from RADARSAT-1 imagery. The value of threshold is chosen manually. In (Csornai et al.2004), SAR (from ESA’s ERS-2) and optical data (Landsat TM, IRS WIFS/LISS, NOAA AVHRR) are used for flood monitoring in Hungary in 2001. To derive flood extent from SAR imagery change detection technique is applied.

Though these methods are rather simple and quick (in computational terms), they posses some disadvantages: need of manual threshold selection and image segmentation, require expertise in visual interpretation of SAR images, require the use of complex models for speckle reduction, spatial connections between pixels are not concerned. To overcome these difficulties we propose neural network approach to flood extent extraction. Our approach is based on SAR image segmentation using self-organizing Kohonen maps and further image classification using additional information on water bodies derived from Landsat-7/ETM+ images and Corine Land Cover (for European countries).

Data sets description

We applied our approach to the processing of remote-sensing data acquired from three different satellites: ERS-2 (flooding on Tisza river on March 2001 (Fig. 1),

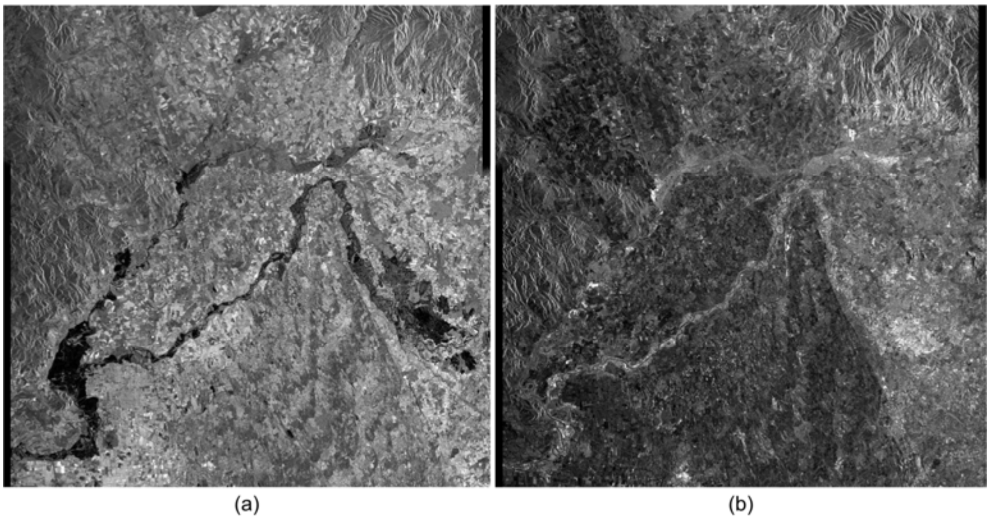


Fig. 1 Flooding (a, date of acquisition is 10.03.2001) and post-flooding (b, 14.04.2001) SAR/ERS-2 images of Tisza river (© ESA 2001)

ENVISAT and RADARSAT-1 (flooding on Huaihe river on July 2007 (Fig. 2). Data from European satellites were provided from ESA Category-1 project “Wide Area Grid Testbed for Flood Monitoring using Spaceborne SAR and Optical Data” (№4181). Data from RADARSAT-1 satellite were provided from RSGS-CAS. Spatial resolution of

ERS-2 images was 4 m (in ENVISAT SLC format (Single Look Complex)), for ENVISAT 75 m and for RADARSAT-1 was 12.5 m.

For more precise geocoding of SAR images and validation of obtained results we used the following set of additional data: Landsat-7/ETM+, European Corine Land Cover (CLC 2000) and SRTM DEM.

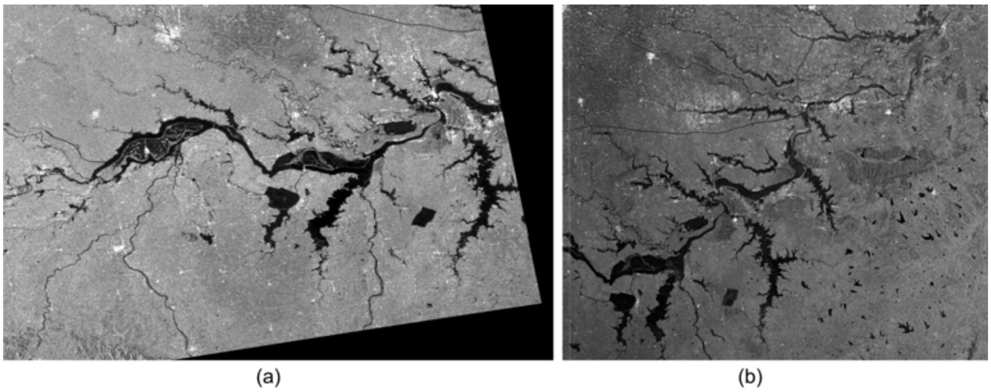


Fig. 2 SAR images acquired from ENVISAT (a, 15.07.2007) and RADARSAT-1 (b, 19.07.2007) satellites during flooding on Huaihe river, China (© ESA 2007; © CSA 2007)

In order to train and calibrate neural network, we manually chose test pixels (with the use of additional data set) that correspond to both territories with presence of water (we denote them as belonging to class “Water”) and without water (class “No water”). The number of test pixels for each of the image is presented in Table 1.

Table 1. Distribution of test pixels for ERS-2, ENVISAT and RADARSAT-1 images

Satellite image/Region	Number of test pixels for images		
	“No water”	“Water”	Total
ERS-2/Ukraine	30016	12939	42955
ENVISAT/China	60575	34493	95068
RADARSAT-1/China	135263	130244	265507

Among test pixels we did not use those ones that relate to boundaries between water and no water lands. Classification of SAR images on more than two classes (e.g. “Water”, “No water”, different levels of water presence) is beyond the scope of this paper and will be investigated in future papers.

Neural network method for flood extent extraction from SAR imagery

Our method for flood extent extraction consists of data pre-processing, image segmentation and classification on two classes using SOMs. These steps are as follows:

1. *Transformation of raw data to lat/long projection*. Level-1 data from ERS-2 and ENVISAT satellites in Envisat format and from RADARSAT-1 satellite in CEOS format were provided with ground control points (GCPs) that were used to transform images to lat/long projection in GeoTIFF format. For this purpose, we used gdalwarp utility from GDAL (Geospatial Data Abstraction Library, http://www.gdal.org).

2. *Image calibration.* In order to calibrate ERS-2/SAR and ENVISAT/ASAR images, we used standard procedures described in (Laur et al. 2004) and (Rosich and Meadows 2004) respectively. As a result of image calibration, the output signal (pixel values) was transformed to backscatter coefficient (in dB). For RADARSAT-1 image, we used original pixel values in DN (digital number).

3. *Geocoding.* We made additional geocoding procedure for ERS-2 image in order to improve the accuracy. This was done by using Landsat/ETM+ and CLC2000 data.

4. *Image processing using self-organizing Kohonen's maps (SOMs)*. SOM is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two dimensional), discretized representation of the input space of the training samples, called a map (Haykin 1999; Kohonen 1995). The map seeks to preserve the topological properties of the input space. SOM is formed of neurons located on a regular, usually 1- or 2-dimensional grid. Neurons compete with each other in order to pass to the excited state. The output of the map is a so called neuron-winner or best-matching unit (BMU) whose weight vector has the greatest similarity with the input sample **x**.

The network is trained in the following way: weight vectors $\mathbf{w}_j$ from topological neighbourhood of BMU vector i are updated according to (Haykin 1999; Kohonen 1995)

$$i(\mathbf{x}) = \arg\min_{j=\underline{1,L}} \left\| \mathbf{x} - \mathbf{w}_j \right\|,$$

$$\mathbf{w}_j(n+1) = \mathbf{w}_j(n) + \eta(n) h_{j,i(x)}(n)(\mathbf{x} - \mathbf{w}_j(n)), \; j = \overline{1, L}, \qquad (1)$$

where η is learning rate, $h_{j,i(x)}(n)$ is neighborhood kernel around the winner unit *i*, **x** is input vector, $\|\cdot\|$ means Euclidean metric, *L* is number of neurons in the output grid, *n* is number of iteration within learning.

As neighborhood kernel $h_{j,i(x)}(n)$, we used Gaussian function. For learning rate we used the flowing expression:

$$\eta(n) = \eta_0 \cdot e^{-\frac{n}{\tau}}, \eta_0 = 0.1, n = 0,1,2,..., \; (2)$$

where τ is a constant.

Kohonen maps are widely applied to image processing, in particular image segmentation and classification (Haykin 1999). Before neural network training, we need to choose image parameters that will be input to neural network. For this purpose, one can choose original pixel values, various filters, Fourier transformation etc (Gonzalez and Woods 2002). In our approach we use sliding window with backscatter coefficient values for ERS-2 and ENVISAT images and DNs for RADARSAT-1 image as inputs to neural network. The output of neural network, neuron-winner, relates to the central pixel of sliding window. In order to choose appropriate size of the sliding window for each satellite sensor, we ran experiments for the following windows: 3-by-3, 5-by-5, 7-by-7, 9-by-9 and 11-by-11.

We, first, used SOM to segment each SAR image where each pixel of the output image was assigned a number of the neuron in the map. Then, we used test pixels to assign each neuron one of two classes ("Water" or "No water") using the following rule. If neuron was activated by majority number of pixels that belong to class "Water", then this neuron was assigned "Water" class. If neuron was activated by majority number of pixels that belong to class "No water", then this neuron was assigned "No water" class. If neuron was activated by neither of test pixels, then it was assigned "No data" class.

For neural network quality assessment, we used two parameters:

– quantization error that is estimated with the following expression

$$QE = \frac{1}{N}\sum_{t=1}^{N} \left\| \mathbf{x}_t - \mathbf{w}_{i(\mathbf{x}_t)} \right\|, \quad i(\mathbf{x}_t) = \arg\min_{j=\underline{1,L}} \left\| \mathbf{x}_t - \mathbf{w}_j \right\|,$$

where *N* is the number of test pixels.

– classification rate that shows relative number of correctly classified test pixels.

Results of image processing

In order to choose the best neural network architecture, we ran experiments for each image varying the following parameters:

- size of sliding window of images that define number of neurons in input layer of neural network;
- number of neurons in output layer, i.e. sizes of 2-dimensional output grid.

Other parameters that were used during image processing are as follows:

- neighborhood topology: hexagonal;
- neighborhood kernel around the winner unit: Gaussian function;
- initial learning rate: 0.1;
- number of training epochs: 20.

Initial values for the weight vectors are selected as a regular array of vectorial values that lie on the subspace spanned by the eigenvectors corresponding to the two largest principal components of input data (Kohonen 1995). Using this procedure, computation of the SOM can be made orders of magnitude faster, since (i) the SOM is then already approximately organized in the beginning, (ii) one can start with a narrower neighborhood function and smaller learning rate. The results of experiments for images are resented in Table 2.

For image with higher spatial resolution (ERS-2 and RADARSAT-1) the best results were achieved for larger input sliding window 7-by-7. In turn, for ENVISAT/ASAR WSM image we used sliding window of smaller size 3-by-3. The use of higher dimension of input window for ENVISAT image led to the coarser resolution of resulting flood extent image and reduced classification rate.

Table 2. Results of SAR images classification using SOMs

Satellite image	Input dimension	Output grid of neurons	Classification rate for test pixels		
			«No water»	«Water»	Total
ERS-2	7-by-7	5-by-5	99.81	99.86	99.90
ENVISAT	3-by-3	7-by-5	100.0	95.70	98.44
RADARSAT-1	7-by-7	5-by-5	99.99	91.92	96.03

The resulting flood extent images for ERS-2, ENVISAT and RADARSAT-1 satellite are shown on Fig. 3-5.

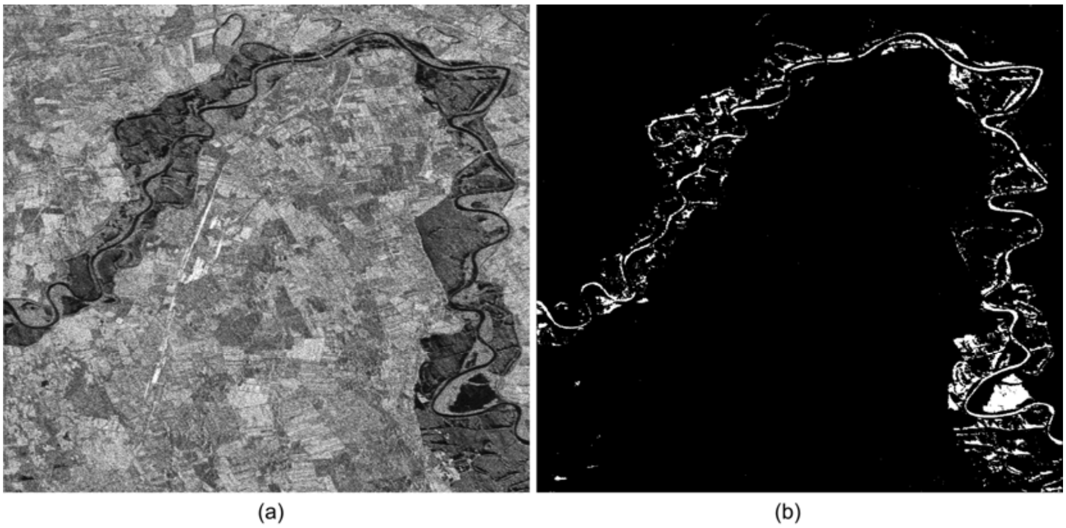


Fig. 3 Raw ERS-2 image (a) and resulting flood extent shown with white color (b) for Tisza river, Ukraine and Hungary (© ESA 2001)

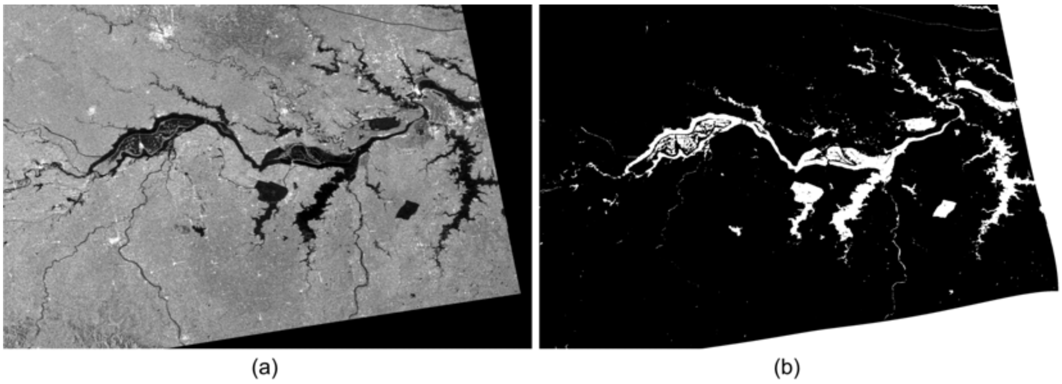


Fig. 4 Raw ENVISAT image (a) and resulting flood extent shown with white color (b) for Huaihe river, China (© ESA 2007)

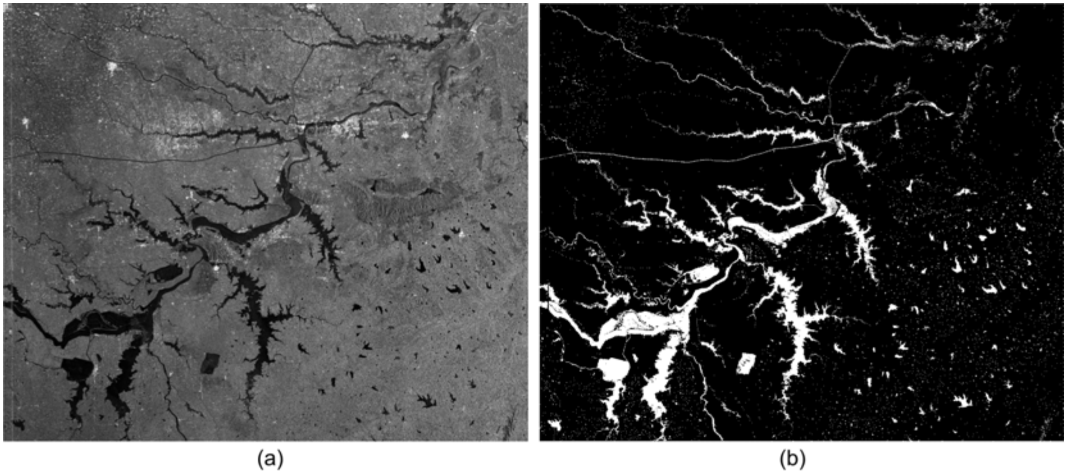


Fig. 5 Raw RADARSAT-1 image (a) and resulting flood extent shown with white color (b) for Huaihe river, China (© CSA 2007)

Conclusions

In this paper we proposed neural network approach to flood extent extraction from SAR imagery. To segment and classify SAR image, we apply self-organizing Kohonen's maps (SOMs) that possess such useful properties as ability to automatically discover statistically salient features of pattern vectors in data set, and to find clusters in training data pattern space which can be used to classify new patterns. As inputs to neuron network, we use a sliding window of image pixels intensities. We ran experiments to choose the best neuron network architecture for each satellite sensor: for ERS-2 and RADARSAT-1 the size of input was 7-by-7 and for ENVISAT/ASAR the sliding window was 3-by-3. The advantages of our approach are as follows: (i) we apply sliding window to process the image and thus considering spatial connection between pixels; (ii) neural network's weight vectors are adjusted automatically by using training data. This enables implementation of our approach in automatic services for flood monitoring. Considering the selection of test pixels to calibrate the neuron network, i.e. to assign each neuron one of the classes, this process can be also automated using geo-referenced information on water bodies for the given region.

We applied our approach to derive flood extent from SAR images acquired by three different sensors: ERS-2/SAR for Tisza river (Ukraine); ENVISAT/ASAR and RADARSAT-1 for Huaihe river (China). Classification rates for manually selected test pixels were 99.99%, 91.92% and 96.03%, respectively. These results demonstrate the efficiency of our approach.

Acknowledgements

This work is supported by ESA CAT-1 project “Wide Area Grid Testbed for Flood Monitoring using Spaceborne SAR and Optical Data” (No. 4181) and by joint INTAS-CNES-NSAU project “Data Fusion Grid Infrastructure” (Ref. Nr 06-1000024-9154).

Bibliography

[Corbley, 1999] Corbley K.P. Radar Imagery Proves Valuable in Managing and Analyzing Floods Red River flood demonstrates operational capabilities. Earth Observation Magazine, 1999, vol. 8, num. 10.

[Csornai et al., 2004] Csornai G., Suba Zs., Nádor G., László I., Csekő Á., Wirnhardt Cs., Tikász L., Martinovich L. Evaluation of a remote sensing based regional flood/waterlog and drought monitoring model utilising multi-source satellite data set including ENVISAT data. In: Proc. of the 2004 ENVISAT & ERS Symposium (Salzburg, Austria, 6-10 September 2004).

[Cunjian et al., 2001] Cunjian Y., Yiming W., Siyuan W., Zengxiang Z., Shifeng H. Extracting the flood extent from satellite SAR image with thesupport of topographic data. In: Proc. of Int. Conf. on Inf. Tech. and Inf. Networks (ICII 2001), vol. 1, pp 87-92.

[Elachi,1988] Elachi C. Spaceborne Radar Remote Sensing: Applications and Techniques. New York: IEEE Press, 1988.

[Haykin, 1999] Haykin S. Neural Networks: A Comprehensive Foundation. Upper Saddle River, New Jersey: Prentice Hall, 1999.

[GEO Work Plan, 2007] GEO Work Plan 2007-2009 “Toward Convergence”.
(www.earthobservations.org/documents/wp0709_v4.pdf)

[Gonzalez and Woods, 2002] Gonzalez R.C., Woods R.E. Digital Image Processing. Prentice Hall, Upper Saddle River, New Jersey, 2002.

[Kohonen, 1995] Kohonen T. Self-Organizing Maps. Series in Information Sciences, Vol. 30. Springer, Heidelberg, 1995.

[Laur et al., 2004] Laur H., Bally P., Meadows P., Sanchez J., Schaettler B., Lopinto E., Esteban D. ERS SAR Calibration. Derivation of the Backscattering Coefficient in ESA ERS SAR PRI Products. ES-TN-RS-PM-HL09 05, November 2004, Issue 2, Rev. 5f

[Rosich and Meadows, 2004] Rosich B., Meadows P. Absolute calibration of ASAR level 1 products generated with PF-ASAR. ESA-ESRIN, ENVI-CLVL-EOPG-TN-03-0010, 07 October 2004.

[Wood, 2001] Wood H.M. The Use of Earth Observing Satellites for Hazard Support: Assessments & Scenarios. Final Report. National Oceanic & Atmospheric Administration, Dep. of Commerce, USA, 2001.
(http://www.ceos.org/pages/DMSG/2001Ceos/overview.html)

[Rees, 2001] Rees W.G. Physical Principles of Remote Sensing. Cambridge University Press, 2001.

Authors' Information

Nataliia Kussul *– Prof., Dr., Head of Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, build. 4/1, Kyiv-187, 03680 Ukraine, e-mail: inform@ikd.kiev.ua*

Andrii Yu. Shelestov *– PhD, Senior Researcher, Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, build. 4/1, Kyiv-187, 03680 Ukraine, e-mail: inform@ikd.kiev.ua*

Serhiy V. Skakun *– PhD, Research Assistant, Department of Space Information Technologies and Systems, Space Research Institute of NASU-NSAU, Glushkov Ave 40, build. 4/1, Kyiv-187, 03680 Ukraine, e-mail: serhiy.skakun@ikd.kiev.ua*

OUTLIERS RESISTANT LEARNING ALGORITHM FOR RADIAL-BASIS-FUZZY-WAVELET-NEURAL NETWORK IN STOMACH ACUTE INJURY DIAGNOSIS TASKS

Yevgeniy Bodyanskiy, Oleksandr Pavlov, Olena Vynokurova

***Abstract**: In this paper an outliers resistant learning algorithm for the radial-basis-fuzzy-wavelet-neural network based on R. Welsh criterion is proposed. Suggested learning algorithm under consideration allows the signals processing in presence of significant noise level and outliers. The robust learning algorithm efficiency is investigated and confirmed by the number of experiments including medical applications.*

***Keywords**: computational intelligence, hybrid architecture, wavelet, fuzzy-wavelet neural network, robust learning algorithm, outliers resistant.*

***ACM Classification Keywords**: I.2.6 Learning – Connectionism and neural nets.*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Nowadays artificial neural networks (ANN) have gained the significant prevalence for solving the wide class of the information processing problems, uppermost for the identification, emulation, intelligence control, time series forecasting of arbitrary kind under significant noise level, and also the structural and parametric uncertainty.

The multilayer feedforward networks of three-layer perceptron type, where the elementary nodes are so-called P- neurons with monotonic activation functions are the most known and popular. The efficiency of the multilayer networks is explained by their universal approximation properties in combination with relative compact presentation of the simulated nonlinear system. It means, that they can be used successfully in the tasks of the simulation (emulation) non-linear systems, which can be described by the equation

$$y(k) = F(x(k)) + \xi(k), \tag{1}$$

where $y(k)$ is the output system signal in k-th instant of discrete time $k = 0,1,2,\ldots,$ $x(k) \in X$ - $(n \times 1)$ is the vector of input signal, including both exogenous variables and previous values of the output signal, $F(\bullet)$ is the arbitrary function, generally in some unknown form, $\xi(k)$ is the unobserved disturbance with unknown characteristics. Usually it is assumed that function $F(\bullet)$ is defined either on the unit hypercube or on the orthotop

$$x_i(k) \in [x_i^{\min}, x_i^{\max}],\ i = 1,2,\ldots,n\,,$$

where $x_i^{\min}$, $x_i^{\max}$ are the known low and upper limits of the i-th input influence variation.

The principal disadvantage of the multilayer networks is the low learning rate which is based on backpropagation algorithm which makes their application in the real time tasks impossible.

Alternative to the multilayer ANNs are the radial basis function networks, having one hidden layer consisting of, so-called, R-neurons. These networks learning is realized on the level of the output layer which is usually represented by the adaptive linear associator [1-6]. Unlike P-neurons, R-neurons conventionally have bell-shaped activation function $f_j(x)$, where the argument is a distance (usually in Euclidean metric) between the current value of input signal $x(k)$ and the center c_j of the j-th neuron, i.e.

$$\varphi_j(x(k)) = \varphi_j\left(\sum_{i=1}^{n}(x_i(k)-c_{ji})^2\right) = \varphi_j\left(\left\|x(k)-c_j\right\|^2\right). \tag{2}$$

The principal advantage of RBFN is the high learning rate in the output layer, because the turning parameters are linearly included to the network description. At the same time the problem of R-neurons centers allocation is remaining, and its unsuccessful solving leads to the «curse of dimensionality» problem. Using clustering techniques though allows reducing the size of the network, but excludes the possibility of on-line operation. Here it can be noted, that in [7] the gradient recurrent procedure of the component-wise tuning parameters c_{ji} is described, but it is characterized by the low learning rate.

Along with neural networks for the arbitrary type signals processing, in the last years the wavelet theory is used sufficiently often [8-9], providing the compact local signal presentation both in the frequency and time domains. At the turn of the artificial neural network and wavelets theories the wavelet neural networks [10-15] have evolved their efficiency for the analysis of nonstationary nonlinear signals and processes.

Elementary nodes of the wavelet neural networks are so-called radial wavelons [16], where the activation functions are the even wavelets with argument in form the Euclidian distance between $x(k)$ and wavelet translation vector c_j, where that every component of distance $\left|x_i(k)-c_{ji}\right|$ is weighted by the dilation parameter σ_{ji} such, that

$$\varphi_j(x(k)) = \varphi_j\left(\sum_{i=1}^{n}\left((x_i(k)-c_{ji})/\sigma_{ji}\right)^2\right), \tag{3}$$

where $\varphi_j(\bullet)$ is wavelet activation function. The receptive fields for such wavelons are hyperellipsoids with axes which are collinear to coordinate axes of the space X.

Taking into consideration the equivalence of radial basis ANN and fuzzy inference systems [17, 18], and also possibility of using even wavelet as a membership function [19, 20], within the bounds of the unification paradigm [16] we can talk about such hybrid system as Radial-Basis-Function-Wavelet-Neuro-Fuzzy Network (RBFWNFN) having the radial-basis function network fast learning ability, fuzzy inferences systems interpretability and wavelet's local properties.

It can be noted, that mostly tuning algorithms based on traditional squared learning criteria in the case of the processing data being contaminated by outliers with unknown distribution law, have shown themselves very sensitive to anomalous outliers. Thus the actual task is a synthesis of the robust learning algorithms, that allow signal processing in presence of anomalous outliers.

This paper is devoted to synthesis of robust learning algorithm for RBFWNN, which has adjustable level of insensitivity to the different kind of outliers, rough errors, non-Gaussian disturbances, has high convergence rate and provides the advanced approximation properties in comparison with conventional computational intelligence systems.

1. Radial-basis-fuzzy-wavelet-neural-network architecture

Let us consider the two-layers architecture shown on fig. 1 that coincides with the traditional radial-basis neural network. The input layer of the architecture is the receptor and in current time instant k the input signal in vector form $x(k) = (x_1(k), x_2(k), \ldots, x_n(k))^T$ is fed on it. Unlike radial basis function network the hidden layer consists of not by R-neurons, but by wavelons with wavelet activation function in the form

$$\varphi_j(x(k)) = \varphi_j\left((x(k)-c_j)^T Q_j^{-1}(x(k)-c_j)\right),\ j = 1, 2, \ldots, h, \tag{4}$$

in which instead of translation parameters σ_{ji} in (1) the positive-definite dilation matrix Q_j is used, i.e. it is not Euclidian distance, but Itakura-Saito metric [21].

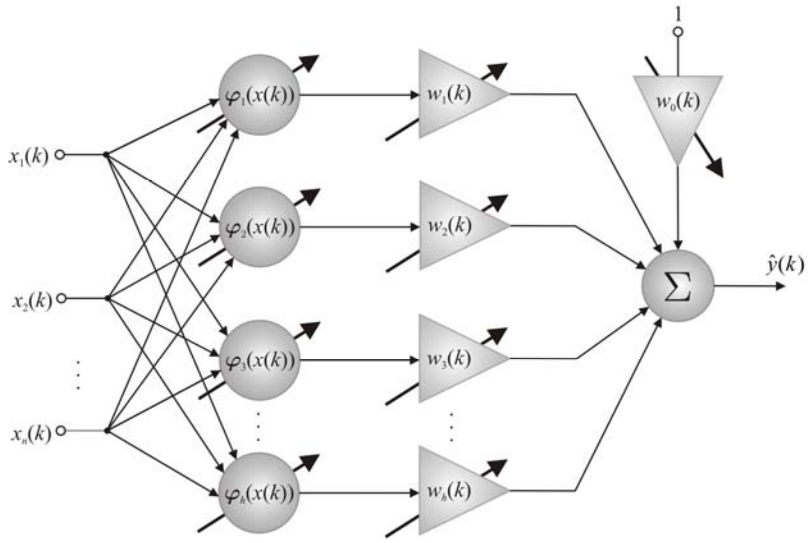
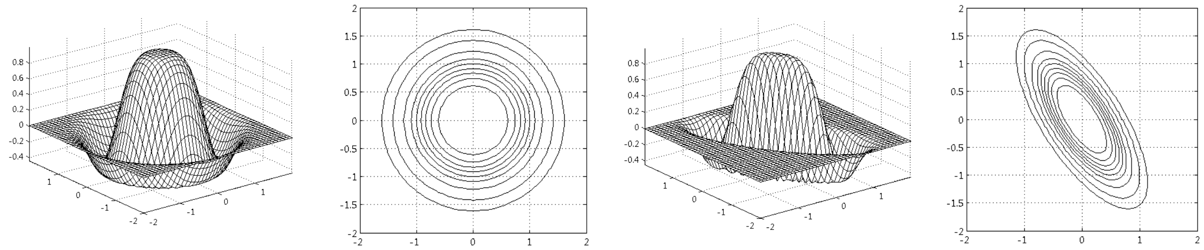


Fig. 1 – Radial-basis fuzzy wavelet neural network

This results to the fact that receptive fields – wavelons hyperellipsoids (2) can have the arbitrary orientation relatively to the coordinate axes of space X, what extends the functional properties of RBFWNN.

Fig. 2 shows the wavelons activation function (2) with arbitrary matrices Q_j.



a) $Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ b) $Q_3 = \begin{pmatrix} 4 & 3 \\ 1 & 2 \end{pmatrix}$

Fig. 2 – Wavelons activation function with arbitrary matrices Q_j

And at last, the output layer is the common adaptive linear associator with tuning synaptic weights w_j

$$\hat{y}(k) = w_0 + \sum_{j=1}^{h} w_j \varphi\left((x(k) - c_j)^T Q_j^{-1}(x(k) - c_j)\right) = w^T \varphi(x(k)), \tag{5}$$

where $\varphi_0(x(k)) \equiv 1$, $w = (w_0, w_1, w_2, \ldots, w_h)^T$, $\varphi(x(k)) = (1, \varphi_1(x(k)), \varphi_2(x(k)), \ldots, \varphi_h(x(k)))^T$.

Thus the tuning parameters of architecture to be determined in the learning process form the set of the $h+1$ synaptic weights w_j, h $(n \times 1)$-vectors c_j and h $(n \times n)$-matrices Q_j^{-1}. In total such network includes h $(1 + n + n^2) + 1$ adjustable parameters.

2. The robust learning algorithm for radial-basis-fuzzy-wavelet neural network

The experience shows that the identification methods based on the least square criterion are extremely sensitive to the deviation of real data distribution law from Gaussian distribution. In presence of various type outliers, an outrage errors, and non-Gaussian disturbances with "heavy tails" the methods based on the least squares criterion loose their efficiency.

In this case the methods of robust estimation and identification [22] which have obtained the wide spread for the learning of the artificial neural networks [23-25] appear on the first role.

Let's introduce into the consideration the learning error

$$e(k) = y(k) - \hat{y}(k) = y(k) - w^T(k)\varphi(k) \tag{6}$$

and robust identification criterion by R. Welsh [26, 27]

$$E(k) = f(k) = \beta^2 \ln\left(\cosh\left(e(k)/\beta\right)\right), \tag{7}$$

where β is a positive parameter, that is chosen from empirical reasons and defining the size of zone of tolerance to outliers. It is necessary to note, that robust criterion (7) satisfies to all metric space axioms.

Further we shall consider synthesis of the learning algorithms. For the synaptic weights and the waveleon parameters (vectors c_j and matrices Q_j^{-1}) tuning we use gradient minimization of criterion (7), thus unlike the component-wise learning considered in [7], we make some corrections in the vector-matrix form, that, firstly is easier from computing point of view, and secondly it allows to optimize learning process on the operation rate.

In general case the learning algorithm can be written in form

$$\begin{cases} w(k+1) = w(k) - \eta_w \nabla_w E(k), \\ c_j(k+1) = c_j(k) - \eta_{c_j} \nabla_{c_j} E(k), \ j = 1, 2, \ldots, h, \\ Q_j^{-1}(k+1) = Q_j^{-1}(k) - \eta_{Q_j^{-1}} \left\{ \partial E(k) \middle/ \partial Q_j^{-1} \right\}, \ j = 1, 2, \ldots, h, \end{cases} \tag{8}$$

where $\nabla_w E$ is vector-gradient of the criterion (7) on w, $\nabla_{c_j} E$ is $(n \times 1)$-vector-gradient criterion (7) on c_j; $\left\{ \partial E(k) \middle/ \partial Q_j^{-1} \right\}$ is $(n \times n)$-matrix, formed by partial derivatives E on components Q_j^{-1}; η_w, η_{c_j}, $\eta_{Q_j^{-1}}$ are the learning rates.

For arbitrary wavelet $\varphi((x(k) - c_j)^T Q_j^{-1}(x(k) - c_j))$ we can write

$$\begin{cases} \nabla_w E(k) = -\beta \tanh\left(e(k)/\beta\right) \varphi_j((x(k) - c_j(k))^T Q_j^{-1}(k)(x(k) - c_j(k))) = -\tanh\left(e(k)/\beta\right) J_w(k), \\ \nabla_{c_j} E(k) = \beta \tanh\left(e(k)/\beta\right) w_j(k) \varphi_j'((x(k) - c_j(k))^T Q_j^{-1}(k)(x(k) - c_j(k))) \cdot \\ \qquad \cdot Q_j^{-1}(k)(x(k) - c_j(k)) = \tanh\left(e(k)/\beta\right) J_{c_j}(k), \\ \left\{ \partial E(k) \middle/ \partial Q_j^{-1} \right\} = -\beta \tanh\left(e(k)/\beta\right) w_j(k) \varphi_j'((x(k) - c_j(k))^T Q_j^{-1}(k)(x(k) - c_j(k))) \cdot \\ \qquad \cdot (x(k) - c_j(k))(x(k) - c_j(k))^T = -\tanh\left(e(k)/\beta\right) J_{Q_j^{-1}}(k), \end{cases} \tag{9}$$

where $\varphi_j'(\bullet)$ is the derivative j-th wavelet on the argument $(x(k) - c_j(k))^T Q_j^{-1}(k)(x(k) - c_j(k))$.

Then the wavelons learning algorithm of the hidden layer subject to (9) is taking the form

$$\begin{cases} w(k+1) = w(k) + \eta_w \beta \tanh\left(e(k)/\beta\right) \varphi_j((x(k)-c_j(k))^T Q_j^{-1}(k)(x(k)-c_j(k))) = \\ \qquad = w(k) + \eta_w \tanh\left(e(k)/\beta\right) J_w(k), \\ c_j(k+1) = c_j(k) - \eta_{c_j} \beta \tanh\left(e(k)/\beta\right) w_j(k) \varphi_j'((x(k)-c_j(k))^T Q_j^{-1}(k)(x(k)-c_j(k))) \cdot \\ \qquad \cdot Q_j^{-1}(k)(x(k)-c_j(k)) = c_j(k) - \eta_{c_j} \tanh\left(e(k)/\beta\right) J_{c_j}(k), \\ Q_j^{-1}(k+1) = Q_j^{-1}(k) + \eta_{Q_j^{-1}} \beta \tanh\left(e(k)/\beta\right) w_j(k) \varphi_j'((x(k)-c_j(k))^T Q_j^{-1}(k)(x(k)-c_j(k))) \cdot \\ \qquad \cdot (x(k)-c_j(k))(x(k)-c_j(k))^T = Q_j^{-1}(k) + \eta_{Q_j^{-1}} \tanh\left(e(k)/\beta\right) J_{Q_j^{-1}}(k), \end{cases} \tag{10}$$

at that convergence rate to the optimal value w, c_j and Q_j^{-1} is completely defined by learning rate parameters η_w, η_{c_j} and $\eta_{Q_j^{-1}}$.

The learning rate increasing can be achieved by using procedures more complex than gradient ones, such as Hartley or Marquardt procedures, that for the first relation (10) can be written in general form [28]

$$w(k+1) = w(k) - \lambda_w (J_w(k) J_w^T(k) + \eta_w I)^{-1} J_w(k) \tanh\left(e(k)/\beta\right), \tag{11}$$

where I is the $(n \times n)$-identity matrix, λ_w is a positive dampening parameter, η_w is a momentum term parameter.

Using the inverse matrices lemma and after applying simple transformations we obtain the effective parameters learning algorithm in the form

$$\begin{cases} w(k+1) = w(k) - \lambda_w \left(\tanh\left(e(k)/\beta\right) J_w(k)\right) \Big/ \left(\eta_w + \|J_w(k)\|^2\right), \\ c_j(k+1) = c_j(k) - \lambda_c \left(\tanh\left(e(k)/\beta\right) J_{c_j}(k)\right) \Big/ \left(\eta_c + \left\|J_{c_j}(k)\right\|^2\right), \\ Q_j^{-1}(k+1) = Q_j^{-1}(k) + \lambda_{Q_j^{-1}} \left(\tanh\left(e(k)/\beta\right) J_{Q_j^{-1}}(k)\right) \Big/ \left(\eta_{Q_j^{-1}} + Tr(J_{Q_j^{-1}}^T(k) J_{Q_j^{-1}}(k))\right). \end{cases} \tag{12}$$

In order to add more smoothing properties, using approach proposed in [29], we can introduce the modified learning procedure:

$$\begin{cases} w(k+1) = w(k) + \lambda_w \dfrac{\tanh\left(e(k)/\beta\right) J_w(k)}{\eta_w(k)}, \qquad \eta_w(k+1) = \alpha_w \eta_w(k) + \|J_w(k+1)\|^2, \\ c_j(k+1) = c_j(k) - \lambda_{c_j} \dfrac{\tanh\left(e(k)/\beta\right) J_{c_j}(k)}{\eta_{c_j}(k)}, \qquad \eta_{c_j}(k+1) = \alpha_c \eta_{c_j}(k) + \left\|J_{c_j}(k+1)\right\|^2, \\ Q_j^{-1}(k+1) = Q_j^{-1}(k) + \lambda_{Q_j^{-1}} \dfrac{\tanh\left(e(k)/\beta\right) J_{Q_j^{-1}}(k)}{\eta_{Q_j^{-1}}(k)}, \eta_{Q_j^{-1}}(k+1) = \alpha_{Q_j^{-1}} \eta_{Q_j^{-1}}(k) + Tr\left(J_{Q_j^{-1}}^T(k+1) J_{Q_j^{-1}}(k+1)\right) \end{cases} \tag{13}$$

(here $0 \le \alpha_w \le 1, 0 \le \alpha_{c_j} \le 1, 0 \le \alpha_{Q_j^{-1}} \le 1$ are the parameters of weighting out-dated information), being nonlinear hybrid of the Kaczmarz-Widrow-Hoff and Goodwin-Ramadge-Caines algorithms and including both following and filtering properties.

3. Results of the experimental research

In the first experiment the developed robust learning algorithm was tested out on the basis of a signal with intensive outliers. The signal had been obtained using Narendra's nonlinear dynamical system (it is a standard benchmark, widely used to evaluate and compare the performance of neural and neuro-fuzzy systems for nonlinear system modeling and time series forecasting) whose output signal is artificially contaminated by random noise generated according to the Cauchy distribution with the inverse transform method described by equation in form

$$F_X^{-1}(x) = x_0 + \gamma tg\left[\pi\left(x - 0.5\right)\right], \tag{14}$$

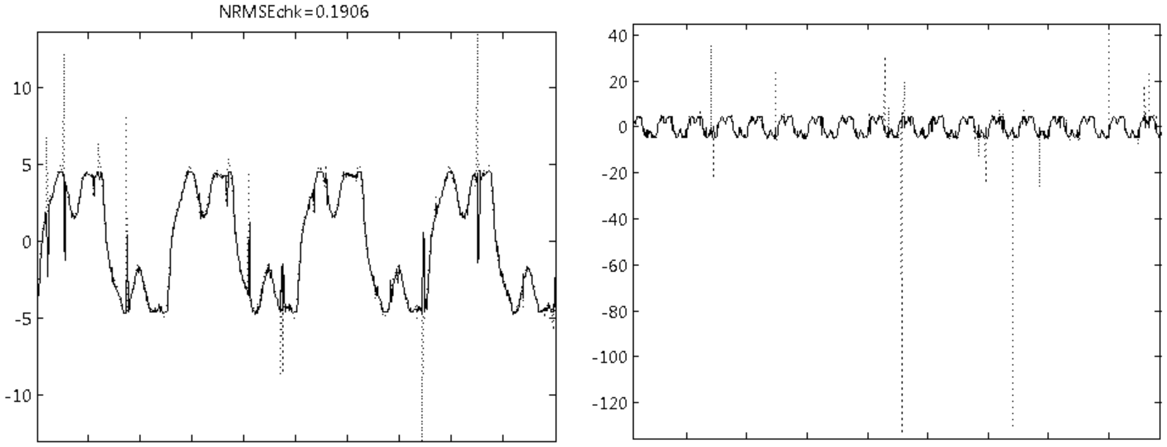
where x_0 is the location parameter, γ is the scale parameter $(\gamma > 0)$, x is the support area ($x \in (-\infty, +\infty)$).

The nonlinear dynamical system is generated by equation in form [30]

$$y(k+1) = 0.3y(k) + 0.6y(k-1) + f(u(k)), \tag{15}$$

where $f(u(k)) = 0.6\sin(u(k)) + 0.3\sin(3u(k)) + 0.1\sin(5u(k))$ and $u(k) = \sin(2k/250)$, k is discrete time. The values $x(t-4), x(t-3), x(t-2), x(t-1)$ were used to emulate $x(t+1)$. In the on-line mode of learning, RBFWNN was trained with procedure (13) for 20000 iterations. The parameters of the learning algorithm were $\beta_w = 1, \beta_c = 0.5,\ \beta_Q = 0.5,\ \alpha_w = \alpha_c = \alpha_Q = 0.99,\quad \lambda_w = \lambda_{c_j} = \lambda_{Q_j^{-1}} = 0.99$. Initial values were $\eta_w(0) = \eta_{c_j}(0) = \eta_{Q_j^{-1}}(0) = 10000$. After 20000 iterations the training was stopped, and the next 1000 points were used as the testing data set. Initial values of synaptic weights were generated in a random way from -0.1 to $+0.1$.

Fig. 3 a shows the results of the noised signal emulation (real values (dashed line) and emulated values (solid line)). Fig. 3 b shows segment of the learning process; as it can be seen the number of outliers with large amplitude, present in the beginning of the sample, didn't have a significant influence on the learning algorithm.



a) b)

Fig. 3 – Results of noised signal processing based on robust learning algorithm

The comparison of emulation results based on robust learning algorithm with results of emulation based on gradient algorithm and the algorithm based on recurrent least squares method where the structure network and the number of tuning parameters were identical was carried out.

Under RBFWNN learning using the gradient algorithm the first outlier in the beginning of the sample, had a noticeable influence on the learning algorithm. Under RBFWNN learning using the recurrent least-squares method the first occurred outlier leads to the covariance matrix so-called "parameters blow-up" what results in inability to emulate signals noised by anomalous outliers. Thus it is obvious that the proposed robust learning algorithm allows signal processing under high level outliers noise conditions.

In Table 1 the comparison results are shown.

Table 1: The results of noisy signal emulation

Neural Network / Learning algorithm	NRMSE
RBFWNN / Proposed robust learning algorithm (13)	0.1906
RBFWNN / The gradient learning algorithm	1.1242
RBFWNN / RLSM	∞

The second experiment has been made on the data set, presented by Government Institution "Institute of General and Urgent Surgery (Academy of Medical Sciences of Ukraine)". It has been carried out studying of the homeostasis indexes dynamic of the patient with the stomach acute injury [31, 32] based on outliers resistant radial-basis-fuzzy-wavelet-neural network. The indexes of oxygen cascade, system hemodynamics, daily pH-measurement, and hypoxia marker and endotoxemia were analyzed. Result of processing studied clinico-laboratory data set was the degree defining of enteral deficiency, that it has allowed to lead the adequate stomach-protect diagnosis and therapy.

Conclusion

In the paper computationally simple and effective all RBFWNN parameters robust learning algorithm is proposed. The robust learning algorithm has following and smoothing properties and allows on-line processing of non-linear signals under a number of outliers and “heavy tails” disturbances. Addition of wavelons receptor fields, including their transformations (dilation, translation, rotation) allows to improve the network approximation properties, that is confirmed by the experiments research results.

Bibliography

[1]. Moody J. Darken C. J. Fast learning in networks of locally-tuned processing units. Neural Computation, 1, 1989, P. 281-294.

[2]. Moody J., Darken C. Learning with localized receptive fields. In: Proc. of the 1988 Connectionist Models Summer School, Eds. D. Touretzky, G. Hinton, T. Sejnowski, 1988, San Mateo: Morgan-Kaufmann, P. 133-143.

[3]. Park J., Sandberg I. W. Universal approximation using radial-basis-function networks. Neural Computation, 3, 1991, P. 246-257.

[4]. Leonard J. A., Kramer M. A., Ungar L. H. Using radial basis functions to approximate a function and its error bounds. IEEE Trans. on Neural Networks, 3, 1992, P. 614-627.

[5]. Sunil E. V. T., Yung C. Sh. Radial basis function neural network for approximation and estimation of nonlinear stochastic dynamic systems. IEEE Trans. on Neural Networks, 5, 1994, P. 594-603.

[6]. Poggio T., Girosi F. A Theory of Networks for Approximation and Learning. A. I. Memo № 1140, C.B.I.P. Paper № 31., Massachussetts Institute of Technology, 1994.

[7]. Bishop C. M. Neural Networks for Pattern Recognition. Oxford: Clarendon Press, 1995.

[8]. Chui C. K. An Introduction to Wavelets. New York: Academic Press, 1992.

[9]. Daubechies I. Ten Lectures on Wavelets. Philadelphia, PA: SIAM.,1992.

[10]. Billings S. A., Wei H.-L. A new class of wavelet networks for nonlinear system identification. IEEE Trans. on Neural Networks, 16, 2005, P. 862-874.

[11]. Zhang Q. H., Benveniste A. Wavelet networks. IEEE Trans. on Neural Networks, 3, 1992, P. 889–898.

[12]. Zhang Q. H. Using wavelet network in nonparametric estimation. IEEE Trans. on Neural Networks, 8, 1997, P. 227-236.

[13]. Bodyanskiy Ye., Lamonova N., Vynokurova O. Recurrent learning algorithm for double-wavelet neuron. Proc. XII -th Int. Conf. "Knowledge - Dialogue - Solution", Varna, 2006, P.77-84.

[14]. Bodyanskiy Ye., Lamonova N., Vynokurova O. Double-wavelet neuron based on analytical activation functions. Int. J. Information Theory and Applications., 14, 2007, P. 281-288.

[15]. Bodyanskiy Ye., Pliss I., Vynokurova O. A learning algorithm for forecasting adaptive wavelet-neuro-fuzzy network Proc. XIII -th Int. Conf. "Information Research & Applications", Varna, 2007, P.211-218.

[16]. Reyneri L. M. Unification of neural and wavelet networks and fuzzy systems. IEEE Trans. on Neural Networks, 1999, 10, P. 801-814.

[17]. Jang J. S. R., Sun C.T. Functional equivalence between radial basis function networks and fuzzy inference systems. IEEE Trans. on Neural Networks, 4, 1993, P. 156-159.

[18]. Hunt K. J., Haas R., Smith R. M. Extending the functional equivalence of radial basis function networks and fuzzy inference systems. IEEE Trans. on Neural Networks, 7, 1996, P. 776-781.

[19]. Mitaim S., Kosko B. What is the best shape for a fuzzy set in function approximation? Proc. 5th IEEE Int. Conf on Fuzzy Systems “Fuzz-96”, V. 2, 1996, P. 1237-1213.

[20]. Mitaim S., Kosko B. Adaptive joint fuzzy sets for function approximation. Proc. Int. Conf. on Neural Networks “ICNN-97”, 1997, P. 537-542.

[21]. Itakura F. Maximum prediction residual principle applied to speech recognition. IEEE Trans. on Acoustics, Speech, and Signal Processing, 23, 1975, P. 67-72.

[22]. Rey W.J.J. Robust Statistical Methods. Lecture Notes in Mathematics, Berlin-Heidelberg-New York: Springer-Verlag, V. 690, 1978.

[23]. Cichocki A., Lobos T. Adaptive analogue network for real time estimation of basic waveforms of voltages and currents IEEE Proc, 139, Part C , 1992, P. 343-350.

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

[25]. Li S.-T., Chen S.-C. Function approximation using robust wavelet neural networks. Proc. of the 14th IEEE Int. Conf. on Tools with Artificial Intelligence, 2002, P. 483- 488.

[26]. Holland P.W., Welsh R.E. Robust regression using iteratively reweighted least squares. Commun. Statist., Theory and Methods., 46, 1977, P. 813 – 828.

[27]. Welsh R.E. Nonlinear statistical data analysis. Proc. Comp. Sci. and Statist., Tenth Ann Symp. Interface. Held at Nat’l Bur. Stds. Gaithersburg, MD, 1977, P. 77-86.

[28]. Bodyanskiy Ye. Identification adaptive algorithms for nonlinear control plant identification. ASU and pribory avtomatiki, Kharkiv: Vyscha shk., 81, 1987, P. 43-46. (In Russian)

[29]. Bodyanskiy Ye., Kolodyazhniy V., Stephan A. An adaptive learning algorithm for a neuro-fuzzy network. Computational Intelligence. Theory and Applications, Ed. by B. Reusch, Berlin-Heidelberg-New York: Springer, 2001, P. 68-75.

[30]. Narendra K.S., Parthasarathy K. Identification and control of dynamic systems using neural networks. IEEE Trans. on Neural Networks, 1990, 1, P. 4-26.

[31]. Cook D.J., Reeve B.K., Guyatt G.H., at all. Stress ulcer prophylaxis in critically ill patients. Resolving discordant meta-analyses. JAMA, 1996, 257, P. 308-314.

[32]. Boyko V.V., Sushkov S.V., Pavlov O.O., Vynokurova O.A. The anesthesia and intensive therapy strategy in the acute gastro ileum hemorrhage, Kharkiv, Eksklusiv, 2007. (In Russian)

Authors' Information

Bodyanskiy Yevgeniy - *Doctor of Technical Sciences, Professor of Artificial Intelligence Department and Scientific Head of the Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, Lenina av. 14, Kharkiv, Ukraine 61166, Tel +380577021890, e-mail: bodya@kture.kharkov.ua*

Pavlov Oleksandr - *Candidate of Medical Sciences, Associate Professor, Doctor- Anesthesiologist, Government Institution "Institute of General and Urgent Surgery (Academy of Medical Sciences of Ukraine)", Balakireva av.,1, Kharkiv, Ukraine, 61018, Tel. +380577021890, e-mail: pavlov73@mail.ru*

Vynokurova Olena *- Candidate of Technical Sciences (equivalent Ph.D.), Senior Researcher of the Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, Lenina av. 14, Kharkiv, Ukraine, 61166, Tel +380577021890, e-mail: vinokurova@kture.kharkov.ua*

INTELLIGENCE ALGORITHMS FOR INCREASING NAVIGATION SYSTEMS ACCURACY

Aleksandr Zbrutsky, Mohamed Rahmouni

Abstract: *Application of neural network algorithm for increasing the accuracy of navigation systems are showing. Various navigation systems, where a couple of sensors are used in the same device in different positions and the disturbances act equally on both sensors, the trained neural network can be advantageous for increasing the accuracy of system. The neural algorithm had used for determination the interconnection between the sensors errors in two channels to avoid the unobservation of navigation system. Representation of thermal error of two-component navigation sensors by time model, which coefficients depend only on parameters of the device, its orientations relative to disturbance vector allows to predict thermal errors change, measuring the current temperature and having identified preliminary parameters of the model for the set position. These properties of thermal model are used for training the neural network and compensation the errors of navigation system in non-stationary thermal fields.*

Keywords: *neural network, navigation system, time model of the sensors errors, errors interconnection function, unobservation, model adequacy, verification, neural network algorithm, increasing the accuracy*

ACM Classification Keywords*: I.5.1 Models - Neural nets*

Conference*: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Neural networks (NN) can solve many problems that could not be approached previously in any practical way. NN have been trained to perform complex functions in various fields of application, where they already have been applied, including control and navigation systems. The basic trends of using this theory are connected with the solution of complicated practical tasks. At present time there are various types of NN, which are assigned to solve diverse tasks. These models differ in connection structure methods of weight determination or teaching principles [Heerman, 1992]. Control systems anyhow using artificial NN are one of possible alternatives to classical control mode. The opportunity of using NN for solving problems of control in many respects is based on that NN consists of two layers, where the first layer is sigmoid and the second layer is linear, can approximate any function of real numbers with the set degree of accuracy [Rauch, Winarske, 1988].

The purpose of paper is to show that in many navigation systems such as inertial navigation systems and strapdown navigation systems, where a couple of sensors are used in the same device in different positions and the disturbances act equally on both sensors, the trained NN can be advantageous for increasing accuracy of such navigation systems. As a particular case a NN to designate the interconnection function of dynamically tuned gyroscope (DTG), which is used in Kalman algorithm to avoid the unobservation of the system, is trained, after that the errors of corrected gyrocompass, that allow to increase the accuracy of course determination is estimated.

Errors time model

Math model of DTG can be presented as follows:

$$\omega_x{}^{dr} = \dot{\alpha}_1 = \dot{\alpha}_0 + \Delta\dot{\alpha}_T, \quad \omega_z{}^{dr} = \dot{\beta}_1 = \dot{\beta}_0 + \Delta\dot{\beta}_T$$

$$\Delta\dot{\alpha}_T = \delta\left(\frac{\Delta c}{H}\right)\beta_0 - \delta\left(\frac{1}{\tau}\right)\alpha_0 + \frac{1}{4}(\delta S_{13} + 4\delta S_{14})(\beta_0^{\,3} + \alpha_0^{\,2}\beta_0) + \delta S_{15}(\alpha_0^{\,3} +$$

$$+\alpha_0\beta_0^{\,2}) + \frac{\Delta c}{H}\delta\beta_0 - \frac{1}{\tau}\delta\alpha_0 + \frac{1}{4}(S_{13} + 4S_{14})(3\delta\beta_0\cdot\beta_0^{\,2} + 2\alpha_0\beta_0\delta\alpha_0 + \alpha_0^{\,2}\delta\beta_0) +$$

$$+ S_{15}(3\alpha_0^{\,2}\delta\alpha_0 + 2\alpha_0\beta_0\delta\beta_0 + \beta_0^{\,2}\delta\beta_0) - \frac{H_1}{H^2}\omega_z\delta H + \frac{\omega_z}{H}\delta H_1 + \delta\left(\frac{M_\beta}{H}\right),$$

$$\Delta\dot{\beta}_T = -\delta\left(\frac{\Delta c}{H}\right)\alpha_0 - \delta\left(\frac{1}{\tau}\right)\beta_0 - \frac{1}{4}(\delta S_{13} + 4\delta S_{14})(\alpha_0^{\,3} + \beta_0^{\,2}\alpha_0) + \delta S_{15}(\beta_0^{\,3} +$$

$$\beta_0\alpha_0^{\,2}) - \frac{\Delta c}{H}\delta\alpha_0 - \frac{1}{\tau}\delta\beta_0 - \frac{1}{4}(S_{13} + 4S_{14})(3\delta\alpha_0\cdot\alpha_0^{\,2} + 2\beta_0\alpha_0\delta\beta_0 + \beta_0^{\,2}\delta\alpha_0) +$$

$$+ S_{15}(3\beta_0^{\,2}\delta\beta_0 + 2\alpha_0\beta_0\delta\beta_0 + \alpha_0^{\,2}\delta\beta_0) - \frac{H_1}{H^2}\omega_y\delta H + \frac{\omega_y}{H}\delta H_1 + \delta\left(\frac{M_\alpha}{H}\right).$$

where $\dot{\alpha}_0, \dot{\beta}_0$ - systematic errors and $\Delta\dot{\alpha}_T, \Delta\dot{\beta}_T$ - thermal errors.

As long as the analysis of thermal errors components shows their mainly linear dependence on temperature, then subject to temperature variation in an unsteady thermal field can be written:

$$\Delta\dot{\alpha}_T = \sum_{k=1}^{m} L_k\left(1 - e^{-kt/\tau_k}\right);\ \Delta\dot{\beta}_T = \sum_{k=1}^{m} R_k\left(1 - e^{-kt/\tau_k}\right); \qquad (1)$$

L_k, R_k - constant coefficients depend generally on ambient temperature, disturbance that acts on gyroscope and gyroscope parameters; τ_k is determined only by gyroscope parameters.

The examination of the character of DTG thermal errors variation (1) enables the functional dependence of errors to be presented as power series

$$\dot{\alpha}_1 = \sum_{j=0}^{n} a_i t^j\ ;\ \ \dot{\beta}_1 = \sum_{j=0}^{n} b_i t^j\ , \qquad (2)$$

which are regression equations.

Findings in (1) show that the errors variation in both measurement channels in time, in unsteady thermal fields, depend in the one way on the temperature change. The received dependences of DTG thermal errors show, that the change of its errors on both axes in time at non-stationary thermal fields is described in workmanlike manner qualitatively by similar dependences, and depend identically on temperature change.

Thus the unity of influence of factors (both determinate and casual), causing the change of disturbing moments of a gyroscope (thermal or magnetic fields, gravitation, etc.) on both measuring axes takes place.

In order to solve the task of algorithmically increasing the accuracy of gyrocompass by compensating DTG thermal errors, it is necessary to make sure that coefficients a_i, b_i (2) for the given device remain constant. With this purpose statistical equality of pairs coefficient b_i, b_i' a_i, a_i', received at various tests of DTG, is verified.

For an estimation of means of distribution $M(a_i)$ and $M(a_i')$ - their best estimations of samples $\overline{a}_i$ $\overline{a}_i'$ are utilized, and for an estimation of a dispersion σ^2 - selective estimations:

$$\widehat{S}_{ai}^{\ 2} = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(a_i - \overline{a}_i)^2\ ,\ \widehat{S}_{a'i}^{\ 2} = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(a_i' - \overline{a}'_i)^2\ .$$

The best estimation for dispersion in this case is $\widehat{S}^2 = \dfrac{\widehat{S}_{ai}^{\ 2}(n_1 - 1) + \widehat{S}_{a'i}^{\ 2}(n_2 - 1)}{n_1 + n_2 - 2}$.

If a hypothesis $M(a_i)$ = $M(a_i')$ it is fair, then random variable ($a_i - a_i'$) submits to the normal law of distribution with average of distribution equal to zero and dispersion equal to $\sigma^2 = (\frac{1}{n_1} + \frac{1}{n_2})$.

As sample estimate of dispersion $D(a_i - a_i')$ usually accept an estimation of $S^2{}_{(a_i - a_i')} = (\frac{1}{n_1} + \frac{1}{n_2})\widehat{S}^2$.

If the random variable $a_i - a_i'$ submits to the normal law of distribution, so statistics

$$t_c = \frac{(a_i - a_i') - M(\overline{a}_i - \overline{a}_i')}{\widehat{S}_{(a_i - a_i')}} = \frac{(a_i - a_i') - M(\overline{a}_i - \overline{a}_i')}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\frac{(n_1 - 1)S_{ai}{}^2 + (n_2 - 1)S_{a'i}{}^2}{n_1 + n_2 - 2}}}$$

has t_c- Student distribution, and $k = n_1 + n_2 - 2$.

Having chosen probability $p = 1 - \alpha$, according to the table t_c of distribution it is possible to determine critical value $t_{cn1+n2-2;\alpha}$, for which $p(|t_c| > t_{cn1+n2-2;\alpha}) = \alpha$.

If the calculated value $|t_c| > t_{cn1+n2-2;\alpha}$ with the probability of $p = 1 - \alpha$ then divergence of a_i, a_i' will be considered to be significant (not casual).

Experimental measurement of DTG drift change at the change of external temperature is shown on fig. 1.

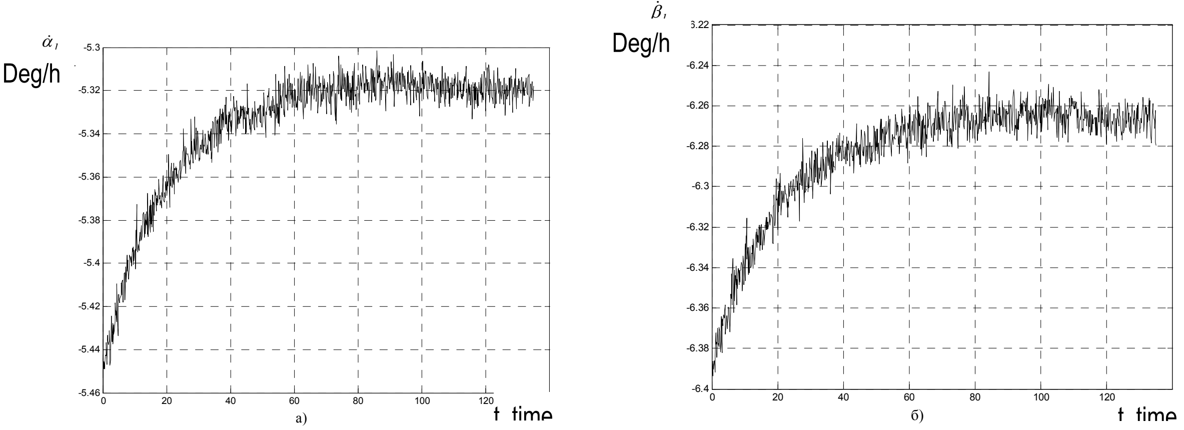


Fig. 1. Dependence of DTG drift at external temperature change on two channels.

The carried out statistical testing of coefficient values a_i, b_i according to the resulted technique has shown, that with the probability of 95 %, regression coefficients of the equations (2) at identical initial conditions are constant.

Function of the gyroscope drift, which approximated by polynomial (1), can be inadequate to observable values of the drift. Therefore it is necessary to check up its adequacy to the experiment data with the help of calculations deviation estimation of function values (2) from experimentally established ones, which are averaged by the number of experiences at factorial space points. For deviations estimation, Fisher's criterion is used.

In the table 1 the results of verification of statistical significance of regression coefficients estimation are adduced. It is seen, that the model (2) is adequate to the experiment at a significance value of $q = 0.05$, since $F_{calc} < F_{tab}$.

The obtained and statistically estimated mathematical model of gyroscopes thermal drift has shown its adequacy to physical process that allows using it for solution tasks of algorithmically increasing the accuracy of gyrocompass.

Verification of DTG drifts has confirmed their repeatability. Let's determine interconnection function between drifts $\dot{\alpha}_1, \dot{\beta}_1$ of a gyroscope in its different channels.

Table 1. Verification of model adequacy

Statistical characteristic	Gyroscopes orientation			
	$\vec{H}$ Vertical		$\vec{H}$ Horizontal	
	I channel	II channel	I channel	II channel
$S^2(\deg/h)^2$	1,975 10^{-2}	1,144 10^{-2}	5,43 10^{-2}	5,83 10^{-2}
$S_{calc.}{}^2(\deg/h)^2$	4,1969 10^{-2}	1,9837 10^{-2}	6,7338 10^{-2}	7,1884 10^{-2}
F_{calc}	2,125	1,734	1,566	1,233
Tabulated value F - criterion at significance value *q*=0.05, $F = 2.52$.				

Neural network algorithm

The possibility of using neural network algorithm to determine the interconnection between the DTG's two drifts in both its channels is approved by training the neural net work and defining weight coefficients and biases, the algorithm's input and output are $\dot{\alpha}_1 = \omega_z{}^{dr}$ and $\dot{\beta}_1 = \omega_x{}^{dr}$ accordingly

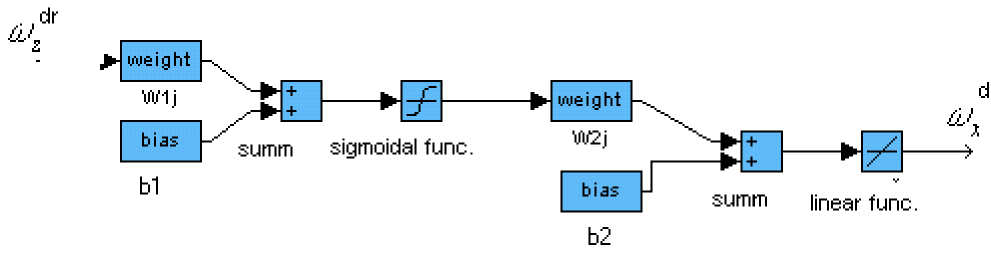


Fig. 2. Model of neural network of a straight propagation for DTG drift approximation

$$\omega_x{}^{dr}(t) = \sum_{j=1}^{N} W_{2j} \cdot (sigm(W_{1j} \cdot \omega_z{}^{dr}(t) + b_{1j})) + b_2\,, \tag{3}$$

where W_{1j}, W_{2j} – weight coefficients; b_1, b_2 – biases, N- number of neurones in net's hidden layer.

The teaching algorithm of neuronet are the next.

1-st step. Weights of the net are given small initial values.

2-d step. The next teaching pair (X, Y) are selected from the teaching ensemble; vector X is delivered to net's input.

3-d step. The output of the net is calculated.

4-d step. The difference between the required (target, Y) and the real (calculated) net's output is calculated.

5-th step. Weights are adjusted so, to minimize the accuracy (in the beginning the weights of the output layer, then with the use of differentiation complicated functions rule and the above mentioned derivative sigmoidal function, then the weights of previous layer and son on)

6-s step. Steps from the 2-nd to the 5-th are repeated for each pair of the teaching ensemble until the error in all ensembles does not reach the acceptable value.

Steps 2 and 3 similar to that carried out in the taught yet net.

The experimental investigations showed that the absolute error of estimating one drift by using the interconnection function (3) and the known other drift is less then 2%.

Increasing the accuracy of gyrocompass

Confining the analyze of those gyrocompass errors that caused by gyroscope drifts only, then the precession equations of gyrocompass motion when the ship is moving at a constant speed, heading and without heaving are [Zbrutsky, Nesterenko, Prokhorchuk, Lukjanenko, 1997]:

$$\begin{aligned} &\dot{\alpha} - \omega_{\eta}\beta = k_x\delta + \omega_z^{dr}; \\ &\dot{\beta} + \omega_{\eta}\alpha = -k_z\delta + \omega_x^{dr}; \\ &\mathrm{T}_n\dot{\delta} + \delta = \beta, \end{aligned} \tag{4}$$

where α , β – deviation angles of the gyrocompass principal axis from the meridian and horizon areas accordingly (α – gyrocompass error); δ – output signal of accelerometer amplifier; T_n – constant time of accelerometer amplifier; ω_{η} – angular velocity northern projection of geographic accompanying trihedron turn; k_x , k_z – torque's pendular and damping slopes; $\omega_{\mathrm{z}}^{\mathrm{dr}}$, $\omega_{\mathrm{x}}^{\mathrm{dr}}$ – gyroscopes drifts angular velocities around vertical and horizontal axes.

The application of Kalman optimal filter method to a gyrocompass in its standard statement (4) shows, that the positive result of increasing the accuracy cannot be achieved because of the nonobservability system. But the use of the offered interconnection function between gyroscope drifts (3) and the application of Kalman optimal filter method allows to estimate the heading identification errors and algorithmically compensate them.

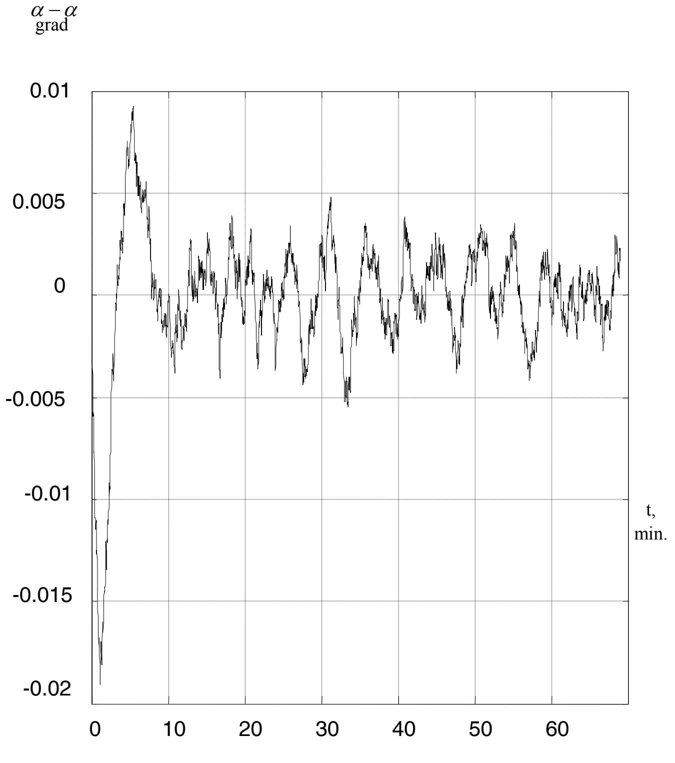


Fig. 3. Kalman optimal filter estimation of the gyrocompass heading error

Then (4) will be

$$\begin{aligned}
&\dot{\alpha} - \omega_{\eta}\beta = k_x\delta + \omega_z^{dr};\\
&\dot{\beta} + \omega_{\eta}\alpha = -k_z\delta + f\left(\omega_z^{dr}\right);\\
&\mathrm{T}_n\dot{\delta} + \delta = \beta;\\
&\dot{\omega}_{\mathrm{z}}^{\mathrm{dr}} = W_1 .
\end{aligned} \qquad (5)$$

where $f(w^{dr}{}_z)$ = $w^{dr}{}_x$ - horizontal drift as function of vertical drift, W_1 - white noise.

It is obvious from the fig. 3, that the estimated fault of the gyrocompass heading error, using the Kalman optimal filter and the proposed interconnection function between the two drifts of DTG, does not exceed 0.02 degrees and the standard deviation of this error is less than 0.003 degrees.

Conclusion

Using the above discussed method we assume that it can be used for improving accuracy characteristics of many navigation systems where two one-component sensor are used in the same device in different positions such as accelerometers, strap down inertial systems etc.

Bibliography

[Heerman ,1992] P.D.Heerman. Neural network techniques for stable learning control of nonlinear systems. Dissertation D.S.University of Texas at Austin, 1992.

[Rauch, Winarske, 1988] H.E.Rauch, T.Winarske. Neural networks for Routing Communication Traffic. In: IEEE Control Syst. Mag., 1988, vol.8.

[Zbrutsky, Nesterenko, Prokhorchuk, Lukjanenko,1997] A.V.Zbrutsky, O.I.Nesterenko, A.V.Prokhorchuk, N.V.Lukjanenko. Small–sized integrated system for the sea mobile objects attitude and navigation. In: Symposium Gyro Technology 1997, Stuttgart, Germany. 1997.

Authors' Information

***Aleksandr Zbrutsky** – Dean of the Faculty, Professor, National Technical University of Ukraine "Kyiv Politechnic Institute", P.O.Box: 37, Peremogy Av., Kyiv, 03056, Ukraine; e-mail: zbrutsky@cisavd.ntu-kpi.kiev.ua*

***Mohamed Rahmouni** – Researcher; National Technical University of Ukraine "Kyiv Politechnic Insuituite", P.O.Box: 37, Peremogy Av., Kyiv, 03056, Ukraine; e-mail: faks@ntu-kpi.kiev.ua*

A DNA CODIFICATION FOR GENETIC ALGORITHMS SIMULATION

Ángel Goñi, Francisco José Cisneros, Paula Cordero, Juan Castellanos

Abstract: *In this paper we propose a model of encoding data into DNA strands so that this data can be used in the simulation of a genetic algorithm based on molecular operations. DNA computing is an impressive computational model that needs algorithms to work properly and efficiently. The first problem when trying to apply an algorithm in DNA computing must be how to codify the data that the algorithm will use. In a genetic algorithm the first objective must be to codify the genes, which are the main data. A concrete encoding of the genes in a single DNA strand is presented and we discuss what this codification is suitable for. Previous work on DNA coding defined bond-free languages which several properties assuring the stability of any DNA word of such a language. We prove that a bond-free language is necessary but not sufficient to codify a gene giving the correct codification*

Keywords: *DNA Computing, Bond-Free Languages, Genetic Algorithms.*

ACM Classification Keywords: *I.6. Simulation and Modelling, B.7.1 Advanced Technologies, J.3 Biology and Genetics*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Since the beginning of computation, John Von Neumann held that the different machine models should try to imitate the functions which take place in living beings. Recently, two paradigms of biological inspiration are being applied very satisfactorily to the resolution of problems: neural nets and genetic algorithms. Nowadays, computer scientist try to go a little bit further by working with the same row material the nature does. That is the case of Leonard Adleman who is the pioneer in this field and solved a problem using real DNA strands. In a short period of time DNA based computations have shown lots of advantages compared with electronic computers. DNA computers could solve combinatorial problems that an electronic computer cannot like the well known class of NP complete problems. That is due to the fact that DNA computers are massively parallel [Adleman, 1994].

Despite all the impressive benefits that DNA computations have, they also have several drawbacks. The biggest disadvantage is that until now molecular computation has been used with exact and “brute force” algorithms. It is necessary for DNA computation to expand its algorithmic techniques to incorporate aproximative and probabilistic algorithms and heuristics so the resolution of large instances of NP complete problems will be possible. Without algorithms DNA computing has linear time solving NP-Complete problems but exponential space.

DNA, deoxyribonucleic acid, is the main motor which moves this new computer paradigm. A DNA molecule consists of two single strands twisted. Each strand is a long polymer of bases. Four different bases are presented in DNA: adenine (A), thymine (T), cytosine (C) and guanine (G). It is the sequence of these four bases that encodes information

Leonard Adleman [Adleman, 1994], an inspired mathematician, began the research in this area by an experiment using the tools of molecular biology to solve a hard computational problem in a laboratory. That was the world’s first DNA computer. A year later Richard J.Lipton [Lipton, 1995] wrote a paper in which he discusses, in detail, many operations that are useful in working with a molecular computer. After this moment many others followed them and started working on this new way of computing.

Adleman’s experiment solved the travelling salesman problem (TSP). The problem consists on a salesman who wants to find, starting from a city, the shortest possible trip through a given set of customer cities and to return to its home town, visiting exactly once each city. TSP is NP-Complete (these kinds of problems are generally believed cannot be solved exactly in polynomial time. Lipton [Lipton, 1995] showed how to use some primitive DNA operations to solve any SAT problem (satisfiability problem) with N binary inputs and G gates (AND, OR, or NOT gates). This is also a NP-Complete problem.

Genetic Algorithms (GA's) are adaptive search techniques which simulate an evolutionary process like it is seen in nature based on the ideas of selection of the fittest, crossing and mutation. GAs follow the principles of Darwin's theory to find the solution of a problem. The input of a GA is a group of individuals called initial population. The GA following Darwin's theory must evaluate all of them and select the individuals who are better adapted to the environment. The initial population will develop thanks to crossover and mutation.

John Holland in 1975 was the first one to study an algorithm based on an analogy with the genetic structure and behaviour of chromosomes. The structure of a basic genetic algorithm includes the following steps. (1) Generate the initial population and evaluate the fitness for each individual, (2) select the best individuals, (3) cross and mutate selected individuals, (4) evaluate and introduce the new created individuals in the initial population. All those steps together are called a generation.

Before generating the initial population, individuals need to be coded. That is the first thing to be done when deal with a problem so that it can be made combinations, duplications, copies, quick fitness evaluation and selection. Nature is a big genetic algorithm in which we are the individuals of the problem. Each of us is coded as a base sequence. We are all different form each other thanks to that sequence. A concrete code must have some characteristics that identify the individual to be more or less qualified.

A GA receives an initial population and after several generations, some codes will disappear and others will appear more often. That is how we can get the solution of the problem without exploring the complete search space.

Previous work on molecular computation for genetic algorithms [J.Castellanos, 1998] shows the possibility of solving optimization problems without generating or exploring the complete search space. A recent work [M.Calviño, 2006] produced a new approach to the problem of fitness evaluation declaring that the fitness of the individual should be embedded in his genes (in the case of the travelling salesman problem in each arch of the path). In both cases the fitness will be determined by the content in G+C (cytosine + guanine) which implies that the fitness of an individual will be directly related with the fusion temperature and hence would be identifiable by spectophotometry and separable by electrophoresis techniques [Macek 1997] or centrifugations.

Gene Characterization

In DNA computing is very important to know that instability of DNA strands can cause undesirable reactions. When facing a problem of any kind, one of the most important things to do is assuring that the data the problem will use is stable. It does not matter whether we are working with DNA or not to carry out this task. During the next sections we will tackle the problem of data encoding in DNA computing problems, concretely, in a simulation of a genetic algorithm with DNA.

The input data for DNA computing must be encoded into single or double DNA strands. Many conditions can cause loss of DNA bases or strand breakage and due to the Watson-Crick complementarity's parts of single DNA strands can bind together forming a double-stranded DNA sequence. Also, several DNA operations like electrophoresis or isopycnic centrifugation, which are absolutely essential for a correct DNA computation, are based on certain characteristics of the DNA strands. Those characteristics can not be altered if we want carry out a problem with DNA. For example, in the case of genetic algorithms, electrophoresis helps us to select the better adapted individuals. That operation is essential for the process of selection of the fittest and we have to take it into account we generating our data: the genes.

We must take care of all these conditions and characteristics so that we can assure the stability of every data of our problem. We can not choose our data randomly making long sequences of bases (A, C, G, and T) because as bigger is our initial data set, more mistakes we will find during the computation.

Taking all into account, we can distinguish two different problems: codification of a stable DNA language and codification of the genes.

Codification of a DNA language.

First of all we have to recall a list of known properties of DNA languages which are free of certain types of undesirable bonds and give a solution as a uniform formal language inequation [Lila Kari, 2004]. That is to create a language from which you can choose any word and be sure of the stability of that DNA molecule.

Codification of the genes.

We want to highlight the possibilities that offer the storage of the information in genes, one word is saved in a different gene, and these genes possess numerous properties (weight, size, ability). Some of the most precise operations that we can realize with the DNA are based on these properties. In our simulation each gene has a different amount of C+G bases. That condition identifies each gene.

When simulating a genetic algorithm, the individuals are formed by several genes and each gene has its own information or characteristic. This characteristic is the content of Cytosine and Guanine they have. An individual would have this aspect:

PCR-primer Np Rep XY RE0 XY RE1 … REn-1 XY Rep Np-1 PCR-primer

XY (gene) is better evaluated as more C+G content

Were the beginning of the individual and the end of it is the almost the same sequence of bases (PCR-primer Np Rep) and the different genes are separated by restriction enzymes (RE).

In this way the individuals are already evaluated. Once they are evaluated we must select them. By isopycnic centrifugation we can select the best suited to their environment. This technique is used to isolate DNA strands basing on the concentration of Cytosine and Guanine they have. The relationship between this concentration and the density (θ) of the strand is:

$$\Theta = 0{,}100[\%(G+C)] + 1{,}658$$

To begin the analysis, the DNA is placed in a centrifuge for several hours at high speed to generate certain force. The DNA molecules will then be separated based primarily on the relative proportions of AT (adenine and thymine base pairs) to GC (guanine and cytosine base pairs), using θ to know that proportion [Gerald Karp, 2005]. The molecule with greater proportion of GC base pairs will have a higher density while the molecule with grater proportion of AT base pairs will have a lower density. In this way the different individuals (different paths or solutions of TSP) are separated and can be easily selected.

With this technique we can identify one gene from the rest. But this work would be useless unless we codify the rest of the gene using a stable DNA language. If not, the genes we define in a genetic algorithm could be altered due to DNA instability.

The first problem can be solved using bond-free languages [Bo Cui, 2007] but those kind of languages do not allow us to codify the genes. We can not choose a word of that language, a sequence of nucleotides, to make a gene because it won't be different from the rest. These methods give a language which assures that any word w_1 of such a language is stable and won't bind together another word w_2 of the same language , but does not assign a proportional weight to the different words of the language. This implies that the genes could not be arranged by weight, preventing from realizing operations with DNA of that properties of the language could take advantage directly.

Every genetic algorithm will need different genes. The genes are the initial data and they must be well defined in order to obtain the correct solution to the problem. However, all the genes would have a similar format. This format must solve both of the problems, codification of a DNA language and codification of the genes. We will use a bond-free language to define most part of the gene but we will add some other characteristic (the fitness of the gene) that must be suitable for a concrete problem. This would be the aspect of a gene:

Bond-Free DNA language (w1) –- Fitness(w3) –- Bond-Free DNA language(w2)

Gene Encoding Language

Words w1, w2, w3: nucleotide sequences

For the language used to surround the fitness we will use a bond-free language. This language assure stability taking care of several conditions like temperature, complementarity, sequences of the same base, concentration of G+C, etc. Any word we take from a bond-free language can be optimal for this task, taking into account that all the words must be different to create different genes. A concrete problem will tell us how long this word must be and how many nucleotides we will use to make this words.

That kind of language could be useful in the resolution of DNA problems that uses ‘brute force’. That is the case of Adleman’s experiment. But if we try to go further in exploring the possibilities of DNA computing we must use algorithms like, for example, genetic algorithms. To complete the Gene Encoding Language we use a certain characteristic (fitness) which is outside ‘stable’ languages but are necessary to give a weight to each gene.

This characteristic which makes a gene different from the rest of the genes of the problem is based on the concentration of Cytosine and Guanine they have. Here also the problem will tell us how long this word must be. Anyway, this word should be much smaller than the other parts preserving the stability.

Example

Now we are going to establishing the notation that would allow us to describe the formalizations. Specifically, we define the terms *node, fitness, gene and language of genes*. For this example we will use the well known Travelling Salesman Problem (TSP), see figure 1. If the salesman starts at city X1, and it he distances between every city are known, what is the shortest path which visits all cities and returns to city X1?

We define *Xi* as a node of the graph (a city), and *Wij* as the length or fitness of a archers in the graph (the distance of the road).

We consider a gene as the minimal data unit of our problem and it is denoted by *Yi*. *Yi* is based on three parameters *{ Xi + Wij + Xj}* which are three different nucleotides sequence. The number of different genes in a problem is always the same as the number of arches in the graph.

$$Yi = \{ Xi + Wij + Xj \} \qquad I,j = 1..n$$

Once we codify all the genes, we have an alphabet of genes which can be used to form paths. A path is composed by a sequence of genes and they are possible solutions of the problem. We represent a path by Zi.

We codify *Xi* using a bond-free language, and *Wij* using a special language that preserves the weight of the gene. In this case, that special language gives each gene a different concentration of C+G.

More concentration of guanine and cytosine represent a shorter way. On the other hand, the lack of these nucleotides means that the gene represents a long way between two cities.

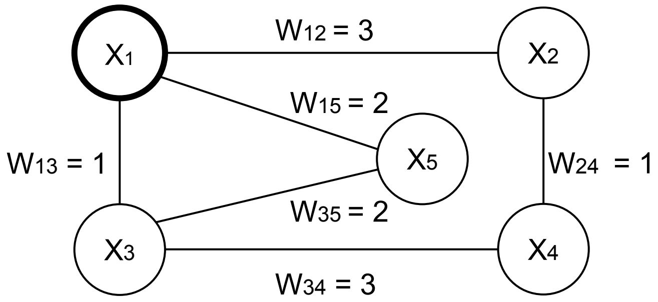


Fig. 1

Fig. 1 represents a map with five cities. Several cities are connected by roads of length *W_{ij}.* The city *X_1* is the initial city. Five nodes are represented *{X_1, X_2, X_3, X_4, X_5}* and only six roads *{W_{12}, W_{13}, W_{24}, W_{34}, W_{45}, W_{15}}*.

X1	=	AATT		
X2	=	ATAT	GGGCCC	= 1
X3	=	TAAA	GGTTCC	= 2
X4	=	TTTA	GTTTTC	= 3
X5	=	AAAA		

Fig. 2

In figure 2 we assign each city a different nucleotide sequence based on a bond-free language in order to make our data set stable. Also we assign sequences to the roads. The sequences of the roads are not chosen randomly from a certain language. They are chosen by the rule explained before: as shorter a road is, more concentration of C+G that gene would have. Only three values are possible for the roads as they are one, two or three kilometres long.

To represent the cities we use a 4-base sequence and to represent the roads we use a 6-base sequence. A gene is the conjunction of a departure city, a road and the arrival city. In this case there are only six different genes. Those genes are:

Y_1	=	X_1	W_{12}	X_2	=	AATT	GTTTTC	ATAT
Y_2	=	X_2	W_{24}	X_4	=	ATAT	GGGCCC	TTTA
Y_3	=	X_4	W_{34}	X_3	=	TTTA	GTTTTC	TAAA
Y_4	=	X_3	W_{13}	X_1	=	TAAA	GGGCCC	AATT
Y_5	=	X_3	W_{35}	X_5	=	TAAA	GGTTCC	AAAA
Y_6	=	X_5	W_{15}	X_1	=	AAAA	GGTTCC	AATT

Fig. 3

The possible solutions of the problem are represented by a sequence of genes which we call a path. A possible path for the TSP shown in figure 1 is the next;

Y_1	Y_2	Y_3	Y_5	Y_6

Way 1

X_1	W_{12}	X_2	X_2	W_{24}	X_4	X_4	W_{34}	X_3	X_3	W_{35}	X_5	X_5	W_{15}	X_1
AATT	GTTTTC	ATAT	ATAT	GGGCCC	TTTA	TTTA	GTTTTC	TAAA	TAAA	GGTTCC	AAAA	AAAA	GGGGCC	AATT

Detail Way 1

Fig. 4

When all the paths are formed in a soup they do not bind together and they do not disappear because of strand breakage. That is due to the fact that this gene encoding uses a bond-free language. Also, they can be easily identified by several DNA operations like gel electrophoresis which select those paths with more C+G. Using this technique will give us the shortest path of the problem

Conclusion

One of the problems DNA computing has is the instability of the DNA strands. Many conditions can cause loss of DNA bases or strand breakage and because of that the problem will probably be doomed to failure. We propose a codification of data when solving a genetic algorithm with DNA. That is, a codification of the genes. In this codification we take care of the conditions that can cause DNA instability and, at the same time, it is preserved the identity of each gene. As a result of such a process we have a DNA codification forming genes that assures stability of DNA and give a concrete property to each gene.

That property (fitness) which makes a gene different from the rest of the genes of the problem is based on the concentration of Cytosine and Guanine they have. However, the base sequence which forms the fitness is a small part of the DNA strand which represents a gene. For assuring DNA stability we use a bond-free language to complete the gene.

Bibliography

[Adleman, 1994] Leonard M. Adleman. Molecular Computation of Solutions to Combinatorial Problems. Science (journal) 266 (11): 1021-1024. 1994.

[Adleman, 1998] Leonard M. Adleman. Computing with DNA. Scientific American 279: 54-61. 1998

[Lipton, 1995] Richard J.Lipton. Using DNA to solve NP-Complete Problems. Science, 268:542-545. April 1995

[Holland, 1975] J.H.Holland. Adaptation in Natural and Artificial Systems. MIT Press. 1975.

[J.Castellanos, 1998] J.Castellanos, S.Leiva, J.Rodrigo, A.Rodríguez Patón. Molecular computation for genetic algorithms. First International Conference, RSCTC'98.

[M.Calviño, 2006] María Calviño, Nuria Gómez, Luis F.Mingo. DNA simulation of genetic algorithms: fitness computation. iTech 2006. Varna, Bulgary.

[Macek , 1997] Milan Macek M.D. Denaturing gradient gel electrophoresis (DGDE) protocol. Hum Mutation 9: 136 1997.

[Dove, 1998] Alan Dove. From bits to bases; Computing with DNA. Nature Biotechnology. 16(9):830-832; September 1998.

[Mitchell, 1990] Melanie Mitchell. An Introduction to Genetic Algorithms. MIT Press, Boston. 1998.

[Lee, 2005] S.Lee, E. Kim. DNA Computing for efficient encoding of weights in the travelling salesman problem. ICNN&B'05. 2005.

[SY Shin, 2005] SY Shin, IH Lee, D Kim, BT Zhang. Multiobjective evolutionary optimization of DNA sequences for reliable DNA computing. IEEE Transactions, 2005.

[Bo Cui, 2007] Bo cui, Stavros Konstantinidis. DNA Coding using the Subword Closure Operation. DNA 13. 13th Internacional Meeting on DNA Computing

[Kari, 2005] Lila Kari, Stavros Konstantinidis, and Petr Sosík. Bond-Free Languages: Formalizations, Maximality and Construction Methods. DNA10, LNCS 3384, pp. 169–181, 2005

[Konstantinidis, 2007] Bo Cui and Stavros Konstantinidis. DNA Coding using the Subword Closure Operation, DNA13, pp. 65–74, 2007.

[Kari, 2005] Kari, L., Konstantinidis, S., and Sosík, P. (2005) Preventing undesirable bonds between DNA codewords. Lect. Notes Comput. Sc., 3384, 182-191.

[Jonoska] Jonoska, N., Mahalingam, K.: Languages of DNA based code words. In: [4], 58–68

Authors' Information

***Ángel Goñi Moreno** – Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain: e-mail: ago@alumnos.upm.es*

***Fco. Jose Cisneros de los Rios** – Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain: e-mail: kikocisneros@gmail.com*

***Paula Cordero** – Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain: e-mail: paula.cormo@gmail.com*

***Juan Castellanos** – Natural Computing Group. Universidad Politécnica de Madrid, Boadilla del Monte, 28660. Madrid, Spain. e-mail: jcastellanos@fi.upm.es*

DISTRIBUTED GENETIC ALGORITHM IMPLEMENTATION BY MEANS OF REMOTE METHODS INVOCATION TECHNIQUE – JAVA RMI

Lukasz Maciura, Galina Setlak

Abstract: *The aim of this work is distributed genetic algorithm implementation (so called island algorithm) to accelerate the optimum searching process in space of solutions. Distributed genetic algorithm has also smaller chances to fall in local optimum. This conception depends on mutual cooperation of the clients which realize separate working of genetic algorithms on local machines. As a tool for implementation of distributed genetic algorithm, created to produce net's applications Java technology was chosen. In Java technology, there is a technique of remote methods invocation - Java RMI. By means of invoking remote methods it can send objects between clients and server RMI.*

Keywords*: Genetic Algorithms and Evolutionary*

ACM Classification Keywords*: G.1.6 Optimalization – Global optimalization*

Conference*: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

To accelerate the optimum searching process in space of solutions for genetic algorithm, distributed genetic algorithm (so called island algorithm [Schaefer, 2002]) was implemented. On the whole, a characteristic feature of classical genetic algorithm is that it has small chances to fall in local optimum. It is very positive feature which distinguish it from the other heuristic algorithms, but nothing is without a defect. Unfortunately, this algorithm has this negative feature, that searching of global optimum lasts much longer than in other heuristic algorithms. It is essential to aim at the acceleration of these algorithms.

To repeatedly accelerate working of any algorithm it is necessary to parallelize or distribute it. Parallelization depends on this, that application which realizes this algorithm has a lot of threads and each of them realizes the separate kind of working. In order to parallelization leads to acceleration of the algorithm this machine on which the application is running must have a lot of processors or multi-threading processor, in such a way that each thread can be run on separate processor or core of processor. Distributing of an algorithm relies on working which is divided on a lot of machines and each of them realizes its separate part. These machines communicate witch each other through the local net or the Internet. The most often there is also the main server on which there are common resources and which manages the work of whole distributed system. Distribution of algorithm has this advantage in comparison to its parallelization that the amount of machines in net unlimited, however, in multiprocessor machine the more processors there is, the more complications occur with the selection of the proper hardware to operate with any amount of processors. Regarding to this distribution of the algorithm not its parallelization was chosen.

Nowadays, there are a lot of technologies which assist in creation of distributed systems. Some of them are independent of platform and programming language as DCOM, CORBA, others are created for specific programming language or platform as RMI mechanism in Java technology [Horstmann, 2003, 2005] or Remote mechanism in .NET platform [Troelsen, 2006]. There is also a possibility of using ordinary TCP/IP sockets but it would be a work from basis in coding/decoding of objects and its packetizing through net, so the best way is to use already checked solution, so as a technology to work out distributed system, in this work Java and its mechanism of Remote Method Invocation (RMI) was chosen.

Working of local genetic algorithm

Genetic algorithm belongs to the group of heuristic algorithm [2], which does not search whole space of solutions, but they work systematically going in some direction or directions of searching, which in particular moment seems to be the most optimal. To the group of heuristic algorithms belong, among other things: Taboo search, ant's algorithms, evolutionary algorithms. Most of these techniques were created on the basis of observation of nature and man. Evolutionary and genetic algorithms were created on the basis of transferring nature evolution methods on computer science area. The area of working genetic algorithms is, among other things, solving optimization problems.

The whole idea of this solution depends on the existence of population of specific amount of individuals and each of them has one or several chromosomes which are a sequence of bits or other data which represent single genes, thanks to which they can intersect with each other and be mutated. Each of the individuals presents a specific solution of the problem which is suitably coded in a chromosome. Besides a chromosome, each of the individuals has a function of the adaptation which determines, which of the individuals (solution of the optimization problem) are better and which are worse. The individuals or descendants of the individuals which have the best function of the adaptation, have the highest chance of passage to next epoch, however, the individuals or descendants of the individuals which have a worse function of the adaptation have small or none chances depending on a method of the selection which was applied. Thanks to it every next epoch we have better and better collection of the individuals – the evolution of whole population lasts. Thanks to this strategy, separate solution is not favouring but a lot of the best solutions that lover chance of falling in local optimum.

Local genetic algorithm that work on single client's machine in presented in this work distributed system, works in the same way as a classical genetic algorithm, with such a difference that sometimes the amount of individuals in population is higher, when taking of the best individuals from server which came from other clients occurred. As a problem of genetic algorithm, necessary to test its working searching of maximum function of two variables which possess a lot of local maximums [3] was chosen.

$$F(x,y) = 2000 - 64\left(\sin\left(\frac{x*\pi}{16}\right) + \sin\left(\frac{y*\pi}{16}\right)\right) - 0.185*((x-64)^2 + (y-64)^2)$$

Values x and y coded in binary chromosome belong to range of <0,128>. If we assumed that C_1 - C_n are genes of chromosome, so values x and y [3] can be work out from formulas:

$$x = \sum_{i=1}^{n} c_i * 2^{-i} * 128 \qquad\qquad y = \sum_{i=n+1}^{n} c_i * 2^{(n-i)} * 128$$

As a selection technique to the next epoch of individuals designed for reproduction the most popular method – roulette was applied. It depends on application of virtual roulette in which each of the individuals has its own segment proportional to value of its function of adaptation. In practice there are ranges of real values from range <0,1>. Then, there is a drawing of a value from this range and checking to what range it belongs. An individual which is associated with this range is admitted to the reproduction. Depending on this, if intersection follows or not, either it or its descendants came to the next epoch. Reproduction occurs always on sorted in this way individuals, regarding to this, if the intersection occurs (the random to this aims value is smaller than probability of the intersection) the descendants of the parents move to the next epoch. However, if the intersection do not occurs, parents move to the next epoch. If the intersection occurs, the position of the intersection is randomized and the first of the descendants receives a fragment of the chromosome from the beginning of the position of the intersection from the first parent, however, from the second parent it receives a fragment of the chromosome from the position of the intersection to the end of the chromosome. This algorithm uses a selection of the parental pool and creates new individuals, as long as the population of the new epoch files. With every reproduction there is some probability that after carried or not carried out operation of intersection mutation occurs. To recapitulate, single genetic algorithm epoch on the local machine operates as the followings:

1. Work out the value of the function of adaptation for each of the individuals in the population.
2. If the best individual in this population is better than the best actual individual from the whole algorithm, then chose it as the best and set a flag which informs about this, that it has to be sending on the server.
3. Do genetic operators (intersection and mutation), as long as, a new population will create from nothing.

The description of the distributed genetic algorithm

The system created in this work bases on mutual communication of clients – agents that realize local genetic algorithm. This communication relies on an exchange of the best individuals among the agents. Each of the clients communicates with the server by means of mechanism of Remote Method Invocation – Java RMI. By means of an invocation of the appropriate methods (functions) on the server, it can place there its best individuals as well as take these which were left there by other agents. Mechanism of serialization of objects in Java technology allows on sending in this way whole structures of objects which are placed on RAM of the computer, not only to reference to them.

The description of the algorithm on client's side

The client contains two threads: thread A realizes an operation of genetic algorithm and thread B realizes a communication with the server and an exchange of individuals.

Algorithm on client side:

1) Client's thread B is logging to the server, invoking the remote method: int login(), and a name tag in the form of number is assign to it.
2) Thread B serially checks, by means of invoking on the server the remote method: boolean permission(), which turns a value 'true' if the expected amount of logged clients to a server will be achieved. In such a case algorithm comes to point 3.
3) To establish the individuals of the population in such a way that every now and then a new the best individual does not occur, thread A carries out specific amount of genetic algorithm epochs, and at the same time updates the best individual from the algorithm's start.
4) After carry out specific amount of epochs threads A and B start to work simultaneously (Fig. 1).

The description of the server

The server serves in this distributed system as a relay of the best individuals among clients. It makes registered on it remote methods available to communicate with clients and transfer of objects between them. Besides remote methods it has a table of individuals on which individuals from clients are saved. In this table each of the clients has assigned its index which receives in the time of logging. Besides this table there is also matrix of value logical type, on which there is saved which of the client load an individual which comes from another specific client. It will be needed in order to a given client does not load repeatedly the same individuals from the server which could overload server and whole algorithm. The best global individual is also saved on the server. After the start of the server, the amount of the clients which should log in to it should be inserting, in order to start whole algorithm after a log in all clients.

The description of the remote methods:

int login() - the method which is used to log in a client to the server and give to it a name tag

boolean permission() - the method which returns the value of the logical type 'true' if a client can start its algorithm and 'false' if not. It depends if the established earlier amount of the logged in clients is achieved and the flag 'start' is set.

boolean endcondition() - the method which returns the value 'true' if the condition of the end of the algorithm was fulfilled. This is established on the server and different conditions of the ending can be considered

void send(int ID,Individual i) – the method which is used to send the best individual to the server. It is placed in a table in a position determined by ID. Additionally, there will be set suitable logical values in matrix that specify which of the clients loads the individuals which comes from individual clients. In the whole column there are set values 'false' because the new individual has not been load by nobody, yet (Table 1). If necessary there is also actualized the best global individual on the server.

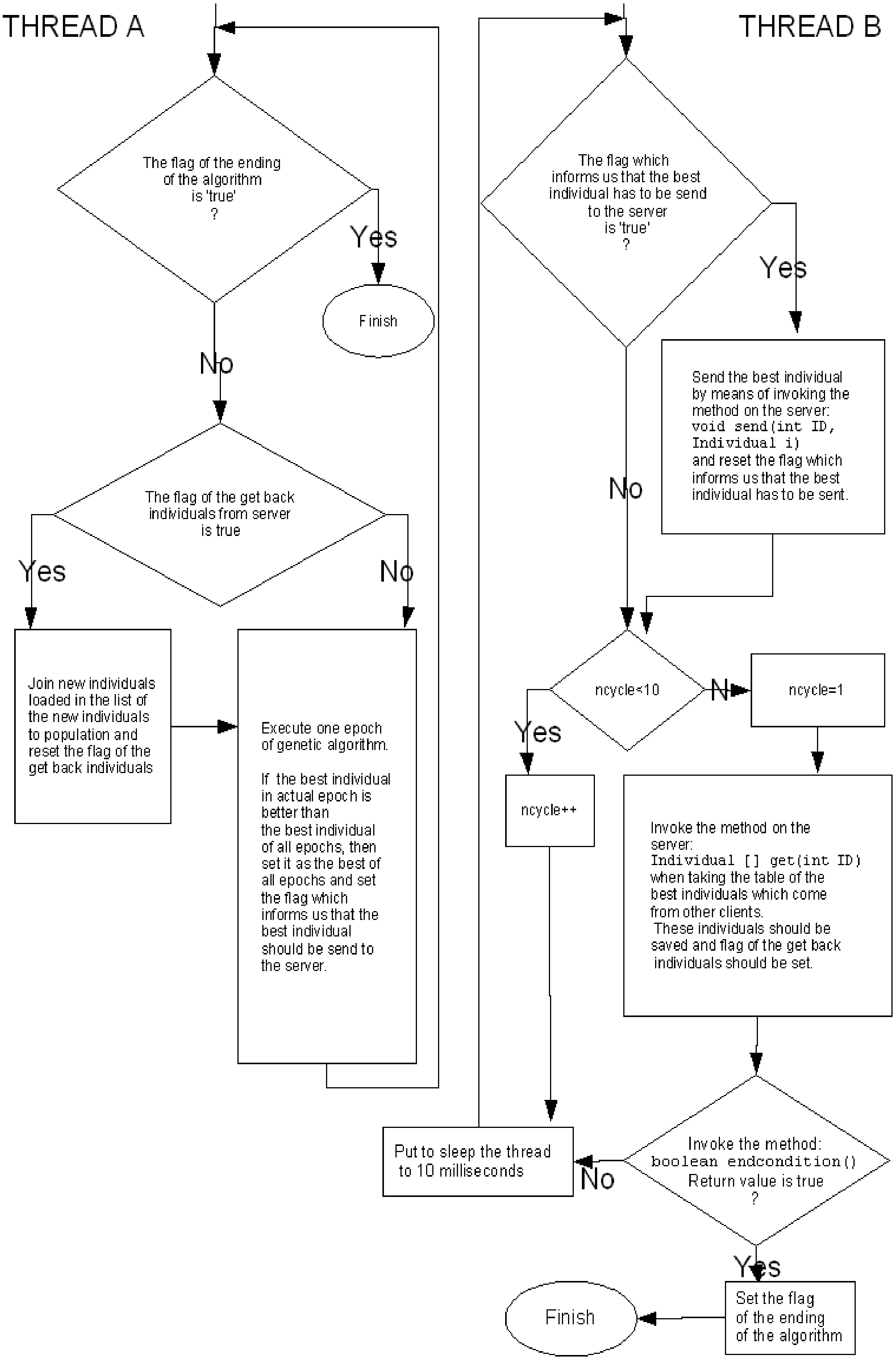

THREAD A
The flag of the ending of the algorithm is 'true' ?
Yes
Finish
No
The flag of the get back individuals from server is true
Yes
No
Join new individuals loaded in the list of the new individuals to population and reset the flag of the get back individuals
Execute one epoch of genetic algorithm.
If the best individual in actual epoch is better than the best individual of all epochs, then set it as the best of all epochs and set the flag which informs us that the best individual should be send to the server.
THREAD B
The flag which informs us that the best individual has to be send to the server is 'true' ?
Yes
Send the best individual by means of invoking the method on the server: void send(int ID, Individual i) and reset the flag which informs us that the best individual has to be sent.
No
ncycle<10
No
ncycle=1
Yes
ncycle++
Invoke the method on the server: Individual [] get(int ID) when taking the table of the best individuals which come from other clients. These individuals should be saved and flag of the get back individuals should be set.
Put to sleep the thread to 10 milliseconds
No
Invoke the method: boolean endcondition() Return value is true ?
Yes
Set the flag of the ending of the algorithm
Finish

Fig. 1. The algorithm of the client

	Individual 0	**Individual 1**	**Individual 2**	**Individual 3**
Client 0	*true*	***false***	*false*	*true*
Client 1	*true*	***false***	*true*	*true*
Client 2	*false*	***false***	*false*	*true*
Client 3	*false*	***false***	*true*	*false*

Table 1. The example of a content of this matrix after sending an individual by client no.1

```
public void send(int ID,Individual i) {
      if(ID>=0 && ID<n_clients && !theend) {
           table_of_individuals[ID]=i;
           for(int ind=0;ind<n_clients;ind++)
           matrix_of_loading[ind][ID]=false;
            i.Print();
      if(i.Value() > threshold) {
         System.out.println("The end, MAX Value="+i.Value());
         theend=true;
      }  }  }
```

Individual [] get(int ID) – the method which is used to load the table of the individuals saved by all other clients except of our own. ID is used to avoid an individual from actual client and to set the suitable value in the matrix that determines which of the individuals were loaded by specific clients. This matrix is especially needed here because it enables us to load a table only of these individuals which were not loaded by a given client (Table 2). Thanks to this it is not possible to load the same individuals by some client. The load is possible only when individual on a given position, that is, this from other client will change.

	Individual 0	**Individual 1**	**Individual 2**	**Individual 3**
Client 0	*true*	*false*	*false*	*true*
Client 1	***true***	***true***	***true***	***true***
Client 2	*false*	*true*	*false*	*true*
Client 3	*false*	*true*	*true*	*false*

Table 2. The example of a content of this matrix after loaded an individual by client no.1

Value 'false' means that a given individual has not been load by a given client, yet, however, 'true' means that there are non individuals yet or a given individual was loaded to a given client. Individual 0 comes from client 0, individual 1 from client 1 etc. When a new individual is send by client (e.g. Client 1) in the whole column 'Individual 1' values 'false' are written.

```
public Individual [] get(int ID) {
  if(ID>=0 && ID<n_clients && !theend) {
    ArrayList list=new ArrayList();
    for(int i=0;i<n_clients;i++) {
      if(!matrix_of_loading[ID][i]) {
        list.add(table_of_individuals[i]);
        matrix_of_loading[ID][i]=true;
      } }  Individual []tab=new Individual[list.size()];
    for(int i=0;i<list.size();i++)
      tab[i]=(Individual)list.get(i);
    return tab;
  } else return null; }
```

The algorithm on server's side:

1) Initiation of all the pools of the matrix that specify which of the clients loads the individuals which comes from individual clients to 'true', in order to block loading from server because there has been no individuals sent by clients, yet.
2) Loading of the required amount of logged in clients
3) Waiting, as long as, the required amount of logged in clients will be achieved.
4) Setting of the flag 'start' thanks to which the clients get to know through the 'permission' method that they can start.
5) In this moment the main programme of a server does nothing, besides of continue checking if the condition of the ending of the algorithm is not fulfilled. If it fulfils the best individual is introduced and the flag 'stop' is set. Remaining work is doing by remote methods invoked by clients on server's objects.

The system testing

To check to what extent the system increase the speed of finding the optimum in the space of solutions series of experiment, which depends on testing the working of an algorithm on many computers, was carried out and in which the amount of the computers was constantly increasing. On the server, there is a threshold. After a crossing of this threshold the algorithm finishes its working and displays the time of working. Thanks to this, we can compare periods of the calculation of the algorithm for different amount of clients which work on the separate computers. For a given number of computers 5 tests were made and median of working time was calculated.

1st series of the experiment:

Settings of the algorithm:

probability of intersection	0.6
probability of mutation	0.1
number of individuals	150
threshold of the finish of the algorithm	2107.417

	The number of clients			
Test	**2**	**3**	**4**	**5**
1	*32,5s*	*1,2s*	*0,7s*	*1,11s*
2	*2,2s*	*4,32s*	*0,7s*	*0,7s*
3	*7,05s*	*1,31s*	*0,8s*	*0,61s*
4	*5,53s*	*1s*	*2,22s*	*1s*
5	*3,11s*	*0,91s*	*0,61s*	*0,61s*
Median	***5,53s***	***1,2s***	***0,7s***	***0,7s***

Table 3. 1st series of the experiment

From these results (Table 3) it is noticeable that when the amount of clients working on separate machines is increasing the speed of finding of the optimum about function of adaptation higher from the given threshold is also increasing. However, it is noticeable that approximately for 4 of 5 computers the time of finding the optimum is the same. It is caused by this, that at the beginning of the algorithm's working there is a big movement in web because very often a new the best individual is found. It slows down the working of the algorithm, but, because finding of the optimum last in this cases short, so this delay is here very essential and it levelled them time of finding the optimum for 4-5 clients. In order to see the difference for a larger amount of computers we should cause the extension of the time of searching the space of the solutions. It can be caused by establishing the threshold of ending of the algorithm which is adequately higher.

2nd series of the experiment:

Settings of the algorithm:	
probability of intersection	0.6
probability of mutation	0.1
number of individuals	150
threshold of the finish of the algorithm	2107.4173

	The number of clients			
Test	**3**	**4**	**5**	**6**
1	*1,92s*	*1s*	*0,52s*	*0,61s*
2	*0,72s*	*1,41s*	*0,91s*	*1,31s*
3	*19,42s*	*1,11s*	*0,7s*	*0,61s*
4	*16,41s*	*0,52s*	*2,02s*	*0,91s*
5	*25,05s*	*0,7s*	*1,31s*	*0,92s*
Median	***16,41s***	***1s***	***0,91s***	***0,91s***

Table 4. 2nd series of the experiment

After this series of the experiments (Table 4) it is noticeable that when the amount of computers is increasing finding of the optimum speeds up, and when the time of searching is small for different amount of clients the changes are invisible.

Conclusion

The distributed model of genetic algorithm in Java technology was implemented. It accelerates the finding of the optimum in the space of the solutions. The speed of searching increases together with the amount of the clients which are working on the separate machines.

In the implemented example this algorithm solves the problem of searching the maximum of the function which can be written by means of mathematical formula, but nothing stands in the way to solve any other problem by this algorithm. It is implemented by means of objected technique of Java language, so it is easy to adapt this through the modification of some classes.

Bibliography

[Schaefer, 2002] R. Schaefer. Basics of global genetic optimization Ed. R. Schaefer. UJ Cracow, 2002

[Rutkowski, 2006] L. Rutkowski. Methods and techniques of artificial intelligence, PWN Warsaw, 2006

[Cytowski, 1996] J. Cytowski. Genetic algorithms. Basis and application, PLJ Warsaw, 1996

[Horstman, 2003] C. Horstman, G. Cornell, Core Java 2 (Volume I), Helion, Gliwice, 2003

[Horstman, 2005] C. Horstman, G. Cornell, Core Java 2 (Volume II), Helion, Gliwice, 2005

[Troelsen, 2006] A. Troelsen. C# language and the .NET platform, PWN, Warsaw, 2006

Authors' Information

***Lukasz Maciura** – PHD Student, The Bronislaw Markiewicz State School Of Higher Vocational Education in Jaroslaw, Czarneckiego Street 16, Poland; e-mail: l_maciura@pwszjar.edu.pl*

***Galina Setlak** – Ph.D., D.Sc, Eng., Associate Professor, Rzeszow University of Technology, Department Of Computer Science , Str. W. Pola 2 Rzeszow 35-959, Poland, Phone: (48-17)- 86-51-433, gsetlak@prz.edu.pl*

AN ADAPTIVE GENETIC ALGORITHM WITH DYNAMIC POPULATION SIZE FOR OPTIMIZING JOIN QUERIES

Stoyan Vellev

Abstract: *The problem of finding the optimal join ordering executing a query to a relational database management system is a combinatorial optimization problem, which makes deterministic exhaustive solution search unacceptable for queries with a great number of joined relations. In this work an adaptive genetic algorithm with dynamic population size is proposed for optimizing large join queries. The performance of the algorithm is compared with that of several classical non-deterministic optimization algorithms. Experiments have been performed optimizing several random queries against a randomly generated data dictionary. The proposed adaptive genetic algorithm with probabilistic selection operator outperforms in a number of test runs the canonical genetic algorithm with Elitist selection as well as two common random search strategies and proves to be a viable alternative to existing non-deterministic optimization approaches.*

Keywords: *genetic algorithms, query optimization, join ordering, randomized algorithms*

ACM Classification Keywords: *H.2.4. Query processing, H.3.4. Performance evaluation (efficiency and effectiveness)*

Conference: *The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Queries in a relational database management system (RDBMS) are defined in a declarative, non-procedural language, such as SQL. This raises the need to transform the declarative query into a procedural, effective plan for its execution. Each query can be mapped to a set of execution plans which are equivalent in terms of the result they generate but the execution cost of the different plans can vary by many orders. The execution plan is selected from the set of all alternatives by a dedicated RDBMS module – the Query Optimizer.

Due to the high processing cost, the evaluation of joins and their ordering are the primary focus of query optimization. Traditionally, the optimization of such expressions is done by complete traversal of the solution space (probably utilizing some pruning techniques). This is a possible approach for most of the classic database applications, where the size of the query (the number of joined relations) rarely exceeds 8-10, but it is completely inapplicable to some contemporary databases (Object-Oriented Databases, Multimedia Databases) and database applications such as Decision Support Systems (DSS), Online Analytical Processing (OLAP), Data Warehousing, Geographical Information Systems (GIS), etc. Queries in such applications may involve tens or even hundreds of joined relations.

This paper is focused on the optimization of a particular type of queries – single flat conjunctive queries, also known as selection-projection-join (SPJ) queries or non-recursive Horn clauses.

The Problem

Each query Q against a relational database defined by some data dictionary $\mathcal{D}$ with a set of relations $\mathcal{R}$ is represented by the ordered tuple (R^q, P^q), where $R^q = \{R_i \mid R_i$ is referenced in $Q\}$, $R^q \subseteq \mathcal{R}$ and $P^q = \{p_i(R^i_j, R^i_k) \mid p_i$ is a join predicate in Q, $R^i_j, R^i_k \in R^q\}$.

A *query execution plan* (QEP) of Q is a binary tree, in which the internal nodes represent join operator implementations (*join methods*), e.g. nested-loop join, merge join or hash join, and the leaves are base relations. Unlike the query itself (that has only declarative semantics), the query execution plan contains the procedural information about how to obtain the query result. Each execution plan has a *cost* that reflects the computational resources needed to evaluate it.

The problem is, given a query Q to find the execution plan (from the set of all equivalents) with the lowest cost that evaluates it (the global optimum). Since the combinatorial explosion makes the exhaustive solution search impossible, the aim will be restrained to finding a good *local* optimum.

Given a query Q of *n* relations against a database supporting a set $\mathcal{S}$ of different join methods, there are $\frac{1}{n}\binom{2(n-1)}{n-1}$ possible QEP tree structures, for each of it its *n* leaves can be ordered in *n*! different ways and each internal node is selectable from *s* different join methods, where *s* = | $\mathcal{S}$ |. In this work we limit our considerations to the space of *left-recursive* solutions (containing all trees for which it holds that each of their nodes has a base relation for a right successor), but there are still $n!s^{n-1}$ different solutions.

The problem of finding the optimum join order can be assumed a static optimization problem – although the database state is dynamic (it may change during the optimization of a query), the cost function depends on database statistics (rather than on the real-time database state) which can be considered static as they are updated in a controlled way and do not interrupt any ongoing optimizations.

There are a number of polynomial-complexity algorithms for solving some special cases of the problem. All of them however impose some major restrictions on the form of the queries, the type of the cost function used, the particular implementation of the join methods, etc. The join ordering problem in its general form is unfortunately $\mathcal{NP}$-complete.

Related Work

Due to the inapplicability of deterministic optimization algorithms (different variations of the classical dynamic optimization with pruning) to the join ordering problem for large queries (where *large* is usually defined as queries with 8 or more joins), the problem has been approached by two classes of non-deterministic algorithms – randomized and genetic.

Two well-known randomized algorithms have been applied to the problem of optimizing large join queries – *Iterative Improvement* and *Simulated Annealing*, as well as a combination of the two [6]. Though the effectiveness of randomized algorithms strongly depends on the shape of the solution space, they generally prove to be a possible alternative to deterministic search for large queries.

Genetic algorithms (GAs) have been first applied to query optimization in [5] and [4]. The fitness function used requires backward transformation from chromosome to tree representation, which is complex and with high computational cost. The chosen crossover operators have a serious flaw – they disrupt the chromosome structure, transforming two valid parent chromosomes into an invalid one, which then needs to be “repaired” to become a correct solution encoding. Despite these shortcomings, the achieved results are promising. Later in [3] some of these disadvantages have been overcome.

A theoretical comparison of randomized and genetic optimization algorithms concluded that many GAs (the canonical GA in particular) are characterized by higher probability of finding good solutions than randomized algorithms, as long as the solution space fulfills several restrictions. These restrictions however are weak and hold for almost any choice of the genetic operators [2].

Currently the only popular non-experimental genetic SQL query optimizer is the GEQO (GEnetic Query Optimizer) in the Postgres (PostgreSQL) RDBMS. It considers only left-recursive solutions, implements an Elitist selection operator, a simple edge recombination crossover and does not apply mutation. The population size is fixed.

Recently, self-adaptation in genetic algorithms (population size adaptation in particular) is receiving great attention [1]. However, no adaptive genetic algorithm has been applied so far to database query optimization.

The Algorithm

We introduce an adaptive genetic algorithm with dynamic population size as an efficient solution to the optimal join ordering problem.

Solution representation (Coding). The coding operator Θ transforms an individual ξ_j into a vector of genes, each gene being an ordered tuple of a relation number and a join method number:

$\Theta(\xi_j) = \Theta(R_{i1} \bowtie_{p1} R_{i2} \bowtie_{p2} \ldots \bowtie_{pn\text{-}1} R_{in\text{-}1} \bowtie_{pn\text{-}1} R_{in}) \longrightarrow ((i_1, p_1), (i_2, p_2), \ldots, (i_n, p_n))$

Mutation. The mutation operator M transforms an individual ξ_j into a new individual ξ'_i by swapping two randomly selected genes γ_x and γ_y and changing the join method of another randomly chosen gene γ_m:

$M(\xi_j) = M(((i_1, p_1), \ldots, (i_x, p_x), \ldots, (i_m, p_m), \ldots, (i_y, p_y), \ldots, (i_n, p_n))) \longrightarrow (((i_1, p_1), \ldots, (i_y, p_y), \ldots, (i_m, p'_m), \ldots, (i_x, p_x), \ldots, (i_n, p_n))); x, y, m \in \{1, 2, \ldots, n\}, x \neq y, p_m \neq p'_m$

The *mutation rate* μ is the probability of an individual ξ_j to mutate on each generation, i.e. $M(\xi_j) \equiv \xi_j$ with probability $(1 - \mu)$. The mutation operator is never applied to the individual with maximum fitness in the population.

Crossover. The crossover operator X combines the chromosomes of two generation-g individuals $\xi^g{}_i$ and $\xi^g{}_j$ to obtain two new generation-(g + 1) individuals $\xi^{g+1}{}_i$ and $\xi^{g+1}{}_j$. Random locus x is chosen, the two parent chromosomes are split at that locus and each of the two offspring receives a whole fragment from one of the parents (the first child - the left and the second child - the right relative to the locus) and the rest of the chromosome is filled up with the missing genes in the order they occur in the second parent. This guarantees both structural and functional similarity of the children with both their parents.

$X(\xi^g{}_i, \xi^g{}_j) = X(((i_1, p_1), \ldots, (i_x, p_x), (i_{x+1}, p_{x+1}), \ldots, (i_n, p_n)), ((j_1, q_1), \ldots, (j_x, q_x), (j_{x+1}, q_{x+1}), \ldots, (j_n, q_n))) \longrightarrow \{((i_1, p_1), \ldots, (i_x, p_x), (j'_{x+1}, p'_{x+1}), \ldots, (j'_n, q'_n)), ((i'_1, p'_1), \ldots, (i'_x, p'_x), (j_{x+1}, q_{x+1}), \ldots, (j_n, q_n))\}, x \in \{1, 2, \ldots, n\}$

Each individual ξ_j in the population selects a crossover partner from its *neighborhood*, defined as its k neighbors by index in the population vector, $k = const$. The probability of an individual ξ_j from the neighborhood to be chosen for a mating partner is proportional to its fitness $\varphi(\xi_j)$.

Selection. The selection operator Σ transforms one population ξ into another population $\xi' \subseteq \xi$:

$\Sigma(\xi) = \Sigma(\{\xi_1, \xi_2, \ldots, \xi_k\}) = \{\xi_{j1}, \xi_{j2}, \ldots, \xi_{jm}\}, m \leq k.$

We propose an algorithm called *Probabilistic Selection with Adaptive Population Size*. All classical selection algorithms keep the population size fixed. This simplifies the algorithms but it is an artificial restriction and does not follow any analogy to biological evolution, where the number of individuals in a population varies continuously in time, increasing when there are high-fit individuals and abundant resources and decreasing otherwise. Intuition hints that it may be beneficial for the population to expand in the early generations when there is high phenotype diversity and there is opportunity to "experiment" with different characteristics of the individuals, and to shrink with the increase of population convergence, when the unification of the individuals in terms of structure and fitness no longer justifies the maintenance of a large population and the higher computational costs associated with it.

The proposed algorithm achieves this control over the population size by defining a *collective* probability for the population to survive (i.e. to expand or to shrink), together with the *personal* probability of each individual to survive.

The personal survival probability for each $\xi_j \in \xi$ is defined as $p_i = \varphi(\xi_j) / \varphi^*$, where $\varphi(\xi_j)$ is the fitness of ξ_j and φ^* is the maximum fitness within the population. In other words, $\xi_j \in \xi'$ with probability p_i for each $\xi_j \in \xi$.

The size N of the population after selection can decrease (in the extreme case – to one, if just the individual with the maximum fitness survives), remain the same or increase (in the extreme case – to 3.N, if all parents and offspring survive). It is desirable that the population size never gets below the initial (i.e. generation-0) population size and the population increase is desirable to be inverse-proportional to the convergence degree of the individuals. The convergence degree can be measured by the ratio between the average fitness and maximum fitness in the population.

The *expected population size* s^E after selection can be roughly approximated by the sum of the survival probabilities of all individuals in the population[1].

The *desired population size* s^D is defined as a linear combination of the two extreme alternatives with coefficients depending on the convergence degree:

$s^D = s^0 c + 3N(1 - c)$,

$c = \varphi' / \varphi^*$,

where s^0 is the desired population size,

N is the current population size,

c is the convergence degree,

φ' is the average fitness and

φ^* is maximum fitness.

The *collective survival probability* p^* is defined as the normalized ratio between the desired and the expected population size ($p^* = s^D / (s^D + s^E)$).

The individual survival probability p_i is then scaled up or down depending on the collective survival probability p^*. In case the population size drops below s^0 in some generation, new random individuals are generated to fill it up to size s^0.

Note that with this algorithm the individual with the maximum fitness survives with probability 1.0, i.e. the proposed selection algorithm always preserves the best individual.

Performance Analysis

The GA proposed in the previous section is convergent, i.e. the fitness of the best individual converges to the global optimum in the solution space as the number of generations tends to infinity. This is a corollary of the properties of the three genetic operators: the selection operator preserves the best individual in the generation with probability 1.0, and the mutation and crossover operators guarantee that every point in the solution space is reachable starting from any randomly selected initial set of points (i.e. individuals in the initial random population). A formal proof of the GA convergence under the above limitations on the properties of the genetic operators can be found in [7]. Convergence is proved by means of homogeneous finite Markov chain analysis.

In order to evaluate the performance of the optimization algorithm, we need to define a performance measure and run a series of statistically significant experiments. The GA proposed in this paper is compared against the most popular classical GA (using Elitist selection) and against the two simplest randomized optimization algorithms, the *Random Search* and the *Random Walk*. In Elitist selection, population size is fixed during the whole optimization process. If the fixed population size is N, Elitist selection simply sorts the individuals in the population in decreasing order of their fitness and preserves the first N of them. The Random Search algorithm generates a finite sequence of random solutions and preserves the solution with the highest fitness found. The Random Walk algorithm starts from a random point in the solution space and on each iteration makes a “move” from the current point to a “neighboring” one (i.e. one that can be reached by applying some mutation operator on the current solution), if this move improves the fitness value.

[1] The exact calculation of the mathematical expectation of the sum of N random variables is significantly more complex and great precision is not needed here.

All experiments were executed using a randomly generated data dictionary and randomly generated queries against it. One and the same simplified cost model and fitness function are used in all experiments. Mutation activity was set to 0.1 and the neighborhood size parameter *k* was set to 6.

A good measure for the performance of the optimization algorithms is the evolution of the maximum fitness as a function of the number of solutions processed. Most performance studies evaluate the evolution of the maximum fitness as a function of the number of generations (for GAs) or iterations (for randomized algorithms) but this metric is inappropriate in our experiments as one of the GAs is characterized by dynamic population size. Experiments with different initial population sizes for the GAs were also executed.

First, we consider the performance of the two GAs and the two randomized algorithms optimizing a query with 7 joined relations. The reason for choosing this particular query size is that it is considered the upper limit for "classical" (or "small") queries and it roughly marks the boundaries of applicability of deterministic query optimization algorithms (for larger queries, deterministic search is not applicable). In fact, this is the largest query whose solution space we were able to fully traverse for a reasonable computation time and so we have the exact values of the global minimum and global maximum of the fitness function.

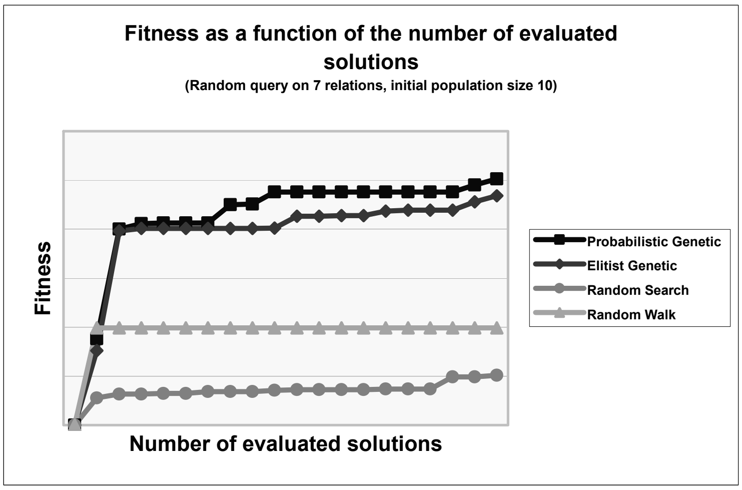


Figure 1. Performance comparison of genetic and randomized optimization algorithms. Random query with 7 joined relations, initial population size 10.

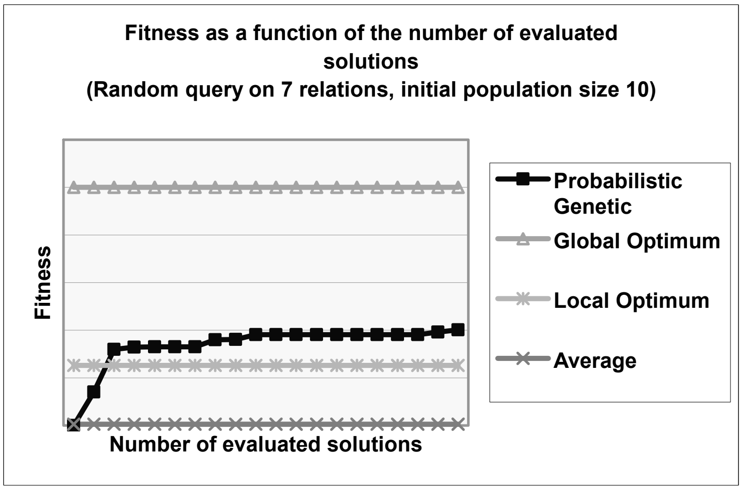


Figure 2. Performance of Probabilistic selection GA. Values compared to the average solution space fitness, a good local optimum and the global optimum.

The two GAs obviously outperform the two randomized algorithms. The random search strategies were easily trapped into local optima and failed to produce solutions that could compete with the best ones found by the GAs. The two GAs show overall similar performance, but the Probabilistic Selection seems a bit better when the number of evaluated solutions gets larger. One pronounced advantage of the Probabilistic Selection is the slower convergence compared to the Elitist selection, which allowed the Probabilistic GA to escape some suboptimal plateaus into which the Elitist GA was trapped. One probable reason is that the populations under Elitist GA are characterized by poorer diversity (very close values of the minimum, average and maximum fitness within the population).

Both GAs performed very well in terms of quality of the solutions found. Figure 2 shows the evolution of the maximum fitness in Probabilistic GA within the boundaries of the average fitness in the solution space and the global optimum. One of the best local optima (into which many of the runs were trapped) is also shown.

Further we test the performance of the four algorithms on a much harder optimization problem – a query with 100 joined relations.

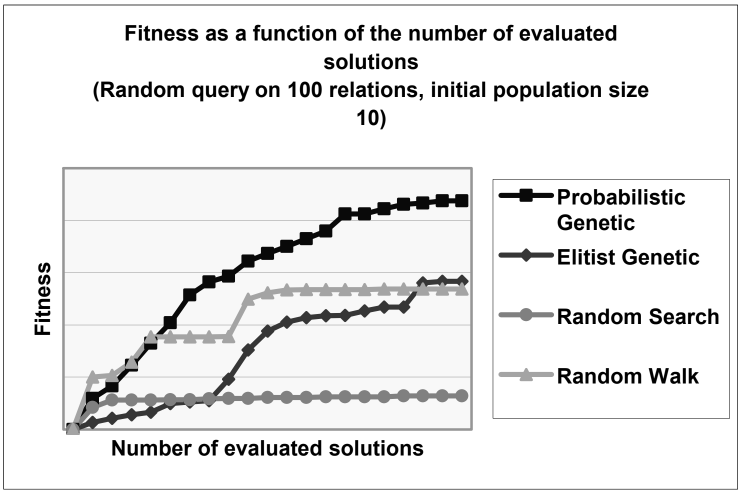


Figure 3. Performance comparison of genetic and randomized optimization algorithms. Random query with 100 joined relations, initial population size 10.

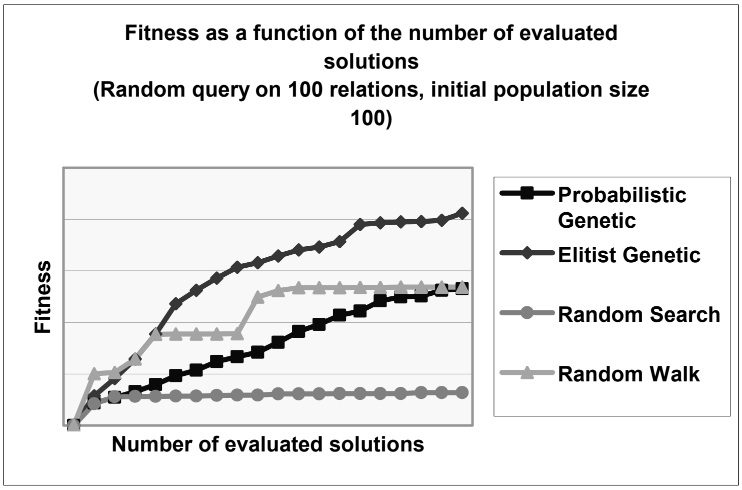


Figure 4. Performance comparison of genetic and randomized optimization algorithms. Random query with 100 joined relations, initial population size 100.

Here the superiority of the two GAs over Random Search is even more pronounced. The Random Walk however performed surprisingly well, with results often comparable to those of the GAs. The relative performance of the

Probabilistic GA and the Elitist GA varied considerably depending on the initial population size, as we can see from the comparison of Figure 3 and Figure 4.

Similar results were observed in experiments with smaller queries also. No rule linking the relative performance of the two GAs and the initial population size could be deduced however – for example, optimizing a query with 10 joined relations, Probabilistic GA was better on initial population size 10, worse on initial population size 20, again better on size 50, etc.

Conclusions and Future Work

In this work an adaptive genetic optimization algorithm is proposed for the join ordering problem. It outperforms the canonical genetic algorithm with classical fixed-size population selection operator such as the Elitist selection in a number of query optimization experiments. Considering the facts that

- Probabilistic Selection is expected to be faster than the Elitist – computational complexity $O(N)$ versus $O(N.\log(N))$
- the performance of the two selection operators are comparable and in many cases Probabilistic selection is better than Elitist
- Probabilistic Selection maintains a population with better diversity and more easily escapes suboptimal plateaus in the solution space,

the proposed optimization algorithm is a viable contender.

Overall, the results unconditionally prove the applicability of genetic optimization algorithms to the join ordering problem. GAs prove to be a competitive alternative to deterministic optimization algorithms even for small solution spaces.

One direction for future work would include experimenting with adaptive mutation and crossover operators. It is reasonable to expect further performance improvements, as it is suggested by a number of recent researches. The proposed algorithm can also be used as a basis for a hybrid optimization algorithm incorporating certain domain-specific heuristics and randomized local search techniques.

Bibliography

[1] A. Eiben, E. Marchiori, V. Valkó. Evolutionary Algorithms with on-the-fly Population Size Adjustment. In: Proc. of the 8th International Conference on Parallel Problem Solving From Nature, 2004.

[2] T. Haynes. A comparison of random search versus genetic programming as engines for collective adaptation. In: Proc. of the ACM Symposium on Applied Computing, 1997.

[3] M. Stillger, M. Spiliopoulou. Genetic programming in database query optimization. In: Proc. of the 1st Annual Conference on Genetic Programming, 1996.

[4] M. Steinbrunn, G. Moerkotte, A. Kemper. Optimizing join orders. In: Report MIP 9307, Universität Passau, 1993.

[5] K. Bennett, M. Ferris. Y. Ioannidis. A genetic algorithm for database query optimization. In: Proc. of the 4th International Conference on Genetic Algorithms, 1991.

[6] Y. Ioannidis, Y. Kang. Randomized algorithms for optimizing large join queries. In: Proc. of the ACM, 1990.

[7] G. Rudolph. Convergence Analysis of Canonical Genetic Algorithms. In: IEEE Transactions on Neural Networks, 1994.

Author's Information

***Stoyan Vellev** – PhD student, Faculty of Mathematics and Informatics, Sofia University, 7 Raiko Alexiev Str, bl. 30, Sofia-1113, Bulgaria; e-mail: stoyan.vellev@sap.com*

ONTOLOGY-BASED CLASSIFICATION OF NEWS IN AN ELECTRONIC NEWSPAPER

Lena Tenenboim, Bracha Shapira, Peretz Shoval

***Abstract**: This paper deals with the classification of news items in ePaper, a prototype system of a future personalized newspaper service on a mobile reading device. The ePaper system aggregates news items from various news providers and delivers to each subscribed user (reader) a personalized electronic newspaper, utilizing content-based and collaborative filtering methods. The ePaper can also provide users "standard" (i.e., not personalized) editions of selected newspapers, as well as browsing capabilities in the repository of news items. This paper concentrates on the automatic classification of incoming news using hierarchical news ontology. Based on this classification on one hand, and on the users' profiles on the other hand, the personalization engine of the system is able to provide a personalized paper to each user onto her mobile reading device.*

***Keywords**: Classification, Electronic newspaper, Information filtering, Mobile device, Ontology, Personalization.*

***ACM Classification Keywords**: H5.2. Information interfaces and presentation, H.3.3. Information Storage and Retrieval; information filtering, H.3.1. Content Analysis and Indexing*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

1. Introduction

Electronic, online newspapers started appearing at about the same time that the Internet became public. An electronic newspaper has many forms. One form is **electronic edition** of the printed newspaper (namely, the publisher publishes its "standard" newspaper on a Website, using e.g. PDF files). The user can read the electronic edition similar to a paper edition; there is no personalization, neither with respect to content nor with respect to layout. Another form of electronic newspaper is **news website**, which enables the user browsing in menus that are organized in subject categories and sub-categories. Yet, another form of electronic newspaper can be seen more like a **search engine**, which enables the user to insert search terms (i.e., topics of interest) and get in response respective news items that are published on the Web by various news providers. In contrast to the previous forms of services, this one does not publish and does not edit news; rather, it searches and provides links to news published elsewhere by news providers/agencies. In addition to searching according to user-defined terms, such systems enable personalization: the user can define a profile by selecting topics of interest, and the system would search for items in the selected topics. Personalization features are supported in some sites via RSS technology, like in portals that deliver news and have built-in news aggregator capabilities. Personalization is sometimes done based on collaborative filtering methods (e.g. Google News). Interactivity is an additional feature, and exists in the form of reader feedback capabilities and in the form of capabilities for "pulling" personalized news and other information. The general picture is that of much diversity and heterogeneity between the various players in the online newspaper branch of the newspapers industry.

Common to most of the above forms of electronic newspapers is that the user is assumed to read the news from a computer screen, while connected via the Internet to a certain news provider or search engine. But these services might not be sufficient for many readers and reading situations. A newspaper reader may be willing to read articles from various favorite daily and weekly newspapers and magazines while being on the move; for example, while on a business trip, or on vacation, or waiting for a fried at a café, etc. A reader would prefer that someone or something would make interesting articles, from favorite sources; accessible to her all the time and anywhere, delivered directly onto a mobile reading device. A reader would like to subscribe to one unique, personalized newspaper that includes the interesting articles from favourite sources, arranged and presented in an order that best fits her interests and reading habits. An advanced electronic newspaper service should enable a reader, with just one click of a button, to check whether updates of her personal newspaper are available, or to choose receiving updates automatically as soon as they are published.

This paper deals with such a service. ***ePaper*** is a prototype of an electronic newspaper system aims at providing personalized newspaper on mobile reading devices. The ePaper is projected to provide a "look and feel" of a newspaper that is run on a medium-format mobile device, providing up-to-date news aggregated from many news providers, and personalized according to each user's preferences. The *ePaper* system is a client-server application: On the server side, it aggregates news coming continuously from many news providers; classifies each news item to subject concepts, based on a news ontology; and then determines the relevancy of the news items to each of the subscribed users (readers), based on both content-based and collaborative filters, and ranks the news items that will be delivered to each user, thus providing personalized newspapers. On the client side, the user, who gets the news on her reading device, enjoys an intuitive interface, enabling easy navigation and browsing, and advanced content adaptation capabilities, including switching and configuring layouts. This paper concentrates on only one part of the system - the classification of news items that are obtained by the system, so that later on the personalization algorithms can determine the relevancy of each item to each user.

The rest of this paper is structured as follows: Section 2 presents the general architecture of the ePaper system; Section 3 describes the content managed layer, which is, among else, in charge of the classification of news. Section 4 surveys some related work on classification methods, and Section 5 describes the classifier implemented in the ePaper. Section 6 summarizes and discusses further research.

2. General Architecture of ePaper System

The ePaper is a research project aimed at developing a prototype system that can be viewed as a central newspapers or magazines provider. On one hand, it obtains news obtained from various providers; on the other hand, it distributes personalized newspapers to subscribed users on specialized mobile reading devices. In this section, we describe briefly the general architecture of the system. Figure 1 presents an overview of the ePaper architecture.

The system is implemented based on client-server architecture. **The server system** consists of five layers: *Aggregator, Content Manager, Personalization, Content Delivery Services,* and *System Management Tools.*

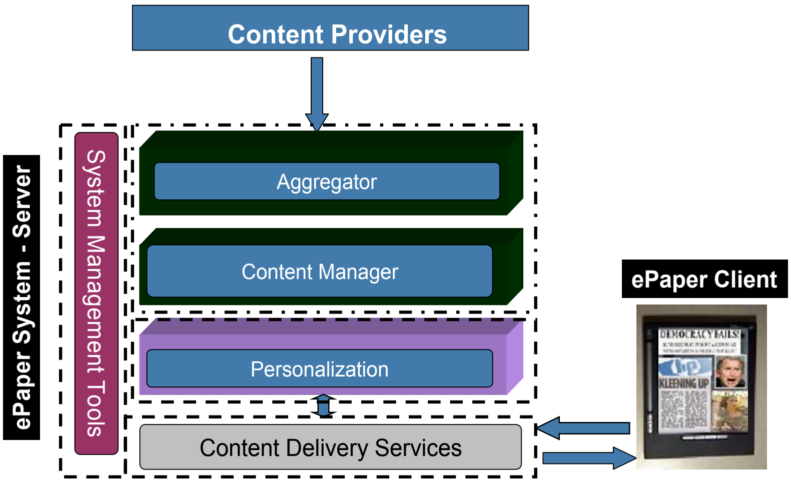


Figure 1. General architecture of the ePaper

The ***Aggregator*** interacts with content (news) providers and imports news item to the ePaper repository. (It may be assumed that the management of the ePaper service has business arrangements with certain providers. The business models with the providers, as well as with the clients/users, are beyond the scope of this project, and are immaterial for the description of the system.) A news item obtained from a provider may consist of one or more text and image files in certain formats, and include various metadata. For example, Reuters uses the NewsML format and its specialized metadata structures. (NewsML - News Markup Language - is a standard for the exchange of news supported by the IPTC (www.iptc.org). The main responsibility of the aggregator is to check the content providers for new news items, download them, create an index for each, and store the new aggregated items in the ePaper's file system.

The ***Content Manager*** processes the content of each news item received from the aggregator, and prepares it for personalization and delivery to relevant users. Its main responsibility is classification of the items: A text classification algorithm is used to analyze the content of each item and determine the concepts that best represent it. For this, the system maintains a hierarchical news ontology, which is based on the IPTC Subject Codes taxonomy. (More details on the Content Manager and the classification process in subsequent sections.)

The ***Personalization*** layer consists of a novel personalization engine that determines the level of relevancy of each news item to each user. The personalization engine applies content-based and collaborative filtering algorithms. The content-based filtering algorithm computes the similarity of the ontology concepts that represent each item, to each user's profile. A user's profile too consists of ontology concepts, which are initially defined by the user (upon registration to the system), and later on are dynamically updated by the system based on implicit user feedback. The measure of similarity between an item's and a user's profile considers the ontological proximity (or distance) between concepts in the two profiles. The collaborative filtering algorithm determines the similar users of each user (based on how many common items they have read), and computes a time-factor which considers how long ago (in hours) each item was read by each user. The final relevancy score of an item is computed as a weighted average of the two filters' rankings, taking into consideration the "maturity" of each item in the system: the more readers an item had, the more weight is given to its collaborative score. The result of the personalization process is a ranked list of items that will be delivered to each user, classified within the main ontology concepts that are of interest to her. The user can ask the system to refresh the items list anytime; as result, more news items may be delivered, and the ranking of all items on the reading device may be updated according to the fresh personalization process. The user can overrule the personalization engine, either by asking to get a "standard edition" of a certain newspaper, or by browsing the news items that exist in the repository, using menus of ontology concepts.

The ***Content Delivery Services*** layer orchestrates the processes of the system. It interacts with the *Personalization* layer, submits requests for personalized news from users, and sends the ranked news items it receives to the user. It also receives feedback from the user (tracking user's behavior data) and sends this data to the *Personalization* layer, which updates the user's profile to reflect the recent user's reading preferences.

The ***System Management Tools*** layer provides standard system tools such as logging and reporting, as well as special tools for the ePaper application. This includes, among else: a) Ontology Editor: this tool enables maintenance of the ontology, i.e. adding new concepts to the ontology, as new concepts may evolve over time. b) Registration subsystem: this is a web-based system where each new user registers and subscribes to preferred ePaper services, including: 1) some demographic and billing information (which are of no interest here); 2) selection of favorite content providers (i.e. newspaper) from whom to receive news; 3) an option to select a "standard" edition of a newspaper (instead of a personalized one); 4) an option not to track the user's reading, which my be desirable by certain users, and by that avoiding implicitly update of the user's profile; 5) definition of an initial content-based profile, by selecting concepts from the hierarchical ontology and determination of their initial weights of importance. As said, the initial profile will be updated dynamically according to implicit feedback from the user's reading device.

The **Client system** interacts with the *Content Delivery Services* layer for receiving data. The data sent to a user includes the user's profile information and a ranked-list of news items as determined by the *Personalization* layer. The Client is in charge of rendering the content and adapting it to preferred layout, and presenting the content to the user. To manage the variety and constraints of different mobile devices, the system supports dynamic content adaptation mechanisms based on the user's device, the user's preferences and local customizations made by each user. Thus, the presentation of content functionality is loosely coupled with the content preparation process, a capability that may scale the number and variety of devices supporting this service easily.

3. Content Management

Figure 2 presents the *Content Management* layer. It receives news items from the *Aggregator*, parses each item to extract content and metadata, classifies its content, and stores it in a repository of "active" items, to be delivered to users. (Items that are not read for some while by users are archived and become inaccessible from the mobile device). The *Content Management* layer consists of several units; some of them are described below.

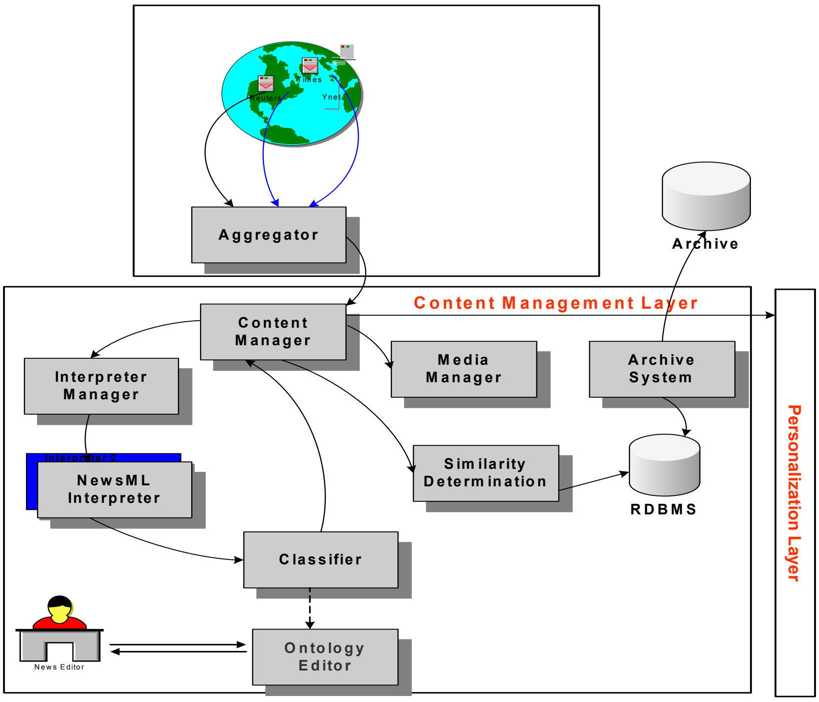


Figure 2. Content Management layer

3.1. The Content Manager

The Content Manager is orchestrating the processes of the *Content Management* layer. It receives news items from the *Aggregator* and sends them to the Interpreter Manager. Then it receives classified data back from the Classifier and sends them to the other functional units of the *Content Management* layer. After an item passes all the functional units, it is stored in the repository of "active" items, ready to be used by the *Personalization* layer.

3.2. The Interpreter Manager

The ePaper system is able to handle news items coming from multiple news providers, in multiple languages and in multiple formats. The Interpreter Manager is responsible for identifying the item's format and activating an appropriate interpreter, i.e., the interpreter that is able to "understand" the item's format and language, and extract from it the relevant metadata. The metadata it extracts from the item includes e.g., the item's provider/source, language and date of creation. Then it passes the item along with these metadata to the Classifier.

Currently, two interpreters have been implemented in the ePaper prototype system: one for NewsML format, which is used by many news providers, e.g. Reuters. The other interpreter is for RSS format. The ePaper can easily be extended to handle other standard formats by developing dedicated interpreters to each standard.

3.3. The Classifier

The ePaper system uses a news-ontology as a common language for content-based filtering. The ontology concepts are used to represent the news items' profiles and the users' profiles; the content-based filter will measure the similarity between the two profiles to determine the level of relevancy of each item to each user.

The ontology of ePaper is based on IPTC NewsCodes. NewsCodes is a set of controlled vocabularies; among other, it includes a **Subject ontology**, which consists of about 1400 concepts, organized in a 3-level hierarchy (termed Subject, SubjectMatter and SubjectDetail). News-providers who use NewsML use these concepts to describe the content of their articles, including them as part of the metadata attached to each news items.

For example, some of the first (Subject) level concepts of the IPTC Subject codes taxonomy are Sport, Politics, Economy, Education, Health and Science and Technology. Some of the second (SubjectMatter) level concepts of Politics are Election, Diplomacy, Defense, Government, and Parties. Some of the third level (SubjectDetail) concepts of Diplomacy are Summit, International Relations and Peace Negotiations.

The Classifier is responsible for determining the ontology concepts that will represent each news item and their weights, i.e., to define its content-based profile. A news item may deal about more than one concept; hence, we are dealing with a **multi-label** classification problem, where a news item may be classified into many concepts and in different levels of the ontology hierarchy. For example, a news item about an attempt of assassination at a presidential elections campaign can be classified to a number of concepts, such as Elections (which is a sub-concept of Politics) and Crime.

The classification process will be described in more detail in Section 5; before that, in the next section we provide a brief overview and related work on classification

4. Related Work on Classification

First, we provide a general introduction on text classification; then we present related work on news classification.

4.1 Text Classification

Text Classification or Categorization (TC) is the task of automatically assigning a text document (in our case, a news) to one or more predefined categories (in our case, ontology concepts) based on its contents. Nowadays, the dominant approach in TC is Machine Learning [Sebastiani, 2002]. According to this approach, a general inductive process automatically builds a text classifier by learning, i.e., by observing the characteristics of a set of previously classified documents - a training set. These characteristics are then used to classify new documents.

Different types of TC tasks can be distinguished: From a category assignment point of view, we distinguish between single-label and multi-label classification. In **Single-label** (also called multi-class) TC, exactly one category must be assigned to a document. In **Multi-label** TC, any number of categories may be assigned to a document. **Binary** categorization is a special case of single-label categorization, in which there is only one category and each document can be assigned to it or not (“yes”, “no”).

TC tasks can also be differentiated by the structure of the predefined categories set. In **Flat** categorization, the predefined categories are treated in isolation and there is no structure defining the relationships among them. Most of the studies in TC have focused on flat classification, and after many years of research flat classification has become a well-established research area and many good classifiers have been developed [Sun and Lim, 2001]. In **Hierarchical** categorization, the predefined categories are organized in a hierarchical structure that reflects relations between them. Most hierarchies are organized in tree-like structures, i.e., there are parent-child relationships between categories. In hierarchical classification, we can distinguish between cases where all documents belonging to a child category also belong to the parent - called *strong subsumption*; and cases where a child category has documents that do not belong to its parent category - called *weak subsumption*.

4.2 Multi-Label Classification

Many classification methods, such as Naïve Bayes, SVM, and Logistic Regression, are of the single-label type. Research on multi-label classification has received much less attention. Some methods that are mainly used for multi-label classification are presented below [based on Tsoumakas and Katakis, 2007].

The most popular approach for multi-label classification is **binary approach** (also called one-against-the-rest). A separate classifier is learned for each category C_i. The original data set is transformed into |C| data sets. The data set for each category C_i contains all examples of the original data set, labeled as c if the labels of the original example contained c, and as ¬c otherwise. For the classification of a new instance x, this method outputs as a set of labels the union of the labels predicted by the |C| classifiers.

This method has two main problems. First, it assumes independence of categories, which is not always true; there may be strong dependence between categories, in particular in hierarchical classification. Also, relations between categories on the same level can exist. For example, the following categories have some dependency: ‘Politics’ and ‘Unrest, Conflicts and War’; ‘Environmental Issue’ and ‘Health’. In such cases, association of an item with one category may influence its probability to be associated with a related category. But the binary approach cannot model such relations. The second problem is that a big number of binary classifiers have to be learned, which may cause memory problems, and take a lot of time because each new instance should be processed by all |C| classifiers.

Another, less popular approach, in multi-label classification, is to **consider each different set of labels** that exist in the multi-label data set as a single label. It so learns one single-label classifier for C` categories, where C` is the power set of initial C categories. One of the negative aspects of this method is that it may lead to data sets with a large number of classes and few examples per class.

Another, yet not well-documented method is to learn one **multi-class classifier**, which can output a distribution of certainty degrees (or probabilities) for all labels in C, and then post-processes this distribution to output a set of labels. One simple way is to output labels for which the certainty degree is greater than a specific threshold (e.g. 0.5). A more complex way is to output labels for which the certainty degree is greater than a percentage (e.g. 70%) of the highest certainty degree. This method too ignores possible dependencies among categories.

4.3 Hierarchical Multi-Label Classification

Hierarchical classification has two main advantages compared to flat classification. First, it enables easy location of required categories when there are a significantly large number of categories; it is much easily to search among some high-level categories and then among some related sub-level categories, than to perform a general search among all existing categories. Second, it reflects the intuition of relatedness of topics that are close to each other in the hierarchy.

Two approaches were adopted by existing hierarchical classification methods: **big-bang**, and **top-down level-based** approach. In the big-bang approach, a document is classified (or rejected) into (or from) a category in the category tree by a classifier in one single step. In the top-down level-based approach, one or more classifiers are constructed at each level of the category tree, and each classifier works as a flat classifier at that level. A document will first be classified by the classifier at the root level into one or more lower level categories. It will then be further classified by the classifier/s at the lower level category/ies, until it reaches one or more final categories, which can be leaf categories or internal categories. (A classifier can stop at an internal node if an item cannot be classified to any of its child categories.)

One of the important works on hierarchical TC is [Koller and Sahami, 1997], who divide the hierarchical classification task into a set of smaller classification tasks, each of which corresponds to some split in the classification hierarchy. They show that this approach enables obtaining significantly higher accuracy compared to a single massive classifier.

4.4 Classification in the News Domain

In this section, some implementations of classification algorithms in news domain are presented.

An interesting implementation of multi-label classification is presented in [Antonellis et al, 2005]. They implemented a news categorization system that decomposes each document into its sentences, computes term to sentences matrix and performs multi-label classification by estimating the similarity of each sentence to the category vectors. Category vectors are computed by combining the columns of the corresponding term to sentences matrix of the training set. If the estimated similarity is above a threshold defined during the training phase, the document is classified to the corresponding category. Thus, multi-label classification is allowed.

[Calvo et al., 2004] applied automatic Naïve Bayes classifier on news stories from Reuters RCV1 Corpus. They performed flat multi-label classification using two different thresholding strategies: score-based, where all categories with a score above some threshold are assigned to a document; and rank-based, where the k top-ranked categories are assigned to a document. The results were compared to a kNN classifier. It was shown that for the rank-based thresholding strategy, the best performance was achieved for k = 1. The explanation is that in this data set, the average of categories that each document is assigned to is approximately 1. It was also noted that the best performance of the rank-based strategy is considerably worse than the score-based thresholding strategy. Regarding the comparison to kNN classifier, it was concluded that Naive Bayes classifiers produce lower quality classification but seem to be better suited for applications were classification needs to be performed at real time; kNN instead produces better classification but places to much load on classification time.

One of the examples for implementation of hierarchical classification in the news domain is presented in [Eilert et al., 2001]. The developed system called Bikini was supposed to classify news from several WWW news sources into concepts hierarchy (ontology). Bikini's ontology was handcrafted and included approximately 130 categories, with a maximum three-depth of four. For classification purpose, it was interpreted as a flat list of concepts. Any node in the tree was interpreted as a leaf tree. For intermediate nodes this was achieved by introducing an artificial ‘miscellaneous’ successor node. The classification was performed by comparing document representation vector with categories representation vectors using simple similarity measure.

5. The ePaper Classifier of News

In ePaper, we implemented hierarchical multi-label classification algorithm utilizing a flat multi-class classifier provided by LingPipe open source software [LingPipe]. The utilized LingPipe's LanguageModel (LM) Classifier is based on statistical language modeling techniques and performs probability-based classification into non-overlapping categories.

The motivation for language modeling has traditionally come from speech recognition; recently it became widely used in many other application areas. Language modeling aims at predicting the probability of naturally occurring word sequences, s = w1w2…wn. It puts high probability on word sequences that actually occur, and low probability on word sequences that never occur. The simplest and most successful approach to language modeling is based on the n-gram model [Peng, 2003]. An n-gram is a sub-sequence of n items from a given sequence. The items can be characters or words. If the language model is based on character sequences, it is called *character* language-model. According to language modeling approach, the probability of any word or character sequence is calculated as frequency of the observed patterns

$$\Pr(w_i|w_{i-n+1}\dots w_{i-1}) = \frac{\#(w_{i-n+1}\dots w_i)}{\#(w_{i-n+1}\dots w_{i-1})}$$

where #( ) denotes the number of occurrences of a specified gram in the training corpus.

The LingPipe's LM classifier constructs a character language-model for each category during the training phase; then, at classification time, it calculates conditional and joint probabilities of each category for the classified object. Also, a score, which is the character cross-entropy rate normalization allowing between-document comparisons, is provided. This score is ordered in the same way as the joint probabilities. Finally, LingPipe classifier returns one best category as result of classification process.

It was found in many researches that the language modeling approach provides competitive and often superior results compared to more sophisticated learning techniques [Peng, 2003]. Moreover, it provides scalable training and classification time performance. Therefore, we utilized LingPipes LM Classifier, allowing classification into one of non-overlapping categories, and enhanced it to satisfy our hierarchical and multi-label requirements.

For **multi-label classification**, we apply an approach based on estimations of a posteriori probabilities of an item to belong to some category. According to this approach, we can think of a classification as "better estimated" if the probability of the destination category is above some threshold. For example, if the probability of an item **D** to belong to category **C** is 0.9 it is better estimated than if it is only 0.4. To determine the threshold for multi-label classification we use the cross-entropy scores provided by LingPipe classifier, as they are better suited for cross-document comparison. Using this principle, we classify item D in n categories Ci, such that

$$\text{score}(Ci,D) > \tau\ ;\ 1 \leq i \leq n \text{ and } 1 \leq n \leq |C|$$

where the threshold τ is learned on empirical runs. Then, the set of categories assigned to item D is {C1,…,Cn} a set of "n best" categories with classification score above the defined threshold.

We chose this approach as it has shown favorable results compared to the more standard binary approach to multi-label classification [Vilar et al, 2004]. Moreover, it is very scalable in terms of memory and time performance: it does not require additional classification models to be loaded into the memory and an additional classification to be performed at run time. Also, it natively ranks the categories assigned to an item according to their level of relevance; this rank is later used by the personalization and content delivery services of ePaper.

For **hierarchical classification**, we apply the top-down level-based approach. We implemented this approach by constructing a separate classification models for each non-leave concepts of the ontology tree. At the first level, there is one model for classification into one or more of 17 Subject concepts. At the next level, there are 17 models for classification into the SubjectMatter concepts; and about 130 models for classification at the third (SubjectDetail) level of the ontology. Hence, the number of models generated is identical to the number of non-leave concepts in the hierarchy plus one for the root node.

After all the above multi-label and hierarchical extensions, the classification of a new item is performed as follows: First, the item is classified into one or more top level (Subject) concepts. Then, if one or more of the Subjects are assigned to the item, it is further classified into one or more child concepts (SubjectMatter) of the Subjects. Then, if one or more of SubjectMatter concepts are assigned to the item, it is further classified into their child concepts (SubjectDetail). The classification process stops when classification to the detailed concept is not confident enough. (Initial confidence thresholds are defined by configuration parameters, depending on empirical runs.)

Once the results of the classifications at each level are obtained, the final classification is determined according to the received concepts' weights and defined confidence thresholds. The most specific concepts having classification scores above the pre-defined threshold are assigned to the item.

The following example will clarify the classification process. Suppose a certain news item has to be classified. First, it is assigned by the root classifier to one or more of the top-level concepts; assume it is assigned to Economy and Science & Technology. Then, the second level classification starts for each of these concepts. The Economy and Science & Technology classifiers are invoked one after the other, attempting to classify the item to one or more of their sub-concepts. Assume that Macro Economics and Energy & Resources concepts are returned by the Economy classifier, and Applied Science concept is returned by the Science & Technology classifier. When Macro Economics and Energy & Resources concepts are received, the third level classification starts for each of them. The Macro Economics and Energy & Resources classifiers are applied on the item one after the other, attempting to classify it to one or more of their sub-categories. The same is for the Applied Science concept: its classifier is applied on the item attempting to classify it to one or more of its sub-categories. Assume that the results of the third level classification are as follows: Macro Economics classifier returns Government Aid concept; but Energy & Resources and Applied Science classifiers cannot classify the item to any of their sub-categories (because the results are below the defined thresholds).

The weight of each assigned concept is calculated based on its cross-entropy score received at the corresponding classification level. Assume the scores of the assigned concepts are as follows: Economy (-1.65), Science & Technology (-1.78), Macro Economics (-1.73), Energy & Resource (-1.75), Applied Science (-1.84), Government Aid (-1.80); and the defined threshold is (-1.86). The item is classified to the most specific concepts according to their location in the hierarchy, i.e., Government Aid, Energy & Resources and Applied Science. We then convert each concept's score onto a weights scale of 0--1, where a concept's weight expresses its absolute importance in the item. This is done as follows: the "best" concept gets the weight 1; the weights of the other concepts are lower, proportionally. In our example, the following concept–weight pairs are assigned to the item: Energy & Resource, 1 (= -1.75), Government Aid, 0.97 (= -1.75/-1.80), and Applied Science, 0.95 (= -1.75/-1.84).

6. Summary and Future Issues

The ePaper prototype system is now undergoing various evaluations. This includes evaluations of the content personalization algorithms and the content adaptation to the reading device and users' preferences. Regarding content personalization, we examine the effect of various parameters of the content-based and the collaborative filtering algorithms, e.g. the optimal scores of the various types of ontological similarity between user and item profiles; the optimal number of concepts to be considered in a user's profile; and the optimal method to update a user's profile based on implicit feedback. For this, we run controlled experiments with users who evaluate the relevancy of news items provided to them, and compare their evaluations to the system's ranking of those items,.

Regarding the classification process, additional work is required for tuning threshold parameters. For this, a combination of multi-label and hierarchical measures will be defined, that will take into consideration specific system requirements. We plan to compare the results of the adopted 'thresholding' strategy to the common binary approach to multi-label classification. Since the construction and classification using about 1400 binary classifiers seems unacceptable for the ePaper system (because of time required for each classification), perhaps the binary approach will be applied only on the top level of the hierarchy (i.e. for the 17 Subject concepts). A method for multi-label classification taking into in consideration dependencies among concepts is also under development.

Acknowledgments

The ePaper project was sponsored by Deutsche Telekom Co., and developed at the Deutsche Telekom Laboratories at Ben-Gurion University.

Bibliography

[Antonellis et al., 2005] Antonellis, I., Bouras, C. and Poulopoulos, V. (2005). Personalized news categorization through scalable text classification. 8th Asia Pacific Web Conference (APWEB '06).

[Eilert et al., 2001] Eilert, S., Mentrup, A., Mьller, M.E., Rolf, R., Rollinger, C.R., Sievertsen, F. and Trenkamp, F. (2001). Bikini: user adaptive news classification in the World Wide Web. Workshop on Machine Learning for User Modeling; 8th Intl. Conf. on User Modeling.

[Calvo et al., 2004] Calvo, R.A., Lee, J.M. and Li, X. (2004). Managing content with automatic document classification. J. Digit. Inf. 5 (2).

[Koller and Sahami, 1997] Koller, D. and Sahami, M. (1997). Hierarchically classifying documents using very few words. Proceedings of the 14th International Conference on Machine Learning, pp.170-178.

[LingPipe] LingPipe – a suite of Java libraries for the linguistic analysis of human language. http://www.alias-i.com/lingpipe/index.html

[Eilert et al., 2001] Peng, F., Schuurmans, D. and Wang. S. (2003). Language and task independent text categorization with simple language models. Proceedings of HLT-NAACL 2003.

[Sebastiani, 2002] Sebastiani F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, Vol. 34 (1), pp.1-47.

[Sun and Lim, 2001] Sun A. and Lim E.P. (2001). Hierarchical text classification and evaluation. First IEEE International Conference on Data Mining, ICDM'01, pp. 521-528.

[Tsoumakas and Katakis, 2007] Tsoumakas, G. and Katakis, I. (2007). Multi-label classification: an overview. International Journal of Data Warehousing and Mining, Vol. 3 (3), pp. 1-13.

[Vilar et al., 2004] Vilar, D., Castro, M.J. and Sanchis, E. (2004). Multi-label text classification using multinomial models. Proceedings of 4th International Conference on Advances in Natural Language Processing (EsTAL 2004), pp. 220-230.

Authors' Information

Lena Tenenboim *(Dept. of Information Systems Engineering - ISE, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: lenat@bgu.ac.il) is Ph.D. student. She holds a M.Sc. in ISE from Ben-Gurion University. Her research interests include Information Retrieval and Filtering, Machine Learning and Data Mining.*

Bracha Shapira *(Dept. of ISE, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: bshapira@bgu.ac.il) is Senior Lecturer of ISE. She holds a M.Sc. in Computer Science from the Hebrew University in Jerusalem and a Ph.D. in Information Systems from Ben-Gurion University. Her research interests include Information Retrieval and Filtering, specializing in various aspects of user profiling and personalization. In addition, she has worked on privacy preservation while browsing and on formal models of Information Retrieval systems. She is leading research projects in these domains at the Deutsche-Telekom Research Lab at Ben-Gurion University.*

Peretz Shoval *(Dept. of ISE, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: shoval@bgu.ac.il) is a Professor of ISE. He earned his Ph.D. in Information Systems from the University of Pittsburgh, where he specialized in expert systems for information retrieval. In 1984, he joined Ben-Gurion University, where he founded the Information Systems Program, and later founded and headed the Dept. of ISE. His research interests include information systems analysis and design methods, data modeling, and information retrieval and filtering.*

DERIVATION OF CONTEXT-FREE STOCHASTIC L-GRAMMAR RULES FOR PROMOTER SEQUENCE MODELING USING SUPPORT VECTOR MACHINE

Robertas Damaševičius

Abstract: *Formal grammars can used for describing complex repeatable structures such as DNA sequences. In this paper, we describe the structural composition of DNA sequences using a context-free stochastic L-grammar. L-grammars are a special class of parallel grammars that can model the growth of living organisms, e.g. plant development, and model the morphology of a variety of organisms. We believe that parallel grammars also can be used for modeling genetic mechanisms and sequences such as promoters. Promoters are short regulatory DNA sequences located upstream of a gene. Detection of promoters in DNA sequences is important for successful gene prediction. Promoters can be recognized by certain patterns that are conserved within a species, but there are many exceptions which makes the promoter recognition a complex problem. We replace the problem of promoter recognition by induction of context-free stochastic L-grammar rules, which are later used for the structural analysis of promoter sequences. L-grammar rules are derived automatically from the drosophila and vertebrate promoter datasets using a genetic programming technique and their fitness is evaluated using a Support Vector Machine (SVM) classifier. The artificial promoter sequences generated using the derived L-grammar rules are analyzed and compared with natural promoter sequences.*

Keywords: *stochastic context-free L-grammar, DNA modeling, machine learning, data mining, bioinformatics.*

ACM Classification Keywords: *F.4.2 Grammars and Other Rewriting Systems; I.2.6 Knowledge acquisition; I.5 Pattern recognition; J.3 Life and medical sciences.*

Conference: *The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Promoters are short regulatory DNA sequences that precede the beginnings of genes. They are common both in prokaryotic and eukaryotic genomes. It is important to distinguish between promoter and non-promoter sequences, because this distinction allows identifying starting locations of genes in uncharacterized DNA sequences. Though there are conserved patterns in the promoter sequences of a species, there are many exceptions which make the promoter recognition a complex problem. Although many complex machine learning methods have been proposed so far for promoter recognition, the problem is still open [Sobha Rani et al., 2007].

The crucial obstacle in finding mammalian promoters is that they usually do not share extensive sequence similarity even when they are functionally correlated, which prevents detection by sequence similarity-based search methods such as BLAST or FASTA. Mammalian promoters can be seen as small structures of coding regions with few functional elements (exons) interspersed in a larger sequence with no known biological function (introns) [Werner, 2003].

Modern promoter recognition tools use machine learning techniques such as Naive Bayes, Decision Trees, Hidden Markov Models, Neural Networks or Support Vector Machine (SVM) [Monteiro, 2005]. Best machine-learning based promoter recognition methods allow achieving up to 98% accuracy [Ranawana and Palade, 2005], however they do not provide any insight on the internal structure of the promoter sequences. Analysis of some promoters suggests that promoter sequences may have a modular structure. Identification of promoter structure may enhance our understanding of gene regulation mechanisms and genome evolution process.

Formal grammars can provide a means of describing complex repeatable structures such as DNA. The structure of promoter sequences can be described using a formal context-free grammar such as *L-grammar* (or *L-system*) [Lindenmayer, 1968], and the problem of promoter recognition can be replaced by grammar induction [Unold, 2007]. Several authors consider grammar induction in the context of bioinformatics as well as L-systems. The possibility of discovering the rewrite rule for L-systems and the state transition rules for cellular automata using genetic programming techniques is discussed in [Koza, 1993]. So-called semi-Lindenmayer systems are considered for the study of protein formation in [Marcus, 1974]. An algorithm to construct a short context-free grammar (also called program or description) that generates a given sequence is proposed in [Jimenéz-Montano, 1984]. The inference of regular language grammar rules based on n-grams and minimization of the Kullback-Leibler divergence is described in [Infante-Lopez and de Rijke, 2004]. O'Neill et al. [2004] generate regular expressions for promoter recognition problem using Grammatical Swarm technique. Each individual swarm particle represents choices of program construction rules, where these rules are specified using a Backus-Naur Form (BNF) grammar. Denise et al. [2003] generate genomic sequences and structures according to a given probability distribution and syntactical (grammatical) parameters. The method is applied for generating basic RNA secondary structures. Stochastic context-free grammars constructed from sample sets of sequences are also considered for modeling RNA sequences [Grate et al., 1994; Sakakibara et al., 1994]. The usefulness of Chomsky-like grammar representations for learning members of biological sequence families is analyzed in [Muggleton et al., 2000].

The aim of this paper is 1) to describe automatic derivation of stochastic L-grammar rules from promoter datasets using Support Vector Machine (SVM), and 2) to apply the derived L-grammar rules for structural analysis of the promoter sequences.

Modeling of DNA Sequences Using L-Grammar

If we treat the genome as a language, we can generalize the structural information contained in biological sequences and investigate it using formal language theory methods [Fernau, 2003]. Some aspects of formal languages are similar to biological processes in general and genetic mechanisms in particular, e.g.:

- *Pure grammars* do not differentiate between *terminal* and *non-terminal* symbols, so that all words generated from the grammar rules are put into the generated language. This notion is well-motivated biologically, because all symbols in a DNA sequence have similar role.
- *Erasing rule* $A \to \varepsilon$ models the death of a cell (which has been in a state A before its disappearance) or a delete mutation in a DNA sequence.
- *Chain rule* $A \to B$ reflects a change of a state of the corresponding cell or a single nucleotide mutation in DNA.
- *Repetition rule* $A \to AA$ models highly repetitive DNA sequences such as tandem repeats.
- *Growing rule* $A \to BC$ models the split of a cell being in a state A, into two children being in state B and C, or growth of a DNA sequence.
- *Stochastic rule* $p : A \to B$ models the random mutations of DNA sequences.

From the biological point of view, the components of any biological organism evolve simultaneously, so we cannot expect that biological processes could be modeled using a sequential approach. It is more likely that the cells that reproduce simultaneously would be modeled by a mechanism that is based on the same behavioral principles. This makes a special class of grammars, called *parallel grammars* [Fernau, 2003], particularly interesting for research of biological sequences such as DNA. L-systems are a special class of parallel grammars that can model the growth of living organisms, e.g. plant development, but also able to model the morphology of a variety of organisms [Prusinkiewicz and Lindenmayer, 1990]. They produce sentences that can be interpreted graphically to produce images of fractals or organisms. L-grammars have been applied also for modeling DNA sequences [Searls, 1993; Yokomori and Kobayashi, 1998; Mihalache and Salomaa, 2001; McGowan, 2002; Gheorge and Mitrana, 2004].

The essential difference between sequential grammars such as BNF and L-systems is that in sequential grammars the production rules are applied sequentially, one at a time; whereas in L-systems they are applied in parallel and may simultaneously replace all letters in a given word. Also, in L-systems there is no distinction between the terminal and non-terminal symbols. All symbols that appear in the grammar are valid in the final string, and any symbol in the alphabet can head a rule [Prusinkiewicz and Hanan, 1989]. The recursive nature of the L-system rules leads to self-similarity and fractal-like forms which is also a property of DNA sequences [Abramson et al., 1999].

For modeling DNA sequences we use a *context-free stochastic L-grammar*, which is defined as a tuple:

$$G = \{V, \omega, R, P\} \tag{1}$$

where:

$V = \{A, C, G, T\}$ is a set of symbols containing elements (nucleotide types, in our case) that can be replaced, i.e. the alphabet.

$\omega = V^K$ is a K-length string of symbols that define the initial state of the system.

$R \subseteq V^1 \times V^L$ is a finite set of production rules defining the way symbols can be replaced with combinations of other symbols. A rule consists of two strings - the *predecessor* (an individual symbol from V) and the *successor* (L-length string composed from V elements). Each production rule refers only to an individual symbol and not to its neighbors.

P is a set of probabilities $p_j \in [0,1]$ that a production rule $r_j \in R$ will be applied.

Machine Learning Using Support Vector Machine

Support Vector Machines (SVM) [Vapnik, 1998] are one of the most popular tools in bioinformatics for supervised classification of genetic data (biosequences, protein structure data, microarray gene expression, etc.). SVM is a structural risk minimization-based method for creating binary classification functions from a set of labeled training data. SVM requires that each data instance is represented as a vector of real numbers in *feature space*. Hence, if there are categorical attributes, we first have to convert them into numeric data. First, SVM implicitly maps the training data into a (usually higher-dimensional) feature space. A *hyperplane* (decision surface) is then constructed in this feature space that bisects the two categories and maximizes the margin of separation between itself and those points lying nearest to it (the *support vectors*). This decision surface can then be used as a basis for classifying vectors of unknown classification.

Consider an input space X with input vectors $x_i \in X$, a target space $Y = \{1, -1\}$ with $y_i \in Y$ and a training set $T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. In the SVM classification, separation of the two data classes $Y = \{1, -1\}$ is done by the *maximum margin* hyperplane, i.e. the hyperplane that maximizes the distance to the closest data points and guarantees the best generalization on new examples. In order to classify a new point x_j, the classification function $g(x_j)$ is used:

$$g(x_j) = \operatorname{sgn}\left(\sum_{x_i \in SV} \alpha_i y_i K(x_i, x_j) + b \right) \tag{2}$$

where: SV are the support vectors, $K(x_i, x_j)$ is the kernel function, α_i are weights, and b is the offset parameter.

If $g(x_j) = +1$, x_j belongs to the *Positive* class, if $g(x_j) = -1$, x_j belongs to the *Negative* class, if $g(x_j) = 0$, x_j lies on the decision boundary and can not be classified.

Here we have a binary classification problem in which the outcomes are labeled either as positive (*P*) or negative (*N*) class. There are four possible outcomes from a binary classifier:

- *True positive* (*TP*) – the outcome from a prediction is *P* and the actual value is *P.*
- *False positive* (*FP*) – the outcome from a prediction is *P* and the actual value is *N.*

– *True negative* (*TN*) – both the prediction outcome and the actual value are *N.*

– *False negative* (*FN*) – the prediction outcome is *N* while the actual value is *P*.

The efficiency of the classification function $g(x_j)$ can be evaluated using several different metrics such as Sensitivity, Specificity, F-measure or Mathew's Correlation Coefficient. Here we use a simple *Accuracy* (ACC) metric, which is a measure of how well a binary classification test correctly identifies or excludes a condition.

$$ACC = \frac{n_{TP} + n_{TN}}{n_P + n_N} \cdot 100\% \qquad (3)$$

where n_i is the number of i cases in the classification results.

The best possible classification method would yield 100% accuracy (all TPs and no FPs are found).

Grammar Rule Inference

Derivation of formal grammar rules, also known as *grammatical induction* or *grammar inference,* refers to the process of inducing a formal grammar (usually in the form of production rules) from a set of observations using machine learning techniques. The result of grammar inference is a model that reflects the characteristics of the observed objects. Here, for inference of L-grammar rules we use a "*trial-and-error*" method [Duda, 2001]. The method suggests successively guessing grammar rules (productions) and testing them against positive and negative observations. The best ruleset is then mutated randomly following the ideas of genetic programming until no improvement in accuracy is found within a certain number of iterations.

The rules of the L-system grammar are applied iteratively and simultaneously starting from the initial string. Let us denote $L(G)$ all those strings over V that can be generated by starting with the start string ω and then applying the production rules in R with probabilities P. We describe the grammar inference problem as a classical optimization problem as follows. Determine G, where V is constant, from a given set of input sequences $X = V^N$ such as to achieve the best classification $c(X')$:

$$c(X') = \max \sum_{X'} g(x_j) \qquad (4)$$

where: $x_j, x_j \in X' \subset L(G)$ are strings produced by production rules R with probabilities P, and $g(x_j)$ is the classification function trained on a set of input sequences $x_i \in X$.

Case Study: Derivation of L-Grammar Rules from Promoter Datasets

Datasets

We use the following datasets:

1) The 2002 collection of data of drosophila (*D. melanogaster*) core promoter regions [Drosophila]. The test file contains 6500 examples (1842 promoters, 1799 introns, and 2859 coding sequences), each 300 bp length.

2) The collection of data of human and additional eukaryotic (vertebrate) promoter regions. The promoters were extracted from the Eukaryotic Promoter Database rel. 50; the negative set contains coding and non-coding sequences from the 1998 GENIE data set [Human]. The test file contains 5800 examples (565 promoters, 4345 introns, and 890 coding sequences), each 300 bp length.

The positive set consists of promoter sequences and the negative set consists of introns and coding sequences.

Description of L-grammar rule generation

The L-grammar rule generation process is performed in two stages as follows:

1) Derivation of a learned classifier. The promoter dataset is used for training a SVM classifier. As a classifier, we use SVM^{light}, which is an implementation of SVM in C [SVMlight]. Nucleotide sequences are mapped onto a feature space using *orthogonal encoding*, where nucleotides in a DNA sequence are viewed as unordered categorical values and represented by the 4-dimensional orthogonal binary vectors: $\{A \to 0001, C \to 0010, G \to 0100, T \to 1000\}$. For training, the linear kernel function is used. The result is a model of a learned classifier that can separate promoter and non-promoter sequences.

2) L-grammar rule induction. The classifier model is used for evaluating the fitness of the generation of L-grammar rules. Rules are generated by L-grammar rule generator using a directed random search method: the best ruleset so far is mutated randomly and used to generate 1000 of 300 bp length L-system strings. The mutation parameters are the start string, successor strings and production rule probabilities. The number of rules as well as the predecessor strings is fixed. There are four rules for each type of nucleotide: A-rule, C-rule, G-rule and T-rule. The generated strings are encoded as 4-dimensional orthogonal binary vectors, fed to the trained SVM classifier and the accuracy of the classification is obtained. If the accuracy of the mutated ruleset is better than the accuracy of the previous best ruleset, the mutated ruleset is saved as the best ruleset; otherwise the previous best ruleset is retained. The process is continued until the required accuracy is achieved.

The structure of the L-grammar induction system is summarized in Fig. 1.

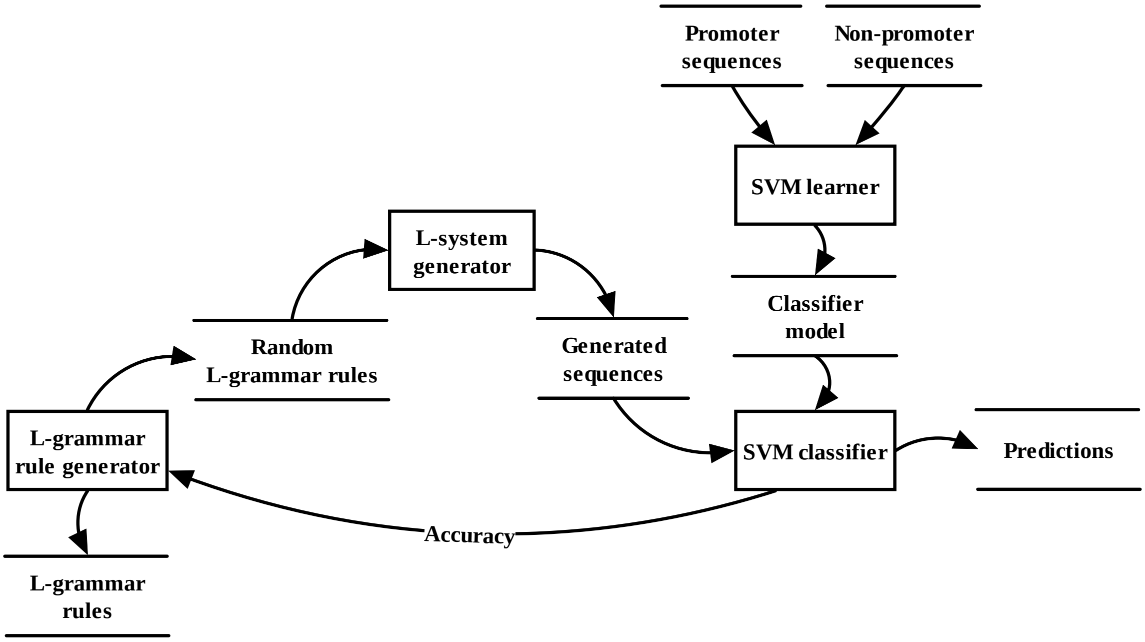


Fig. 1. Structure of the L-grammar induction system

Results

The classification was done in two stages: (1) SVM was trained using promoter vs. non-promoter sequences, and the trained classifier was used to classify (2a) random sequences labeled as promoters, and (2b) artificially generated (from the induced L-grammar rules) sequences labeled as promoters. The classification results using the Accuracy metric (see Eq. 3) are presented in Table 1.

We can see that artificially generated sequences can be classified vs. non-promoter sequences almost as good as real (natural) promoters. That suggests that induced L-grammar rules indeed capture some essential dataset properties of promoter sequences. The results are worse for the vertebrate dataset, because vertebrate promoters have more complex patterns with more irregularities and exceptions.

The derived stochastic L-grammar rules are presented in Fig. 2, a) and Fig. 2, b) for drosophila and vertebrate promoter sequences, respectively. Note that in drosophila grammar rules C and in vertebrate grammar rules T symbols are completely missing from the right (successor) side of production rules.

Table 1. Classification results

Classified sequences	No. of sequences	Classification accuracy	
		Drosophila dataset	Vertebrate dataset
Promoters vs. non-promoters	6500/5800	99.74%	94.67%
Random sequences	1000	3.3%	1.2%
Artificially generated sequences	1000	93.40%	92.10%

a)

```
Variables: A, C, G, T
Start:     AAACTAAT
Rules:     0.85: (A -> TATA),
           0.94: (C -> TA),
           0.91: (G -> TAG),
           0.10: (T -> TGA)
```

b)

```
Variables: A, C, G, T
Start:     A
Rules:     0.50: (A -> CGGAA),
           0.86: (C -> CCCCG),
           0.19: (G -> ACGG),
           0.50: (T -> AA)
```

Fig. 2. Stochastic L-grammar rules for generating
a) drosophila, and b) vertebrate promoter-like sequences

Evaluation

After analyzing the induced promoter sequence production rules, we can conclude that these rules can fairly good characterize the subsequences that are typical for promoters. The drosophila promoter production rules feature the TATA successor string, which matches the TATA-box (TATAA or TATAAA) typical for the drosophila promoters. Other subsequences typical for the promoter sequences are produced by the successive application of the production rules, e.g., the *Pribnow* box (TATAAT) is produced by two successive applications of A-rules.

The vertebrate promoters are typically more complex than the drosophila promoters, because there are many TATA-less promoters, which are characterized by other more complex subsequences. These subsequences also can be produced from the induced grammar rules, e.g., the CACG subsequence characteristic to the *E-box* (CACGTG) is produced by the successive application of the A-rule and G-rule. The AACC subsequence characteristic to the *Y-box* (GGGTAACCGA) is produced by the successive application of the G-rule, A-rule, and C-rule. Human promoter sequences are also characterized by the occurrence of DPE (*downstream promoter element*), a distinct 7-nucleotide subsequence (A/G)G(A/T)CGTG that is similar to the derived G-rule.

Conclusion

The advantage of the machine learning method combined with formal grammar is that additionally to recognition which does not provide any structural information, the modular structure of the analyzed genetic sequences can be identified. The structural analysis of the derived L-grammar production rules allows identifying common promoter sequence elements (so called “boxes”) such as TATA-box, Pribnow box, E-box, Y-box and DPE.

The classification results of the artificial promoter sequences generated using the derived L-grammar rules are almost as accurate as that of the natural promoters, thus showing a great deal of similarity between the discriminating features of both types of sequences.

Bibliography

[Abramson et al., 1999] Abramson, G., Cerdeira, H.A., Bruschi, C.: Fractal properties of DNA walks. Biosystems 49(1): 63-70 (1999)

[Denise et al., 2003] Denise, A., Ponty, Y., Termier, M.: Random generation of structured genomic sequences. Proc. of Int. Conf. on Research in Computational Molecular Biology (RECOMB'03), Berlin, Germany, 10-13 April (2003)

[Drosophila] Berkeley Drosophila Genome Project. Drosophila promoter dataset. Available at: http://www.fruitfly.org/seq_tools/datasets/Drosophila/promoter/

[Duda, 2001] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2 ed.), John Wiley & Sons (2001)

[Fernau, 2003] Fernau, H.: Parallel Grammars: A Phenomenology. Grammars 6(1): 25-87 (2003)

[Gheorghe and Mitrana, 2004] Gheorghe, M., Mitrana, V.: A formal language-based approach in biology. Comparative and Functional Genomics 5: 91–94 (2004)

[Grate et al., 1994] Grate, L., Herbster, M., Hughey, R., Haussler, D.: RNA modelling using Gibbs sampling and stochastic context-free grammars. Proc. of the 2nd Int. Conf. on Intelligent Systems for Molecular Biology. AAAI/MIT Press (1994)

[Human] Berkeley Drosophila Genome Project. Human promoter dataset. Available at: http://www.fruitfly.org/seq_tools/datasets/Human/promoter/

[Infante-Lopez and de Rijke, 2004] Infante-Lopez, G., de Rijke, M.: Alternative approaches for generating bodies of grammar rules. Proc. of 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004), Barcelona, Spain, 21-26 July, 454-461 (2004)

[Jiménez-Montaño, 1984] Jiménez-Montaño, M.A.: On the Syntactic Structure of Protein Sequences and the Concept of Grammar Complexity. Bull. Math. Biol. 46: 641-659 (1984)

[Koza, 1993] Koza, J.R.: Discovery of Rewrite Rules in Lindenmayer Systems and State Transition Rules in Cellular Automata via Genetic Programming. Symposium on Pattern Formation (SPF-93), February 13, Claremont, CA (1993)

[Lindenmayer, 1968] Lindenmayer, A.: Mathematical models for cellular interactions in development. Journal of Theoretical Biology 18: 280-315 (1968)

[Marcus, 1974] Marcus, S.: Linguistic structures and generative devices in molecular genetics. Cahiers Ling Theor. Appl. 1: 77–104 (1974)

[McGowan, 2002] McGowan, J.F.: Nanometer Scale Lindenmayer Systems. Proc. of SPIE Vol. 4807 (2002)

[Mihalache and Salomaa, 2001] Mihalache, V., Salomaa, A.: Lindenmayer and DNA: Watson-Crick D0L Systems. In G. Paun, G. Rozenberg, A. Salomaa (Eds.), Current Trends in Theoretical Computer Science, 740-751 (2001)

[Monteiro, 2005] Monteiro, M.I., de Souto, M., Gonçalves, L., Agnez-Lima, L.F.: Machine Learning Techniques for Predicting Bacillus subtilis Promoters. Advances in Bioinformatics and Computational Biology. LNCS 3594. Springer (2005)

[Muggleton, 2001] Muggleton, S.H., Bryant, C.H., Srinivasan, A., Whittaker, A., Topp, S., Rawlings, C.: Are grammatical representations useful for learning from biological sequence data? - a case study. Journal of Computational Biology 8(5), 493-522 (2001)

[O'Neill et al., 2004] O'Neill, M., Brabazon, A., Adley, C.: The Automatic Generation of Programs for Classification Problems with Grammatical Swarm. Proc. of Congress on Evolutionary Computation CEC 2004, Portland, USA, 104-110 (2004)

[Prusinkiewicz and Hanan, 1989] Prusinkiewicz, P., Hanan, J.: Lindenmayer Systems, Fractals, and Plants (Lecture Notes in Biomathematics). Springer-Verlag (1989)

[Prusinkiewicz and Lindenmayer, 1990] Prusinkiewicz, P., Lindenmayer, A.: The Algorithmic Beauty of Plants. Springer-Verlag: New York (1990)

[Ranawana and Palade, 2005] Ranawana, R., Palade, V.: A neural network based multiclassifier system for gene identification in DNA sequences. Journal of Neural Computing Applications 14: 122–131 (2005)

[Sakakibara et al., 1994] Sakakibara, Y., Brown, M., Hughey, R., Mian, I.S., Sjoelander, K., Underwood, R., Haussler, D.: Stochastic context-free grammars for tRNA modelling. Nucleic Acids Res 25: 5112–5120 (1994)

[Searls, 1993] Searls, D.B.: The computational linguistics of biological sequences. In Hunter, L. (ed.): Artificial Intelligence and Molecular Biology, 47–120. AAAI/MIT Press (1993)

[Sobha Rani et al., 2007] Sobha Rani, T., Durga Bhavani, S., Bapi, R.S.: Analysis of E.coli promoter recognition problem in dinucleotide feature space. Bioinformatics 23(5):582-588 (2007)

[SVMlight] SVMlight. Available: http://svmlight.joachims.org/

[Unold, 2007] Unold, O.: Grammar-Based Classifier System for Recognition of Promoter Regions. In Adaptive and Natural Computing Algorithms, LNCS Vol. 4431. Springer Berlin/Heidelberg (2007)

[Vapnik, 1998] Vapnik, V.: Statistical Learning Theory. Wiley-Interscience, New York (1998)

[Werner, 2003] Werner, T.: The state of the art of mammalian promoter recognition. Briefings in Bioinformatics 4(1):22-30 (2003)

[Yokomori and Kobayashi, 1998] Yokomori, T., Kobayashi, S.: Learning local languages and their application to DNA sequence analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence 10(20): 1067-1079 (1998)

Authors' Information

***Robertas Damaševičius** – Assoc. Prof.; Software Engineering Department, Kaunas University of Technology, Studentų 50, LT-51368, Kaunas, Lithuania; e-mail: robertas.damasevicius@ktu.lt*

AUTOMATIC GENERATION OF TITLES FOR A CORPUS OF QUESTIONS

Jesús Cardeñosa, Carolina Gallardo

***Abstract**: This paper describes the followed methodology to automatically generate titles for a corpus of questions that belong to sociological opinion polls. Titles for questions have a twofold function: (1) they are the input of user searches and (2) they inform about the whole contents of the question and possible answer options. Thus, generation of titles can be considered as a case of automatic summarization. However, the fact that summarization had to be performed over very short texts together with the aforementioned quality conditions imposed on new generated titles led the authors to follow knowledge-rich and domain-dependent strategies for summarization, disregarding the more frequent extractive techniques for summarization.*

***Keywords**: Summarization, text processing, subjective clustering.*

***ACM Classification Keywords**: H.3.1 Content Analysis and Indexing*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction to the Problem: context and antecedents

Management information based on summaries is frequent in libraries, documentation centers, press or document collections based on short texts. Some libraries have at their disposal information based on summaries of articles or abstracts that have been not generated by the authors but by documentalists. These abstracts have been made in such a way so that information search can be done over the complete document or over the abstract instead. Frequently, these summaries are indexed with some keywords that identify them thematically. Another case is given by news. News are relatively short documents but are identified by means of a title. The title is defined by the author of the new and it serves to search over this news. Titles can be conceived of the most condensed abstraction of the contents of a document. Thus, summaries do not serve a unique function: they can be used for information searches or as indexes of their documents, mainly.

In the area of automatic summarization, summaries can vary in their *form*. Form relates to the way that the summary has been produced, thus the summary can be composed of extracts of the document (extractive summarization) or by and abstract (a concise summary of the central subject matter of a document). Obviously, techniques used for extractive summarization are different from those used for abstracting and more frequent as well. Extractive summarization tries to identify the relevant sentences of the document. To distinguish which sentences are relevant from irrelevant ones, several criteria are used: a positional criterion (e.g., sentences at the end or beginning of the paragraph are considered to be more relevant) was used in [Brandow et al. 95]; [Lin and Hovy, 1997] considered as relevant those sentences that contained signature words (id est, key words that defined by means of frequency measures like the tf-idf schema); on the other hand, [Osborne, 2002] classifies sentences as relevant/non relevant according to the existence of certain word pairs, sentence length, sentence position and discourse features. In essence, extractive summarization typically uses statistical techniques to extract the relevant sentences. As such, they constitute domain-independent techniques.

Non extractive trends in single-document summarization rely on knowledge-based techniques and tend to be domain-dependent approaches. A paradigmatic system, SUMMONS described in [Radev and K. McKeown, 1998], is restricted to the summarization of news about terrorism. SUMMONS firstly extracts relevant information (like places, victims, authors, date, etc.) from texts using predefined templates. Then the extracted information is passed through a language generator module which is also based on templates. Other knowledge-based approaches make use of linguistic processing, like [Tucker and Sparck Jones, 2005], together with domain knowledge, exemplified by the works of [Saggion and Lapalme, 2002] or [Hahn and Reimer, 1999]. In any case, during the last decade there has been much less research and work in knowledge-based summarization (see [Spärck, 2007] for an up-to-date review of the summarization area).

In our specific case, the problem is defined by the necessity to assign a title to the questions that belong to the opinion polls that the Centre for Sociological Research (CIS – *Centro de investigaciones sociológicas*) of the Spanish government. A question is a short text with the objective of collecting the value or one of more sociological variable. A sociological variable can vary from a specific piece of information about the interviewee (like labour situation, education level, social class, number of cars that the interviewee has, etc.) to the interviewee's opinion about a given issue, institution or public person. Questions include a number of different answer options, then interviewees have to choose between one or more, free answers are not allowed. Thus, a title for a question is more than an identifier: it is a summary of the content of the question that permits the understanding of the distribution of the frequency of answers without reading the complete text of the question.

Bearing in mind the particular nature of our problem, any extractive technique will prove inadequate for our problem. Since there is a need to **generate** new titles that are a condensed abstracting of the question, the strategies to follow will have to be knowledge-rich and domain-dependent. This article will describe the process that we have followed to automatize the process of title generation. Although our proposal is based on domain-dependent criteria, it is applicable to similar problems.

Analysis of the Domain

The specific problem is as follows. The complete corpus of different questions belonging to the opinions polls is composed of:

- 39257 questions with a title, assigned manually by experts.
- 47964 questions with no title (hereafter, they will be referred as untitled questions).

Since there is a corpus of already titled questions, the generated titles have to be coherent and similar with the manually assigned ones. To do that, we performed a thorough analysis of the domain with the main objective of searching for relations (of any kind) between the existent questions and their titles and extrapolate such relations to the corpus of untitled questions. Although there are many aspects that can be analyzed in this particular problem, we focused on two main aspects: linguistic features of titles and relations between titles and questions. The analysis is completed with the estimation of frequencies of the different types of titles and relations.

I. Linguistic features of titles

The set of titles (without questions) were thoroughly examined. Aspects like length of the title, existence of repeated titles, linguistic constructions and tackled themes were taken into account. For our purposes, we established two broad categories of titles: **subjective topic titles** and **objective specific titles**.

Subjective topic titles

This sort of titles refers to the judgement about a given topic of the interviee. The judgement can be an approval, rejection, preference, evaluation, etc. of a topic; and this sort of judgement is included in the title, together with the object of the judgement. It is remarkable that from a thematic point of view, the most frequent tackled themes are vote, elections and politicians. Besides, the titles dealing with these three topics shows less variation that others. In general, the structure of subjective topic titles follow the general schema of: *Type of judgement + Nexus + Topic.* Where type of judgement is a word like "opinion", "preference", etc., "nexus" is a the preposition that requires the initial word, and topic is the nominal group denoting what the question is about. Table 1 shows some examples.

Table 1. Example of Subjective topic titles

Type of Judgement	Nexus	Topic
Approval	of	the labour of Felipe Gonzalez as Prime Minister
Preference	between	different alternatives of territorial organization of the State
Satisfaction	with	the job of the interviees that do not study

Objective Specific Titles

These titles asked for an **objective** piece of information about the interviee. Thus, in their linguistic structure there is not an initial word denoting a judgement. There are two types:

- *Fixed*: they seem to be obligatory in all surveys and usually refer to sociodemagraphic aspects of interviees like labour situation, social class, etc. Some examples are:
    - Age of Interviewee; Head of the familiy
- *Specific*: they are specific to a given survey and usually refer to habits like smoking, sports, leisure, etc. For instance:
    - Starting Age for smoking; Number of mobile phones of the interviewee

II. Linguistic relation between titles and questions

The degree of success in automatizing the process of generating titles is directly related with the relation between a question and its title. We have identified three main types of relations:

a) The title is a **summary** of the question. As in the following example:

TITLE: Acceptance of fraudulent behaviour regarding the National Institute of Employment
QUESTION: In our society, there are happen things that are completely acceptable for some people and absolutely unacceptable for others. I am going to read you some of those things and I'd like to know whether they are acceptable or not for you. [….]
- To evade taxes | - To receive an unemployment subsidy while working. | ….

b) The topic of the title is a **nominalization** or **paraphrasis** of a part of the question.

TITLE: Frequency of attendance to religious services
QUESTION: How often do you go to mass or attend religious services if you have other religion?
- Never | - Several times a year | - Sometimes in the month

c) The topic of the title is a **literal fragment** of the question.

TITLE: Attitude towards the creation of an International Court
QUESTION: In any case, are you in favour or against the creation of an International Court of these characteristics.
- In favour | - Against | - Not know – not answer

III. Frequency of the different types of titles & questions

Our work is based on the hypothesis that the analysis results obtained from the existent corpus of titled questions are valid for the corpus of untitled questions. In this way, we assure similarity of the generated titles with the existent ones. It is important to estimate the frequency of the different types of titles and of the different types of relations between titles and questions. To estimate frequencies, we established five categories of questions that resulted from the reorganization of the classification from the analysis of the linguistic features of titles and from the relation between title and question, namely:

- From the linguistic analysis of titles, there results four categories. *Subjective topic* titles branches into two categories: a) **Topics about vote, elections and politicians**; and b) **Rest of topics**. Whereas *Fixed Objective Specific* titles divides into other two categories: c) **Fixed**; and d) **Specific.**
- From the relation between title and question, we are just interested in identifying the untreatable questions, since these are the ones that will pose more problems to the task. Thus, only one category is posed: **Untreatable questions**.

Thus, we will postulate 5 categories: (1) Topics questions about vote, elections and politicians; (2) Rest of topics questions; (3) Fixed question; (4) Specific questions; and (5) Untreateble questions. In order to estimate the frequency of these five categories, we extracted six samples of 40 questions from the corpus of titled questions. For each sample, we classified the question –along with its title- under one of the five postulated categories. Table 2 shows the estimated frequency of different types of titles.

Table 2. Frequency of different types of titles

Type of Question	Frequency
Elections, vote and politicians	14.58 %
Topic questions	12,92%
Specific questions	10.83 %
Fixed questions.	37,08 %
Untreatable questions	24,58 %

Under the hypothesis that untitled questions behave as titled ones, these frequencies helped us to estimate that around 25% of the corpus of untitled questions would remained untitled.

Development of the work: Methodology

We imposed the following working hypothesis: only questions whose title can be generated from a nominalization, paraphrasis or exact wording of a fragment of the question will be assigned a title. We will not include deep NLP processing technology but shallow language analysis and domain-dependent patterns to identify the relevant fragments of the sentence and to generate the corresponding title.

The main reason to reject deep natural understanding techniques are given by the number of language resources required, such us dictionaries, grammars for analysis and grammars for generation. A domain-dependent strategy instead does not require either deep natural language understanding or big language resources. On the other hand, it requires an exhaustive domain-analysis and a representative corpus of examples.

Let's examine the following example:

TITLE: Religiosity of the interviewee
QUESTION: How would you considered yourself regarding religion?
- Practising catholic. | - Non practising catholic | - Other religions | - Non believer | - Do not answer

This specific question and its title are repeated around 800 times in the corpus of titled questions however it will fall into the category of untreatable questions. It represents a simpler case where to fairly repeated string it is assigned a fixed title, disregarding any kind of linguistic analysis.

Thus the chosen technique was a domain-dependent one. This option was supported by the following facts: a) high frequency of fixed questions; b) homogeneity in the existent titles; and c) the results of the analysis of the domain already pointed out at patterns.

Strategies for the different types of questions

Each type of questions have a different strategy. For example, fixed questions are the simplest, their treatment imply looking for a specific string and assign another string, while the rest of categories require to look for the topic, type of opinion, etc. Let's have a look at how each different type have been dealt with.

I. Questions about vote, elections and politicians

The questions about these topics fairly frequent in the corpus of both titled and untitled questions. They show little variability in their wording, being the variable elements the type of election (European, general, regional or municipal), the date of the election or the politician that is being evaluated.

The **strategy** for generating the titles for this type of question is a) to associate a specific sequence of words in the question to a specific **partial title**; b) identify the **variable elements** (type and date of election, name and position of the politician); and c) concatenate the different elements. That is, these titles have the following general schema:

TITLE → Partial title + Variable element

For example, pattern 1 shows one specific pattern for the identification of a partial title.

PATTERN 1. PARTIAL TITLE: "INTENTION OF VOTE IN …"

```
[Q]/Do you think you will vote in/ → [PT] "Intention of vote in"
```

That is, if the string "do you thing you will vote in" matches the question [Q], then identify the partial title [PT] as "Intention of vote in".

Pattern 2 identifies one variable element (type of election).

PATTERN 2. VARIABLE ELEMENT: TYPE OF ELECTION

```
[Q] /<these|the|forthcoming> elections X PUNCTUATION/ → [VE] "the X
elections"
```

In this case, if the question matches any of the words <these|the|forthcoming> followed by the string "elections" and any string (denoted by X) until a punctuation mark, then identify the variable element [VE] as the string "elections X".

The following question and title illustrate these patterns:

QUESTION: Do you think you will vote in the elections to the European Parliament, to be celebrated next 15th June.
TITLE: Intention of vote in the elections to the European Parliament.

II. Rest of topic Questions

The **strategy** to generate the titles for these questions follows two steps: a) identify the sort of judgement that is being required to the interviee (that is, an opinion, approval, reason or evaluation); and b) identify the topic of the question. The final title will be the concatenation of the type of judgment and the topic:

TITLE → Type of judgment + topic

Both elements are extracted with the help of regular expressions, but in some cases some linguistic knowledge is also used. Pattern 3 shows an specific example to identify the type of judgment, whereas pattern 4 identifies the topic of a question.

PATTERN 3. JUDGEMENT TYPE: "DEGREE OF AGREEMIENT WITH"

```
[Q] /TELL <the|your> degree of agreement/ → [JT] "Degree of agreement with"
```

That is, if the question is composed of any form of the verb "to tell" followed by determiner "the" or pronoun "you", followed by the string "degree of agreement with", identifies the judgement type as "degree of agreement with".

PATTERN 4. TOPIC

```
[Q] /going to read (you) some <opinions | statements> about X FINAL_MARK/ →
[TOPIC]: "some opinions about X"
```

That is, if the question matches the string "going to read you some opinions about" and its possible variations (presence or absence of "you", disjunction between "opinions" or "statements") followed by any string (represented by X) until a final mark, identify the topic of the sentence as "some opinions about X".

FINAL MARK

```
[FINAL_MARK]: <;|:|.|CLAUSE>
```

That is, a final mark can be any of the following: a semi-colon, a full stop, colon or the beginning of a new clause (for example: pronoun + verb).

These patterns show how (very) shallow linguistic knowledge aids to the task of identifying patterns. They can be directly applied to the following question, where both the **type of judgement** and **topic** are highlighted:

QUESTION: Now, I am going to read **some opinions about the development of the State of Autonomies** and I would like you to tell me your **degree of agreement with** each of them. Autonomous Communities …
- Have contributed to …

TITLE: “degree of agreement with the development of the State of Autonomies”.

III. Specific Questions

These questions usually ask for objective data about the interviee. Thus they present more variability in the topics of the questions. For this reason, the strategy is slightly different from the previous ones: patterns for these questions do not rely on the identification of a particular string in the sentence but on the identification of specific linguistic constructions. Thus, for the process to be successful, it is required that the linguistic constructions present in the questions are homogeneous. Consequently, this type of questions are the ones that require more linguistic knowledge for their processing, in particular, the recognition of the shallow structure of wh- questions. Patterns 5 and 6 are two paradigmatic examples for specific questions.

PATTERN 5. TITLE: “NUMBER OF TIMES …”

```
[Q]: /How many times (in total) have you X?/ → [T] “Number of times that the interviee has X”
```

That is, if the question matches the string “how many times have you” followed by any string (represented by “X”) and a question mark, identify the title [T] of the question as “Number of times that the interviee has X”.

This pattern has been directly applied to the following question:

QUESTION: How many times have you been hospitalized in the last twelve months?
TITLE: Number of times that the interviee has been hospitalized in the last twelve months.

PATTERN 6. TITLE: “PERSON/ENTITY THAT …”

```
[Q]: /Who do you <think|believe> that X <:|?>/ → [T] “Person/Entity that X”
```

That is, if the question matches the string “who do you think that” (or “believe”, instead) followed by any string (represented by “X”) and a question mark, identify the title [T] of the question as “Number of times that the interviee has X”.

Pattern 6 generates the following title:

QUESTION: Who do you think that should provide information about the social and sanitary assistance and services for old people?
TITLE: Person or entity that should provide information about the social and sanitary assistance and services for old people

IV. Fixed Questions

The strategy for the generation of this titles does not require any linguistic knowledge. The process merely consists in the assignment of a fixed title (without variations) to questions that present a particular wording with hardly any variation. Patterns 7 and 8 deal with two different types of fixed questions.

PATTERN 7. FIXED TITLE

```
[Q]: /what is your social class/ → [T]: “Subjective social class of the interviee”
```

That is, if the question matches the string “what is your social class”, identify its title as “Subjective social class of the interviee”

PATTERN 8. FIXED TITLE

```
[Q]: /Which of the following describes your current situation?/ → [T] “Labour situation of the interviee”
```

That is, if the question matches the string “Which of the following describes your current situation?”, identify its title as “Labour situation of the interviee”.

V. Untreatable questions

These questions do not have any feature to identify them. In this case, we do not attempt to generate a title.

Finally, in addition to untreatable questions, there is a set of questions that are left untitled intentionally. These questions does not have enough content to generate a title. In particular, they present any of the following characteristics:

a) The question has less that 4 words. For example:

QUESTION: And why not?

b) The question ends with suspensions dots. For example:

QUESTION: What do you think about …?

c) The questions contains enclitic pronouns and general terms like “that”, “statement”, “reason” in their topic. For example:

QUESTION: Do you agree with that?

Thus, any question with any of the aforementioned characteristics is left out from the process of title generation. They are referred to as **filtered questions**.

Results

There are two main aspects to be evaluated: the quantity and the quality of the generated titles. Quantitative results are summarized in table 3. As can be seen, at the end of the process, we were able to generate 22347 titles and leaving apart 1627 questions as filtered ones. This means that we automatically titled around 47% of the questions.

Table 3. Results for untitled questions

Type of Question	Number
Titled questions	22347
Untitled questions	23990
Filtered Questions	1627
TOTAL	**47964**

We also reviewed the quality of the generated titles. To do that, we evaluate the quality of titles of six samples of 50 titles each. The average percentage of correct titles for all the samples was **96%.**

Thus after the evaluation, we can ask ourselves again whether our initial hypothesis were correct. From the quantitative point of view, our hypothesis about the frequency of the different types of questions is only partially correct. Untreated questions represented 24% of the titled questions, whereas they represent 50% of untitled questions. Fixed questions are also less numerous in the corpus of untitled questions. However, from a qualitative point of view, we have assured homogeneous and correct titles.

Conclusions

There are several conclusions from this work. The first one made us think about the differences in the distribution of the frequency of the different types of questions. This shift is sometimes due to the evolution of society. The sociological changes are reflected in the topics of the questions.

The second one refers to the followed techniques and strategies. In this work, domain-analysis and ad hoc patterns prevails over domain-independent linguistic processing. Linguistic processing is kept to a minimum, since the linguistic resources are expensive. On the other hand, domain-dependent strategies prove to be highly efficient while quick to be developed.

Bibliography

Brandow, R., Mitze, K., and Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. Information Processing and management, 31(5), 675–686

Hahn, U., and Reimer, U. (1999). ‘Knowledge-based text summarisation: Salience and generalisation for knowledge base abstraction’. In Mani and Maybury, eds. (1999). Advances Advances in automatic text summarisation. Cambridge, MA: MIT Press. (pp. 215–222).

Lin, C., and Hovy, E. (1997). Identifying topics by position. In Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP--97), 283--290.

Osborne, M. (2002). Using maximum entropy for sentence extraction. In Proceedings of the Acl-02 Workshop on Automatic Summarization - Volume 4

Radev, D. R. and McKeown, K. (1998). Generating Natural Language Summaries from Multiple On-Line Sources. Computational Linguistics, vol 24(3): 469-500.

Saggion, H. and Lapalme, G. (2002). Generating informative-indicative summaries with SumUM. Computational Linguistics, 28(4), 497–526.

Spärck Jones, K. (2007) Automatic summarising: The state of the art. Information Processing and Management, 43: 449-1481

Tucker, R. I. and Sparck Jones, K. (2005). Between shallow and deep: An experiment in automatic summarising, Technical Report 632, Computer Laboratory, University of Cambridge.

Authors' Information

Jesús Cardeñosa – *Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es*

Carolina Gallardo – *Group of Validation and Industrial Applications. Escuela Universitaria de Informática. Universidad Politécnica de Madrid. Carretera de Valencia Km.7. 28041 Madrid; email: cgallardo@eui.upm.es. http://www.vai.dia.fi.upm.es.*

EXPERIMENTAL SUPPORT OF ARGUMENT-BASED SYNTACTIC COMPUTATION

Velina Slavova, Alona Soschen

Abstract: *Linguistic theory, cognitive, information, and mathematical modeling are all useful while we attempt to achieve a better understanding of the Language Faculty (LF). This cross-disciplinary approach will eventually lead to the identification of the key principles applicable in the systems of Natural Language Processing. The present work concentrates on the syntax-semantics interface. We start from recursive definitions and application of optimization principles, and gradually develop a formal model of syntactic operations. The result – a Fibonacci-like syntactic tree – is in fact an argument-based variant of the natural language syntax. This representation (argument-centered model, ACM) is derived by a recursive calculus that generates a mode which connects arguments and expresses relations between them. The reiterative operation assigns primary role to entities as the key components of syntactic structure. We provide experimental evidence in support of the argument-based model. We also show that mental computation of syntax is influenced by the inter-conceptual relations between the images of entities in a semantic space.*

Keywords*: natural language, mathematical modeling, cognitive modeling*

ACM Classification Keywords*: I.2 Artificial Intelligence, 1.2.0. Cognitive simulation,1.2.7. Natural language processing - Speech recognition and synthesis****.***

Conference*: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

We use mathematical formalism of Generalized Nets to develop a stage-simulating model of NLP. This formal approach allows a more exact representation of information flows during the stages of processing, expressed as the transitions Z_1—Z_{29} of the Net (Slavova 2004). The analyses performed on this basis suggest that information treatment consists of the operations that use two types of Long Term Memory knowledge (syntactic and semantic) in parallel. As an example, this is the case of transition $Z_{27,}$ which expresses the stage when the system builds the syntactic structure of a sentence after its last word-form was stored in Working Memory (figure 1.). A detailed examination of the incoming information flow allows us to suggest that the procedure, running on $Z_{27,}$ must use semantic and syntactic knowledge in parallel. We assumed that **syntactic structure is better clarified** when it receives semantic justification.

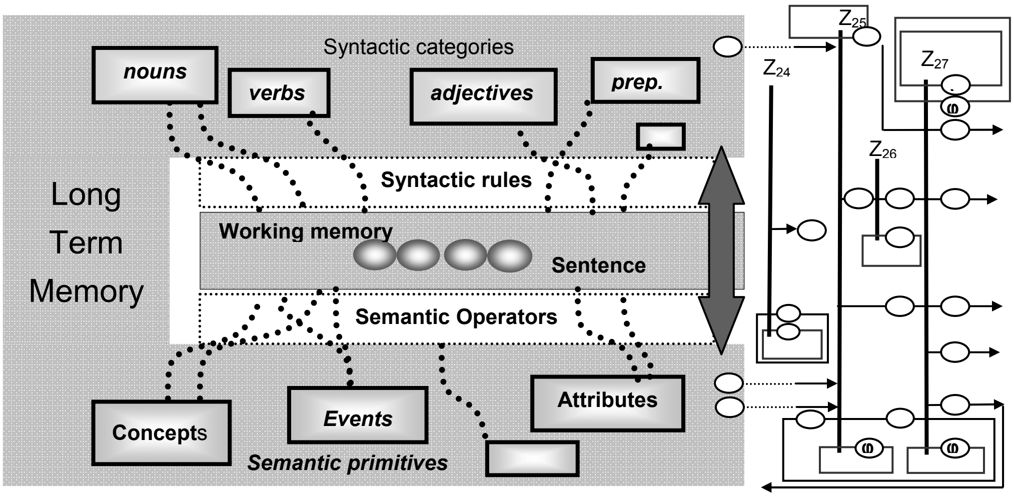


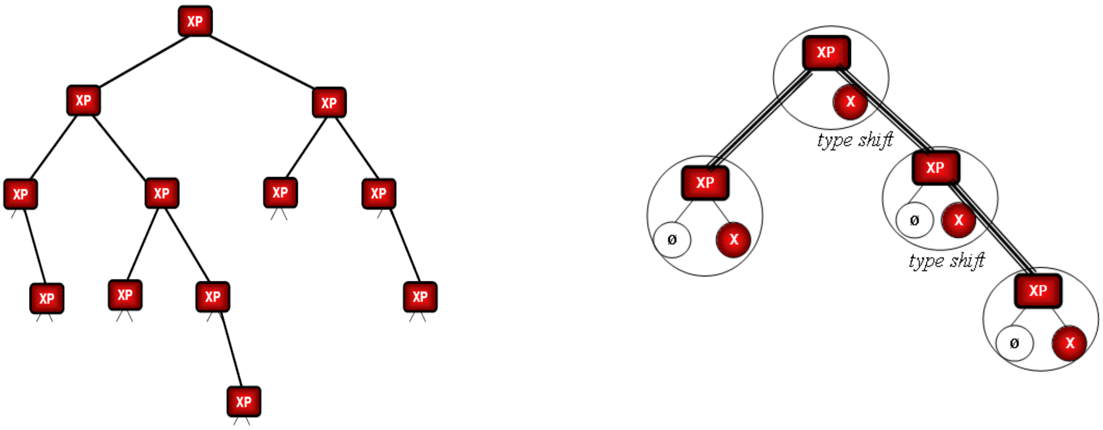
Figure 1. Information treatment of a sentence, based on language and semantics

For further analyses, the two types of knowledge stored in Long Term Memory were modeled by means of a database structure that shows the interconnection of syntactic rules, semantic primitives, and semantic operators (Slavova, Soschen, Immes, 2005). The assumption was that language units (word-forms) have images as semantic primitives such as "concepts", "attributes", "events" etc, and that grammatical rules comply with semantic operations on these primitives. This formalization of the Language as a "joint" Information System was used to study a particular language rule - secondary predication in Russian[1]. This rule was modeled by means of the formal approach described above. That led to a coherent and well-defined formal procedure and confirmed that the rule entails operations on semantic primitives.

Further efforts are put forward to obtain the proof that **semantic knowledge** and **syntax** are interrelated. The question so far is how syntax is related to operations on semantic primitives – concepts, events, attributes, etc. This is one of the most important questions in contemporary linguistics and cognitive science.

Syntax as Computation

Following one of the widely accepted linguistic theories, the key component of Faculty of Language (FL) is a computational system (narrow syntax) that generates internal representations and maps them into the conceptual-intentional interface by the (formal) semantic system (Hauser et al., 2002). There is a consensus that the core property of FL is *recursion*, which is attributed to narrow syntax. In other words, the process of mental generation of syntactic structures relies on the capacity of the human brain to perform specific operations in compliance with the principles of efficient computation. The claim in the recent theories is that this computation is based on a primitive operation that takes already constructed objects to create a new object. This basic operation, called "Merge", provides a "language of thought", an internal system to allow preexistent conceptual resources to construct expressions (Chomsky, 2006). Although these questions receive a lot of attention, there are no convincing proposals yet concerning the precise type of resources on which such computation is performed in a recursive manner to build syntactic structures.



a. Figure. 2. b.

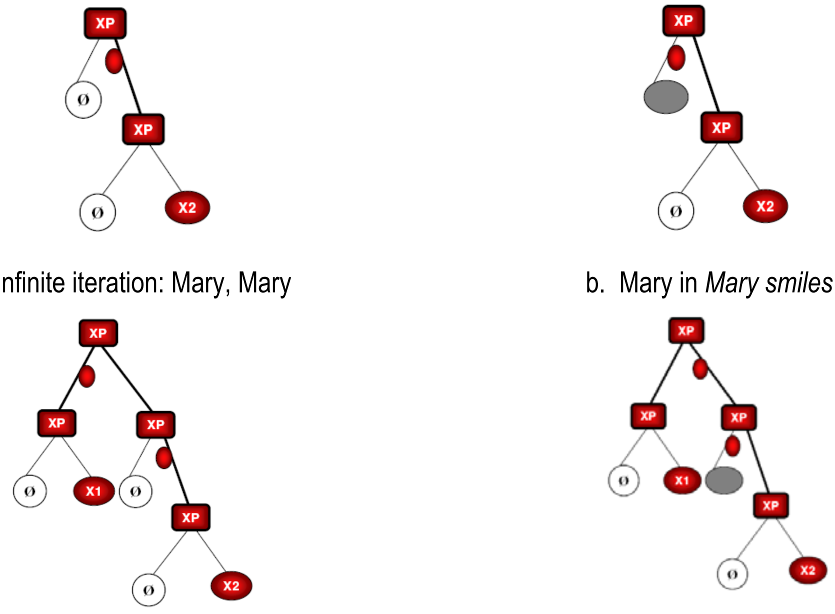
Following from the above, the study of syntactic recursion by mathematical means may provide valuable insights into the principles underlying the human language. One step in this direction was provided in Slavova and Soschen (2007). Syntactic structures, presented in the traditional sense of Chomskyan theory (Bare Phrase Structures, XP-structures), were re-defined in terms of finite recursive binary trees. The "traditional syntactic tree" does not correspond to the finite nature of a sentence; consequently, it cannot be defined recursively as a finite object. Another reason to introduce this modification is to build a structure that complies with the principles of optimization, namely with the principle of efficient growth (Soschen 2006, 2008). The tree was modified; the

[1] The linguistics theories don't provide a consistent explanation of Secondary Predication in Russian.

nodes related to syntactic role of verbs were discarded. The structure obtained in this way is a tree of Fibonacci (figure 2. a).

This tree can be seen as is an operator – it "performs" a bottom-up Merge (figure 2.b.), its nodes are the results of Merge. In the model under development, XPs are *sets*, Xs are 'unbreakable' *entities*, and Merge can be applied to two non-equivalent substances (the tree has ordered nodes). These formal transformations of the traditional tree result in a structure that incorporates two operations of fundamental importance in the syntactic model. The first is "Ø-Merge", operation that takes place at the point where Xs as initial substances form *singleton sets,* ready for further syntactic computation. The second is *type-shift,* which results in a transition from *sets* (XPs) to *entities* Xs and expresses a property of the dual mental representation of XP as either consisting of two separate elements or as an 'unbreakable' whole (part of a larger unit).

The Fibonacci-like tree shows the patterns of relating arguments (Soschen 2006, 2008). An important question is the height h of the XP Fibonacci-tree, since it refers directly to the memory, necessary for the computation. The tree is a recursive object; the same patterns of Merge are repeated at its levels. It is easy to show that merge-patterns start to reiterate when h>3 and that any tree with h>3 can perform more than one merge-pattern. We defined the tree with h=3 as the basic tree (fig. 2.b). We interpret its properties as follows: the basic tree defines the maximal number of Xs that can be merged in a procedurally unambiguous way. It could be suggested that this structure is determined in the same way as the number of nodes and relations that can be treated by the human brain within a semantically meaningful argument space.



a. Infinite iteration: Mary, Mary

b. Mary in *Mary smiles*.

c. Two arguments Mary, John in Mary loves John.

d. Three arguments Mary, John, apple in Mary gave John an apple.

Figure 3.

The tree represents a bare (label-free) syntactic structure that has no lexical input; what it has are the paths that connect smaller units in order to produce a larger meaningful unit. We called the tree in (fig. 2.b) "the Argument-Based Syntactic Tree".

According to the hypothesis put forward in Soschen (2005, 2006, 2008), a general rule governing efficient growth applies in syntax in such a way that minimal syntactic constituents incorporate arguments (*agent, recipient, theme*) which are related to each other. In the Fibonacci-tree model, the type of merge configuration determines the type of relation between arguments. The maximal configuration (fig. 3.d) corresponds to thematic roles *agent*, *recipient*, and *theme*. The “syntactic meaning” of the schemes in (fig. 3) corresponds to configurations offered in (Soschen 2006, 2008).

These schemes represent all possible configurations and relations between arguments in the human theta-role Semantic Space. Carnie (2006) shows convincingly that the number of arguments in a thematic domain is necessarily limited to three, a fact that has not found an explanation in linguistics so far. The model under development suggests that the number of arguments is limited in a particular way in compliance with the principles of efficient growth, which are, in our terms, the principles of efficient computation as well.

Of importance to linguistic theory is our proposal that the argument-based model of syntax has a fundamental character. This model shows that syntax utilizes recursive calculus to connect *arguments* and express *relations* between them. The argument-based model assigns a primary syntactic role to *entities*, usually expressed as nouns. This viewpoint is in contrast with *verb*-centered models of syntax.

Our efforts are focused on the experimental evidence that supports the argument-based model. The difficulty of designing an appropriate experiment is that mental computation runs on a deep (pre-linguistic) level and cannot be captured on the lexical level by a standard experiment. One possible way to extract some information about the primary mechanisms is to force the mental system to solve ambiguities on the lexical level and to analyze the system’s response.

Experimental Design

Bulgarian is the only Slavic language which, during the last 10 centuries, has undergone a transition from synthetic to analytical language. Prepositions replaced case-flections, and a suffixed definite article appeared. One interesting result of the transition is that the Genitive and Dative cases are both expressed by means of the preposition *‘на’* (na). “Na” has several meanings: to, of, on. Our experiment is based on the following two meanings of “Na”:

1. *Of – meaning* (whose, Slavonic Genitive)

The X *на* *the* Y means “the X of the Y” i.e. “the Y’s X”, as in:

The X	*На*	The Y	
Къщ*ата*	*На*	Куч*ето*	
The house	*Of*	*The dog*	*The dog's house*

2. *To – meaning (to whom, Slavonic Dative)*

Subject Verb *на* *the* Y means that the subject S acts To the Y. For transitive verbs, *на* assigns the syntactic role of a *Recipient*:

S Verb O		The Y (Recipient)
Той донесе стол	*На*	Директора
He brought a chair	*To*	*the director*

In the example above, Object is not marked with an article. Such sentences always have the meaning S-(V)-O-R (three arguments: agent, theme, and recipient).

When the Object is marked with an article, the sentence becomes:

Subject Verb *the X (*Object) *на* *the Y* .

and its second part fits the Genitive construction the X на the Y . In result, the available grammatical rules of the language assign to the noun Y two possible roles:

1. Subject Verb *the* X (Object) *to* *the Y (*Recipient). S-(V)-O-R, Resipient

2. Subject Verb *the* X (Object) *of* *the* Y (Possessor). S-(V)-O-of-P, Possessor

In such sentences, preposition *на* indicates that the noun that follows it is either Recipient (argument), or it is the object's owner/ Possessor. The difference between these two interpretations is crucial, as the basic syntactic structure of two sentences is completely different - in the former, there are three arguments, and in the latter, there are two (corresponding respectively to the trees on fig. 3.d and 3.c). In Bulgarian, all the sentences of type:

Subject Verb *the* Object *на* *the* Y.

are ambiguous: they assign two different meanings to Y - *Recipient* and *Possessor*.

In normal listening or reading-comprehension conditions, native Bulgarian speakers interpret one of these meanings depending on the context. The sentence "Mary gave the book на the boy." in the context "Mary entered holding a book and she saw a boy" is interpreted as "Mary gave the book to the boy." And, in the context "The boy left his book. Mary was asked about the book." the very same sentence is interpreted as "Mary gave the boy's book to someone else." Speakers of Bulgarian are never mistaken about the conveyed meaning. However, as our experiment has shown, they are not even aware of the existence of the two meanings. It appears that in the cognitive space such "*на*-sentence" acts as a Necker Cube – one may "see it" in either of the two ways. The context makes one of the meanings explicit, while the subjects are not aware of the other meaning. And, in fact, as is the case with Necker's Cube, if one concentrates long enough on an isolated *на*–sentence, one will discern that it has two meanings.

Our goal is to study the mechanisms of mental computation of the syntactic structure of an isolated sentence, with regard to the role of the verb and the arguments.

1. If the assumption is correct that the argument-centered computation is the key to mental operations, an isolated *на*-sentence will be constructed by assigning to Y the role of Recipient.
2. The *на*-sentences are ambiguous; if the role of *entities* (nouns in this case) is primary, semantic relations between their images in the conceptual nets will influence the final result of the syntactic computation.

Experiment

In what ways an isolated *на*–sentence is interpreted? We prepared 13 examples of *на*-sentences (Table 1). Each of these sentences has an argument that conveys either of the two meanings – Recipient (Rc) vs. Possessor (Ps). All the verbs used in the test examples are transitive and allow Recipient. All the sentences can exist as complete sentences without Possessor and without Recipient. The verbs are in the past tense, Perfective form.

Table 1.

200.Ex	Иван	Продаде	Къщата	На	баща си
	Ivan	*Sold*	*The house*	*to/of*	*his father*
201.Ex	Мария	Продаде	Колата	На	Съседката
	Mary	*Sold*	*the car*	*to/of*	*the neighbour*
202.Ex	Михаил	Продаде	Къщата	На	съседа си
	Mihail	*Sold*	*The house*	*to/of*	*his neighbour*
203.Ex	Елена	Продаде	Къщата	На	Кучето

	Elena	Sold	The house	to/of	the dog
204.Ex	Анна Anna	Продаде Sold	Ябълките The apples	На to/of	Момчето the boy
211.Ex	Анна Anna	Подаде Gave	Стола the chair	На to/of	Директора the director
212.Ex	Петър Peter	Донесе Brought	Стола the chair	На to/of	Директора the director
220.Ex	Мария Mary	Показа Showed	Колата the car	На to/of	Съседката the neighbour
221.Ex	Иван Ivan	Показа Showed	Пътеката the wolk	На to/of	Баща си his father
222.Ex	Петър Peter	Показа Showed	Къщата The house	На to/of	Баща си his father
231.Ex	Кумчо Вълчо The Big Bad Wolf	Продаде Sold	Къщата The house	На to/of	Кучето the dog
232.Ex	монтьорът The fitter	Показа Showed	Колата The car	На to/of	Съседката the neighbour

We need to find out which of the two meanings of these isolated sentences is obtained FIRST, i.e. in the most natural way. That can provide information about the mechanisms of mental computation of the basic syntax.

The difficulty in designing an efficient experiment is that when asked to explain the meaning of such a sentence, subjects usually reply by repeating the very same sentence. For them, in the first moment, the sentence has only one meaning that can be put into words in one particular way only. The subjects do exactly what they were asked to do: they express the meaning by using words. Further efforts to make them reveal the meaning make them focus on the sentence for a longer period of time. As a result, they discover that the sentence has one more meaning, and they report that the sentence can mean two different things.

This difficulty was overcome in a tricky way. We used the fact that sentence structure, including word order, is exactly the same in French. The crucial difference is that the preposition *на* is translated in French as “à” (to) for the Recipient-meaning and as “de” (of) for the Possessor-meaning.

The subjects of our experiment were the students in the masters program of the Francophone Institute for Management in Sofia, all of them fluent speakers of French. The subjects, 62 students with different backgrounds (economists, sociologists, biologists, linguists, engineers etc.), were: native speakers of Bulgarian - 39, of Ukrainian - 6, of Rumanian – 5, of Russian - 3, of Georgian – 3, of Albanian – 3, of Macedonian – 2, and of Arabic – 1. Some of the non-native Bulgarians spoke Bulgarian fluently, some were less fluent.

The statements in Bulgarian were presented in a written form to the subjects, on small separate pieces of paper, with the only instruction “Translate into French”. It was done at the end of regular classes, under circumstances implying that “it is not something you should worry about, do it speedily”.

Each statement was presented to 10-12 different subjects. Each subject was given 2 different statements in a random manner, while the statements did not contain the same verb or the same noun. The 23 non-native Bulgarian speakers could ask the experimenter about the meaning of Bulgarian words. There were a few questions about the meaning of “монтьор” (fitter), “тапицер” (upholster) and “пътека” (path) as well as about the corresponding French-tense of the verbs (Past-perfect forms are translated with “passé composé”). There were no questions about the meaning of *на*.

The 124 written translations of the test statements were stored in a database. Table 2 contains the proportion of the Recipient- and Possessor-meanings assigned to each statement (Of% and To%).

Table 2.

	Subject	Verb	Object	на	Y	Of%	to%	Tendency
204.Ex	Anna	Sold	The apples	To/of	the boy		**100**	Y = Recipient
202.Ex	Mihail	Sold	The house	To/of	his neighbor	29	71	Y -> Recipient
201.Ex	Mary	Sold	The car	To/of	the neighbor	30	70	Y -> Recipient
231.Ex	The Big Bad Wolf	Sold	The house	To/of	the dog	33	67	Y -> Recipient
200.Ex	Ivan	Sold	The house	To/of	his father	67	33	Y -> Possessor
203.Ex	Elena	Sold	The house	To/of	the dog	**100**		Y = Possessor
221.Ex	Ivan	showed	The path	To/of	his father		**100**	Y = Recipient
220.Ex	Mary	showed	The car	To/of	the neighbor	11	89	Y = Recipient
222.Ex	Peter	showed	The house	To/of	his father	33	67	Y -> Recipient
232.Ex	The fitter	showed	The car	To/of	the neighbor	50	50	Equivalence
211.Ex	Anna	gave	The chair	To/of	the director		100	Y = Recipient
212.Ex	Peter	brought	The chair	To/of	the director	13	88	Y = Recipient
233.Ex	The upholster	brought	The chair	To/of	the director	50	50	Equivalence

This experimental design was successful in the sense that only 4 subjects, native Bulgarian speakers, became aware that a given sentence has 2 meanings. It is interesting that some of these subjects noticed the double meaning of one of the statements that they had to translate, but not of the other. They were asked to put down the two possible translations in the order in which the meanings came to their minds, and only the first one was taken into account for further analyses.

The results in Table 2 show that, in spite of the “Necker’s cube property” of each statement, one of its possible meanings is interpreted by the subjects more often than the other. The second observation is that for some statements the preferred interpretation is the Recipient-meaning and for others – the Possessor-meaning. The third observation is that these changes do not depend on the verb. For one and the same verb, the interpretation “switches” from one to the other meaning. For example, as one can see in Table 2, “Sold” appears in statements varying from 100% of Recipient-meaning, to 100 % of Possessor-meaning.

Based on the available experimental data (at least ten trials for each statement from different subjects), we assume that the experiment has captured some major tendencies in the interpretation of the test statements. This experiment allows us to further explore the principles of mental operations underlying interpretation of the basic syntactic argument structure. So far, a linguistic theory that would explain the observed tendencies in obtaining some particular result, “computed” by the subjects, has not been developed. Our experiment has shown that the explanation can be provided by using the argument-oriented model derived in compliance with the principles of efficient computation.

First Analyses of Experimental Results

The experimental results show that the interpretation of the syntactic structure depends on entities (in this case, nouns). The verb itself does not predetermine the type of structure: either S-(V)-O-R (three arguments) or S-(V)-OofY (two arguments). Many of the contemporary linguistic theories mostly consider predicate-based and verb-centered syntactic structures. Actually, if the verb does not allow a recipient, the syntactic structure of the *на*-sentence is calculated as S-(V)-O of Y.

Suppose that mental calculus depends solely on the type of the verb. Then in the cases where the verb allows Rc, на would ALWAYS imply a S-(V)-O-R structure. But that is clearly not the case in the last four examples given in Table 3:

Table 3

Subject	Verb	Object	Ha	Y	of%	To%	Tendency
Anna	Sold	the apples	to/of	the boy		100	Y = Recipient
Ivan	showed	the path	to/of	his father		100	Y = Recipient
Anna	Gave	the chair	to/of	The director		100	Y = Recipient
Mary	showed	the car	to/of	the neighbor	11	89	Y = Recipient
Peter	brought	the chair	to/of	The director	13	88	Y = Recipient
Mihail	Sold	the house	to/of	his neighbor	29	71	Y -> Recipient
Mary	Sold	the car	to/of	the neighbor	30	70	Y -> Recipient
The Big Bad Wolf	Sold	the house	to/of	the dog	33	67	Y -> Recipient
Peter	showed	the house	to/of	his father	33	67	Y -> Recipient
the fitter	showed	the car	to/of	the neighbor	50	50	Equivalence
the upholster	brought	the chair	to/of	The director	50	50	Equivalence
Ivan	Sold	the house	to/of	his father	67	33	Y -> Possessor
Elena	Sold	the house	to/of	the dog	100		Y = Possessor

As it is shown in Table 3, when the verb allows a Recipient, *на* implies preferably, but not necessarily the structure S-(V)-O-R (three arguments). The noun Y selects the Rc role in most cases. If mental operations were not dependent on the calculus which relies on the arguments as primary substances, all the statements of the experiment would be with around 50% interpretation of Y as Rc and 50% - Y as Ps.

We conclude that the argument-centered representation of syntax is the key to syntactic analyses.

The next question is: if the argument S-(V)-O-R structure is calculated first, what are the reasons that lead the calculus to take another route and assign a S-(V)-O of Y structure to a similar sentence? Our assumption is that the sentence is kept in working memory (figure 1.) and that the final "solution" about basic syntactic roles is assigned to all its parts after semantic verification. If that was not true, the word order would be the key factor in the syntactic computation and the observed differences in the interpretation would not appear.

Let us analyze why the statement:

Elena	Sold	The house	*to/of*	The dog.	100% of	Y = Possessor,

is interpreted as having S-(V)-O of Y structure. The reason for that seems very clear: the noun dog is rejected as Rc of "sold". The noun takes upon itself the role of the owner of the house. If this is the right mechanism, it is sufficient to provide "the dog" with the possibility to be the Rc of the house, or to modify a noun: "Elena sold the house to a dog-buyer".

The argument-centered syntactic model attests to the fact that syntactic relations depend on the relations between concepts that exist in the semantic space. In fact, as the experimental results show, it is sufficient to replace the subject noun with the one that can be related to the dog as a buyer in a fairy tale context:

The Big Bad Wolf	Sold	The house	*to/of*	The dog	To 67%	Y -> Recipient

This result indicates that mental calculus takes into consideration not only the meaning of the noun but also the relations between the nouns. Thus:

Ivan	Showed	The path	*To/of*	His father	To	100%	Y = Recipient
Ivan	Sold	The house	*To/of*	His father	67%	Of	Y -> Possessor

The three possible arguments of both sentences correspond to concepts that exclude relations such as “fathers have paths” or “sons sell houses to their fathers”. Note that sentences reveal the relations between all the three of the arguments. The predominant meaning in the semantic space of the second sentence is ‘fathers have houses and sons operate their father’s property’.

These dependencies between the basic concepts expressed as Subject and Object are shown as two pairs of statements below:

Subject	Verb	Object	Ha	Y	of%	To%	Tendency
Mary	Showed	The car	to/of	the neighbor	11	89	Y = Recipient
The fitter	Showed	The car	to/of	the neighbor	50	50	Equivalence
Peter	Brought	the chair	to/of	the director	13	88	Y = Recipient
The upholster	Brought	the chair	to/of	the director	50	50	Equivalence

When Mary shows the car, she shows it TO the neighbor; when the fitter shows the car, there is a high probability that this is the neighbor’s car. In the semantic space, fitters operate on cars, while neighbors have cars. That same tendency is observed in the second in pair (upholsters and a director’s chair). Once again, argument structure is influenced by the inter-conceptual relations.

These examples provide evidence about the nature of the primary elements - participants in mental operations. It becomes clear that syntactic computation depends on the meaning of the nouns and inter-conceptual relations.

Conclusions and Future Work

Assumptions about how the argument structure is computed have led to the development of the argument-based model of basic syntax. We applied the methods of cognitive, information, and mathematical modeling, and linguistic theory. An experiment designed to test our ideas confirmed that the argument-centered model is the key to mental operations. The semantic role of entities (nouns) is primary in syntax. The semantic relations between the nouns’ concept-images in the conceptual nets influence the final result of syntactic computation.

The role of the noun has proven to be primary from the point of view of evolution, language acquisition, and other factors of major importance for language. The proposal that arguments (nouns) play the key role in syntax has been supported by experimental evidence. Further study requires a more precise picture of the dependencies between semantic primitives, lexical items, and syntactic rules. That will lead to an advanced modeling of the phenomenon under examination.

Bibliography

[Carnie, 2006] Carnie, Andrew, (2006) Syntax: A Generative Introduction. Blackwell.

[Chomsky, 2004] The Generative Enterprise Revisited. Mouton de Gruyter.

[Chomsky, 2006] Noam Chomsky. Biolinguistic Explorations: Design, Development, Evolution, 2006, West Hall, Bathish Auditorium, AUB.

[Hauser et al., 2002] M. Hauser, N. Chomsky and W.T. Fitch. The Faculty of Language: What is it, who has it, and how did it evolve? In: Science Vol. 298.

[Slavova, 2004] Slavova, Velina (2004) A generalized net for natural language comprehension. In: Advanced Studies in Contemporary Mathematics, vol 8, Ku-Duk Press, 131-153.

[Slavova, Soschen, 2007] Slavova, V., and A. Soschen (2007), A Fibonacci-tree model of cognitive processes underlying language faculty, in: Proc. Of 3-rd international conference in Computer Science, NBU, University of Fulda, Boston University, pp. 196-205.

[Slavova, Soschen, Immes, 2005] Slavova, V., Soschen A. and L. Immes, (2005), Information processing in a cognitive model of NLP, in: International Journal Information theories & applications, vol 12, N3, pp 157 - 166

[Soschen, 2005] Soschen, Alona. 2005. Derivation by phase: Russian Applicatives. Canadian Linguistic Association Conference proceedings.

[Soschen, 2006] Soschen, Alona (2006). Natural Law and the Dynamics of Syntax (MP). Linguistics in Potsdam 25. Optimality Theory and Minimalism: a Possible Convergence? Hans Broekhius and Ralf Vogel (eds.): ZAS, Berlin.

[Soschen, 2008] Soschen, Alona (2008). On the Nature of Syntax. To appear in Bio-linguistics Journal, V.2/2.

[Soschen, Slavova, 2007] Soschen A., and V. Slavova (2007), Cognitive modeling of recursive mechanisms in syntactic processing. In: proc. of the IX international conference in Cognitive Modeling n Linguistics, Text processing and cognitive linguistics, pp 334-343

Authors' Information

Velina Slavova *- New Bulgarian University, Department of Computer Science, 21 Montevideo str., 1618 Sofia, Bulgaria, e-mail: vslavova@nbu.bg*

Alona Soschen *- Massachusetts Institute of Technology, MIT Department of Linguistics and Philosophy, 77 Massachusetts Ave. 32-D808Cambridge, MA 02139-4307, USA, e-mail: soschen@mit.edu*

PERSONALIZED QUESTION-ANSWERING MOBILE SYSTEM

Lee Johnston, Vladimir Lovitskii, Ian Price, Michael Thrasher, David Traynor

Abstract: *Mobile messaging is an integral and vital part of the mobile industry and contributes significantly to worldwide total mobile service revenues. In today's competitive world, differentiation is a significant factor in the success of the business communication. SMS (Short Message Service) provides a powerful vehicle for service differentiation. What is missing, however, is the availability of personalized SMS messages. In particular, the exploitation of user profile information allows a selection and content delivery that meets preferences and interests for the individual. Personalization of mobile messages is important in today's service-oriented society, and has proven to be crucial for the acceptance of services provided by the mobile telecommunication networks. In this paper we focus on user profile description and the mechanism for delivering the relevant information to the mobile user in accordance with his/her profile.*

Keywords: *mobile text messages, user profile, personalization, question-answering system*

ACM Classification Keywords: *I.2 Artificial intelligence: I.2.7 Natural Language Processing: Text analysis.*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

This paper represents results of our further research in the text data mining and the natural language processing areas [1-6] restricted by mobile phone text-based SMS messaging. SMS actually accounts for approximately 75% to 80% of non-voice service revenues worldwide [7]. Last year we represented the Question-Answering Mobile ENgine (QAMEN) [6], which is able to support now its mobile users with personalized situation-aware services. Moreover, QAMEN frees users to have an expensive mobile phone with a web browser. Internet connections from mobile devices remain expensive.

Let us distinguish four different types of Mobile Message (MM):

1. **Person↔QAMEN MM** (MM_{PS}) wherein QAMEN receives user's search query and immediately sends back a text message with the carefully selected result. User's Profile (UP) might be involved to meet the user's demands for searching.
2. **QAMEN (UP)→Person MM** (MM_{SUP}) when user describes in User's Profile what kind of information he/she wants to receive what kind of events need to be taken into account to generate the MM_{SUP}, and when MM_{SUP} should be sent to user. QAMEN, in accordance with those descriptions, generates replies and sends them to user. For example, user wants to know "*the weather in Doncaster on the day of the horse races*".
3. **External MM** (MM_E) when MM_E is sent by some external organisation e.g. "*dental appointment reminder*".
4. **Person-to-Person MM** (MM_{PP}) is ordinary MM when one person sends MM to another person and QAMEN is not involved.

Only MM_{PS} and MM_{SUP} will be considered in this paper.

The success of using MM (MM without index means MM_{PP} and MM_E) is clearly described by Metcalfe's Law [8] – *"The usefulness, or utility of a network equals the square of the number of users"* i.e. put simply, the more users on a network, the more useful and successful it is. This is clearly demonstrated by the success of national SMS interworking – national SMS traffic grew nearly eight times in nine months once the four UK networks were fully interconnected [9].

Mobile question answering differs from standard information retrieval methods. First, it needs to retrieve specific fact information rather than whole documents. Secon, it should select among the found facts the shortest and appropriate fact to meet the 160 characters requirement. In short what a user really wants is a precise answer to a question. For instance, given the question *"When Alexander Pushkin was born?"*, a user wants to get the answer *"In 1799"*, but not to read through lots of documents that content the words *"Alexander"*, *"Pushkin"* and *"born"*. QAMEN takes MM$_{PS}$ as input, classifies it, transforms it into enquiry taking into account UP and current events. When a set of relevant facts is retrieved, the QAMEN extracts from them the most appropriate one and sends it to user's mobile. Search technologies of QAMEN are evolving to provide users with appropriate results despite of unstructured web content. The reasons for web content data remaining unstructured are:

- Data comes from multiple unstructured repositories (file servers, document management systems, intranet sites, internet sites, etc.).
- Data in unstructured documents is of widely varying quality.
- Different types of unstructured data vary greatly from area to area.

That is why processing of personalized MM$_{PS}$ has to take into account **Who** uses MM$_{PS}$ and in **What area**.

Who uses MM and How often?

According to a recent BBC report, SMS has taken the lead as the most popular function for a mobile phone amongst young people. Some **80%** of people under **25** would rather send an MM than make a call, but the number reduces to **14%** among those aged **55 years** and above. When considering gendered differences, the data shows that while 36% of the men reported daily use, more than 40% of the women said that they send MM on a daily basis. The mean number of words per message for men was 5.54. By contrast, the mean number was 6.95 words per MM for women.. Using abbreviations in their MM text-messages: F = 89%; M = 57%.

MM survey

A survey was undertaken by SMS text-messaging company, KAPOW [10]. A summary of survey findings are presented below:

- **How many MM do you receive per day?** (**a**) None – 9%; (**b**) 1-5 – 59%; (**c**) 5-10 – 17%; (**d**) +10 – 15%.
- **Have you ever received MM from the following?** (**a**) Mobile-phone operators – 45%; (**b**) Mobile-phone resellers – 17%; (**c**) Adult-content providers – 4%; (**d**) Doctors/dentists (for appointments etc) – 3%; (**e**) Banks – 10%; (**f**) Charities – 1%; (**g**) Bars & Clubs – 9%; (**h**) Other – 11%.
- **Has MM helped you to remember a meeting, work commitment or any other appointment?** (**a**) Yes – 65%; (**b**) No– 35%.
- **If you opt to receive sales information how do you prefer to receive it?** (**a**) via phone call – 3%; (**b**) via e-mail – 62%; (**c**) via MM – **16%;** (**d**) via instant messaging – 1%; (**e**) via post – 18%.
- **For which service would receiving MM be most useful?** (**a**) Football scores – 14% (**b**) Confirmation of appointments – 28% (**c**) Entertainment services such as ringtones and logos etc – 4% (**d**) Bank account balances – 18% (**e**) Insurance quotes/confirmation – 4% (**f**) Bar and club promotions – 6% (**g**) Travel information – 19% (**h**) Other 7%.
- **Do you agree that MM will become a much bigger part of our working and domestic lives over the coming years?**(**a**) Yes – 87% (**b**) No– 13%.
- **84%** of users expect a MM$_{D}$ response in **five minutes**.

In What area is MM used?

MM is being used in increasingly sophisticated ways, and is fast becoming a huge money earner for operators as well as a tool for businesses. Growth in the MM market is directed towards the area of value-added MM services.

These range from downloads of simple ring tones to news and sports updates. MM is increasingly also being used for finance based transactions. Some might say internet businesses could even consider the technology a means to accept micro-payments for content and services. With premium-priced MM customers simply find something they wish to purchase from a website, and then send a text message to a specified number, including a product code, and moments later a reply is received with an access code. Once a code is used for a purchase, via the phone, a charge is debited on a customer's phone bill or - in the case of a pre-paid mobile phone - directly.

MM is a low-cost communication exchange method that is relatively stable. For example, there is an increased use of one-way outbound **alert notifications for crisis** because MM is more secure, it's faster, and it enables users to reach a wide array of citizens and alert them to pending dangers.

MM is an ideal way for advertisers to reach target markets and establish a one-to-one relationship with the consumer, which is every advertiser's ultimate aim. For example, weather application is defined as personalized, localized weather prediction according to user location, personal profile. Weather related advertisement system knows how to match the right add to the right weather where the advertisement is most effective. For example, implementing a decision to start a soft drink campaign when the temperature approaches, 32°C / 90°F according to user location (if the user is close to the beach and experiencing higher levels of effective temp he will enjoy different add in different temperatures). Such approaches would help the advertiser to optimise its advertising campaign.

As for MM$_{\mathbf{PS}}$ **34%** users use MM$_{\mathbf{PS}}$ for ***news and sport***, **25%** - for ***map and location***, **21%** to ***search*** some data, and **20%** for ***checking weather*** [11].

The most usable areas for MM$_{\mathbf{E}}$ are meeting reminder, sales management, work order, system alert, appointment confirmation, job dispatch, workflow management, information update, payment reminder, customer notification, marketing message, stock and fund quotes, travel information, local weather.

Intelligent MM

The MM has quickly become a boon to the business world as well as to consumers, but so far developers are only scratching the surface of its potential business usage. To fully realize the benefits of MM, businesses must integrate it into their business processes, and into their existing IT systems. When MM is used as part of an overall business process that interacts with consumers, for example, then the enterprise has moved from traditional MM (TMM) to intelligent MM (IMM).

IMM may be differentiated from TMM in these ways [12]:

- The service application is typically a rich enterprise application with business process data, compared to "lightweight" application such as queries for TMM.
- The transaction is "pushed" by the service application, compared to the mobile user "pull" method of TMM.
- An IMM is typically interactive between the service application and the user, whereas TMM is typically a one-way action.
- An IMM allows the user to respond to a message with a "one button" response, where TMM requires the keying in of a response message.
- Authentication of the user with the server application is embedded and automatic to IMM, whereas TMM may be based upon the mobile phone number, plus codes that must be entered by the user.

We have some experience of IMM implementation. 2ergo launched of its MultiSend messaging suite, a range of products that will introduce a new level of interaction and engagement between organisations and their target audience [13]. Design and build a scalable MultiSend solution that would be capable of transmitting up to 40 million messages per month (SMS, MMS and Email). Companies that have already signed up for the MultiSend suite include the internationally renowned travel company Thomas Cook, the major UK car rental company, National Car Rental, and the trans-national publishers, Reed Business International. The suite gives organisations the capability to automate many of their regular outbound communications and to engage in one-to-one dialogue with their target audience, not only to encourage rapid responses, but to also conclude many forms of business transactions. For example, appointment and payment reminders, membership and subscription renewals, or marketing campaigns and customer surveys.

The central question to be addressed by this paper, however, is how to provide the response to a personalized user's MM (MM_{PS} **and** MM_{SUP}). Let us underline that in this paper we consider the precise situation when an MM is sent to the QAMEN i.e. to the artificial system, but not to another person (MM_{PP}). It is important to notice because there is a significant difference between MM_{PS}, which is very similar to internet enquiry, and MM_{PP}.

Difference between MM_{PS} and standard text

There are several elements that lead us to think that MM_{PS} is more like speaking than writing. Firstly, MM_{PS} is intended for immediate response. Secondly, as with most spoken language MM_{PS} makes the assumption of informality. In addition, as a rule, MM_{PS} is ungrammatical:

- Dropping '?' at the end of MM_{PS}.
- Not using any punctuation at all.
- Dispensing with the verb e.g. "*2ergo address*" instead of "*What is 2ergo address?*", or "*Where is 2ergo located?*". Specific questions are used as a rule to find out date: "*When Pushkin born*", or place: "*When Pushkin born*".
- Deletion of articles.

In the next sections a definition of UP is given and elements of UP are described.

User Profile

Various, quite different definition of personalization can be found in [15] and [16]. However, throughout this paper the definition from [14] is used:

"Personalization of a service is the ability to allow a user **U** to adapt, or produce, a service **A** to fit user **U**'s particular needs, and that after such personalization, all subsequent service rendering by service **A** towards user **U** is changed accordingly."

Personalization is provided by the user by means of the user profile (UP). A UP is a group of settings that define how QAMEN is set up for a particular user. Simply stated, the UP serves as a bridge between the generic queries from the diverse users and the heterogeneous data. The main task of UP creation is to establish UP structure. There are, as yet, no standards for representing these, because there is no general agreement on what UP should contain. That is why we were free to offer our vision of UP structure. The UP is uniquely identified by a mobile phone number. Its content composes of three parts: (1) a collection of personal data, (2) set of frames representing the user demands, and (3) history activities. The history activities of the user is a crucial feature in order to automate UP improving process i.e. provide self-improved UP. In this paper we focus solely on the description of user demands and the process of the self-improving UP is not therefore considered.

The personal data of the user consists of the following items:

- Mobile No and Password;
- First Name, Last Name, Date of Birth, and Gender.

User's requirement to MM_{PS} and MM_{SUP} is represented by Demand's Frame (DF), i.e. by DF_{PS} and DF_{SUP} accordingly. DF should take into account the fact that different users may expect different answers to the same query and the same user for the same query may expect different answers in different periods of time. That is why the possibility to define the desire date and time in UP becomes relevant. DF is defined by a Type (DFT) and a set of attributes.

The general form of a DF is the following:

DFT<AI>,<VAI><Priority>[Value][City Country]<Web Site><Date and/or Time><Event><SMS>,

where:

- **DFT** is a type of DF and will be described in the next section;
- **AI** stands for area of interest and is represented by the first level of *Areas of Interests* tree (see Figure 1);
- **VAI** is the value that is associated with AI;
- **Priority** has just one meaning ***Default*** and can be used only in one DF$_{\mathbf{PS}}$;
- **[Value]** and **[City Country]** represent the DF slots. They might be predefined during UP creation, or be empty;
- **Web Site** allows the user to define the desirable site for searching;
- **Date** and/or **Time** is used to set require date and/or time only in DF$_{\mathbf{SUP}}$, i.e. user can specify the delivery date and/or for incoming messages;
- **Event** indicates that list of events for current day (see Figure 2) must be involved for both MM$_{\mathbf{PS}}$ and MM$_{\mathbf{SUP}}$ modification;
- **SMS** is used only in DF$_{\mathbf{SUP}}$ and designate that MM$_{\mathbf{SUP}}$ must be created (if **Event** is mentioned), QAMEN should search for reply, and send the found response to user.

The general requirements to UP creation are:

- Only one **Default** DF might be used in DF$_{\mathbf{PS}}$;
- Duplication of DFT is not allowed in DF$_{\mathbf{PS}}$, but there is no any restriction in using the same DFT in DF$_{\mathbf{SUP}}$;
- Empty MM$_{\mathbf{PS}}$ might be used only for **Default** DF$_{\mathbf{PS}}$. For example, if DF$_{\mathbf{PS}}$ is described as:

Weather [Default] [Varna Bulgaria]

it would be enough for user just to send empty MM$_{\mathbf{PS}}$ to receive the proper information about weather in Varna.

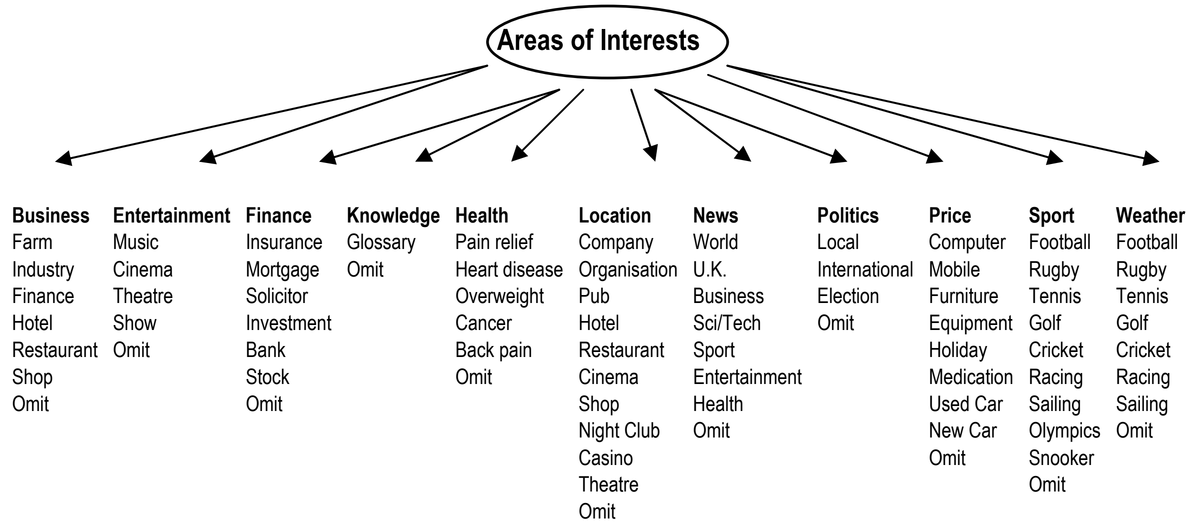


Figure 1. Areas of Interests

The described structure of UP may change in the future but unless some of the requirements were missing it seems the existing choice is simple and flexible enough that no big change should be needed in the future. UP can be easily created and updated via the web-based interface (see Figure 3).

Keywords and Short Code of DFT

QAMEN is an SMS oriented engine which allows the user to enter requests in the shortest form, which **provides the user a better response to their enquiry**. For example, instead of entering the full enquiry: *"I'm looking for pizza hut restaurants in Preston"* it's enough to type in: *"Pizza hut Preston"*, or instead of *"What is the address of 2ergo?"* better to enter the request *"2ergo address"*. Only specific questions ***When*** and ***Where*** should be used e.g. *"When Pushkin was born?"* and *"Where Pushkin was born?"* but not *"Who is Pushkin?"* because you should enter just *"Pushkin"* to received the proper answer.

There are several **Keywords** and **Key symbols** (short code of DFT), which significantly simplify the request presentation. The selection of these keywords and DFT was initiated by areas of interests (see Figure 1). The following keywords and key symbols should be used as DFT:

- By default, i.e. for ANY USER, any request without DFT is considered by QAMEN as a request for searching General Knowledge. Firstly, QAMEN is searching in the Local Knowledge Base (**LKB**), and then, if the result of searching was not success, QAMEN is searching in the Internet. If (for any reason) searching in the LKB need to be omitted DFT **q** should be used e.g. **q** ***British Civil War*** instead of ***British Civil War.***
- **Weather** (or simply **w**), followed by the location. Usually a city name will be enough, but to avoid an ambiguity better to include the country as well e.g. ***weather Plymouth UK*** or (**w** ***Plymouth UK***).
- **Location** (or simply **a**), followed by shop's (or organisation's) name, city and country (just in case to avoid an ambiguity) e.g. **a** ***used car Preston***, or **a** ***HSBC Nice France***, or **a** ***opera London***, or **a** ***NINO's Rawtenstall***) provides an **Address** and/or **Telephone**.
- **News** (or simply **n**) followed by the searchable values e.g. **n** ***Manchester United***, or **n** ***Tony Blair***.
- **Sport** (or simply **s**) followed by the searchable values e.g. **s** ***tennis Sharapova***.
- **Price** (or simply **p**), followed by the product description e.g. **p** ***coffee maker***, or **p** ***Dell XPS***.
- **Finance** (or simply **f**), followed by company's name e.g. **f** ***2ergo plc***.
- To have result of searching in specific file type request should starts with searchable values followed by space, semicolon and file type e.g. ***David Traynor*** **:pdf**.
- To provide searching within the local site request should start with searchable values followed by the www address e.g. ***David Traynor*** ***www.2ergo.com***.
- **Population**, followed by the country e.g. ***population of UK*** (or ***population UK***).
- **Evaluation of Mathematical Expressions** e.g. ***sqrt(34^7/356)*sin(pi/2.3)***.
- **Currency Conversion** e.g. ***10 GBP in Bulgarian money***.
- **Measurement Conversion** e.g. ***61 F in C***, or ***16 stones in kg.***

```
22-28.10.2007<*>DARTS<*>Dublin<*>Skybet World Grand Prix
22-28.10.2007<*>TENNIS<*>Basle<*>Swiss Indoors
22-28.10.2007<*>TENNIS<*>St Petersburg<*>St Petersburg Open
22-28.10.2007<*>TENNIS<*>Lyon<*>Grand Prix
22-28.10.2007<*>TENNIS<*>Linz<*>Generali Women's Open
25-28.10.2007<*>GOLF<*>Majorca<*>Mallorca Classic
26-27.10.2007<*>HORSE RACING<*>Doncaster<*>Trophy meeting
26-28.10.2007<*>DARTS<*>Bridlington<*>World Masters
27.10.2007<*>RUGBY LEAGUE<*>Huddersfield<*>First Test, GB v NZ
27.10.2007<*>HORSE RACING<*>Oceanport<*>Breeders' Cup
```

Figure 2. List of Events for 27.10.2007

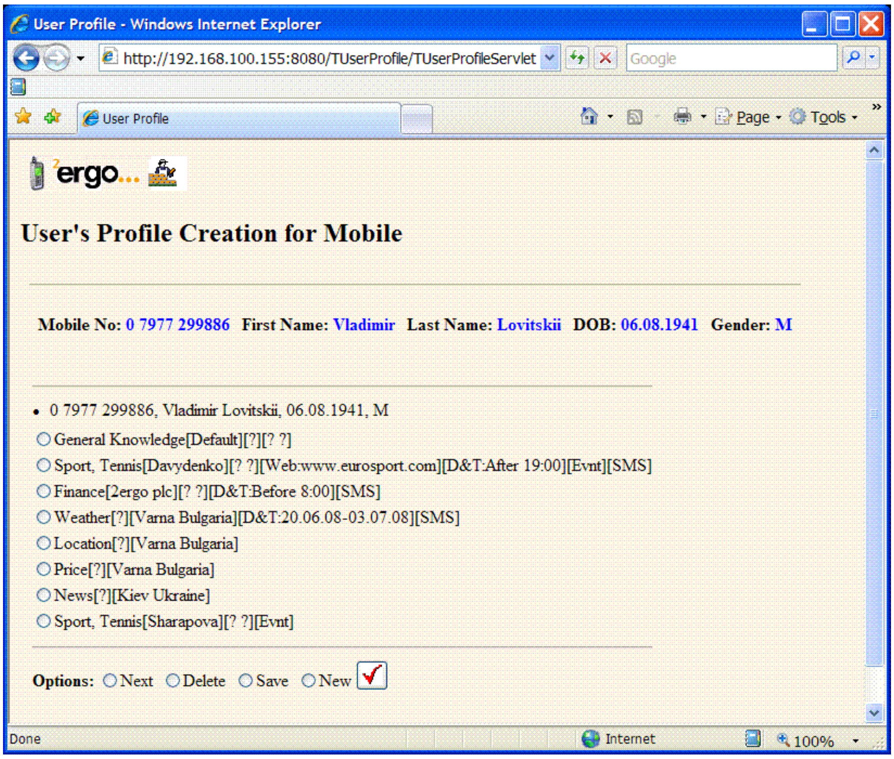
MM_{PS} Processing

The purpose of MM_{PS} processing is to match MM_{PS} against the UP (for explanation UP shown on Figure 3 will be used) and modify MM_{PS} in accordance with the corresponding DF_{PS} to query (Q_{PS}). In the result of searching response (RPS) is produced and is sent to user i.e.

$$MM_{PS} \oplus DF_{PS} \mapsto \{Q_{PS}\} \mapsto \{R_{PS}\} \text{ and } MM_{PS} \varnothing DF_{PS} \mapsto Q_{PS} = MM_{PS} \mapsto R_{PS}$$

where symbol ⊕ means that MM_{PS} match against DF_{PS}, and symbol ∅ has an opposite meaning.

$\{Q_{PS}\}$ and **$\{R_{PS}\}$** designate finite sets of Queries and Response accordingly. **$\{Q_{PS}\}$** might be empty if, on the one hand, DF_{PS} requires to take into account Events, but, on the other hand, at the current day required event does not exist. **$\{R_{PS}\}$** might be empty if in the result of both KB and Internet searching information was not found.


User Profile - Windows Internet Explorer
http://192.168.100.155:8080/TUserProfile/TUserProfileServlet
Google
User Profile
Page
Tools
ergo...
Done
Internet
100%

User's Profile Creation for Mobile

Mobile No: 0 7977 299886 **First Name:** Vladimir **Last Name:** Lovitskii **DOB:** 06.08.1941 **Gender:** M

- 0 7977 299886, Vladimir Lovitskii, 06.08.1941, M

○ General Knowledge[Default][?][? ?]
○ Sport, Tennis[Davydenko][? ?][Web:www.eurosport.com][D&T:After 19:00][Evnt][SMS]
○ Finance[2ergo plc][? ?][D&T:Before 8:00][SMS]
○ Weather[?][Varna Bulgaria][D&T:20.06.08-03.07.08][SMS]
○ Location[?][Varna Bulgaria]
○ Price[?][Varna Bulgaria]
○ News[?][Kiev Ukraine]
○ Sport, Tennis[Sharapova][? ?][Evnt]

Options: ○ Next ○ Delete ○ Save ○ New ✓

Figure 3. Example of User's Profile

A general process for MM_{PS} modification can be explained by means of the following examples:

- MM_{PS}=*"a pizza hut"*. DF_{PS}="Location [?] [Varna Bulgaria]" (see Figure 3). In the result of MM_{PS} parsing QAMEN placed *pizza hut* into value's slot **[?]** i.e. Q_{PS}=***"a pizza hut Varna Bulgaria"***.
- MM_{PS}=*"a HSBC Nice France"*. DF_{PS}="Location [?] [Varna Bulgaria]". In the result of MM_{PS} parsing QAMEN recognised *Nice* as a city and *France* as a country, and replaced the contents of slot **[City Country]** i.e. Q_{PS}=***"a HSBC Nice France"***.
- MM_{PS}=*"p"*. DF_{PS}="Price [?] [Varna Bulgaria]". Value for slot is not given and that is why QAMEN generate the R_{PS}=*"What to you want to buy in Varna Bulgaria?"* and send it to user.
- MM_{PS}=*"s"*. DF_{PS}="Sport, Tennis [Sharapova] [? ?][Evnt]". If there is not any tennis events at the current day then Q_{PS}=**nil**. Suppose, MM_{PS} has been sent at 27.10.2007 (see Figure 2). For this day four different tennis events occurred and therefore for queries have been generated by QAMEN: Q_{PS}={***"Sharapova Basle Swiss Indoors", "Sharapova St Petersburg Open", "Sharapova Lyon Grand Prix", "Sharapova Linz Generali Women's Open"***}.

DF_{SUP} Processing

The main purpose of DF_{SUP} processing is to generate the set of Q_{SUP} in accordance with Date and/or Time, and Events (if given), get responses and send them to users i.e.

$$DF_{PS} \mapsto \{Q_{PS}\} \mapsto \{R_{PS}\}$$

The subset of **$\{R_{PS}\}$** is shown on Figure 4.

MOBILE: 0 7764 446240
SMS = Weather for Wigan UK 11C Mostly Cloudy Wind NE at 10 km/h
Humidity: 82
Temperature:
Thu 10C - 3C
Fri 12C - 8C
Sat 16C - 13C
Sun 15C - 6C

MOBILE: 0 7764 446240
SMS = Broca PLC (BROC). 69.00p Down 1.00p (-1.43%). Market cap: £25.965m

MOBILE: 0 7977 299886
SMS = BASEL Switzerland - David Nalbandian lost to Stanislas Wawrinka in the first round of the Swiss Indoors on Wednesday three days after beating Roger ...

MOBILE: 0 7977 299886
SMS = PETERSBURG Russia - Top-seeded Nikolay Davydenko defeated Filippo Volandri 6-1
6-1 Wednesday to advance to the second round at the Petersburg Open ...

MOBILE: 0 7977 299886
SMS = Top-seeded Andy Roddick was upset by Fabrice Santoro in the first round of the Lyon Grand Prix on Wednesday at Lyon France. Santoro 34 hit three aces in...

MOBILE: 0 7977 299886
SMS = Linz Austria (Sports Network) - US Open semifinalist Anna Chakvetadze was an easy second-round winner Wednesday at the $600000 Generali Ladies Linz...

Figure 4. Result of Responses

Conclusion

In this paper we have proposed a profile-based approach to improve the efficiency of SMS. We turned our attention towards the UP creation and its possible application in a mobile environment. The object of our research is to improve query response by creating UP. Most importantly, the structure of UP and general process of personalization was given. It is important to offer and realize some ideas (not necessarily the best) when there are as yet no standards for representing UP, because there is no general agreement on what these profile should contain. Of course, the ultimate criterion of “good” UP is that a user should be satisfied with search results without the necessity of understanding the structure of UP, MM_{PS} modification, search methods etc.

Bibliography

[1] G.Coles, T.Coles, V.A.Lovitskii, "Natural Interface Language", *Proc. of the VIII-th International Conference on Knowledge-Dialogue-Solution: KDS-99*, Kacivelli (Ukraine), 104 -109, 1999.

[2] T.Coles, V.A.Lovitskii, "Text Searching and Mining", *Journal of Artificial Intelligence, National Academy of Sciences of Ukraine, Vol 3,* 488-496, 2000.

[3] D.Burns, R.Fallon, P.Lewis, V.Lovitskii, S.Owen, "Verbal Dialogue Versus Written Dialogue*", International Journal "Information Theories & Applications", Vol 12(4)*, 369-377, 2005.

[4] Ken Braithwaite, Mark Lishman, Vladimir Lovitskii, David Traynor, "Distinctive Features of Mobile Messages Processing",*International Journal "Information Theories & Applications", Vol 14(2)*, 154-160, 2007.

[5] Guy Francis, Mark Lishman, Vladimir Lovitskii, Michael Thrasher, David Traynor, "Instantaneous Database Access", *International Journal "Information Theories & Applications", Vol 14(2)*, 161-168, 2007.

[6] Vladimir Lovitskii, Michael Thrasher, David Traynor, "Automated Response To Query System", *Proc. of the XIII-th International Conference on Knowledge-Dialogue-Solution: KDS-2007*, Varna (Bulgaria), 534 - 543, 2007.

[7] www.portioreserch.com

[8] Robert Metcalfe: http://en.wikipedia.org/wiki/Metcalfe's_Law.

[9] Jeff Wilson, Chairman, www.telsis.com

[10] www.kapow.co.uk.

[11] Wireless World Forum: http://de.w2forum.com/i/.

[12] Jukka Salonen, BookIT Oy: www.bookit.fi

[13] www.2ergo.com (MultiSend™)

[14] Blom, J., "Personalization – A Taxonomy", *Conference on Human Factors in Computing Systems (CHI).* Hague, Netherlands, 1-6 April 2000. ISBN:1-58113-248-4.

[15] Lankhorst, M.M., Kranenburg, van H., Salden, A., Peddemors, A.J.H., "Enabling Technology for Personalizing Mobile Services", *Proc. of the 35th Hawaii International Conference on System Science, 2002.*

[16] Jorstad, I., van Do, T., Dustdar, S., "Personalisation of Future Mobile Services", *9th International Conference on Intelligence in Service Delivery Networks*. Bordeaux, France, 18-21 October 2004.

Authors' Information

***Lee Johnston** – 2 Ergo Ltd, St. Mary's Chambers, Haslingden Road, Rawtenstall, Lancashire, BB4 6QX, UK, e-mail: lee.johnston@2ergo.com*

***Vladimir Lovitskii** – 2 Ergo Ltd, St. Mary's Chambers, Haslingden Road, Rawtenstall, Lancashire, BB4 6QX, UK, e-mail: vladimir@2ergo.com*

***Ian Price** – Broca Communications Ltd, St. Mary's Chambers, Haslingden Road, Rawtenstall, Lancashire, BB4 6QX, UK, e-mail: ian.price@brocaplc.com*

***Michael Thrasher** – University of Plymouth, Plymouth, Devon, PL4 6DX, UK, e-mail: mthrasher@plymouth.ac.uk*

***David Traynor** – 2 Ergo Ltd, St. Mary's Chambers, Haslingden Road, Rawtenstall, Lancashire, BB4 6QX, UK, e-mail: david.traynor@2ergo.com*

METHODOLOGY FOR LANGUAGE ANALYSIS AND GENERATION IN CLOSED DOMAINS: PHARMACEUTICAL LEAFLET

Jesús Cardeñosa, Carolina Gallardo, Adriana Toni

***Abstract**: The best results in the application of computer science systems to automatic translation are obtained in word processing when texts pertain to specific thematic areas, with structures well defined and a concise and limited lexicon. In this article we present a plan of systematic work for the analysis and generation of language applied to the field of pharmaceutical leaflet, a type of document characterized by format rigidity and precision in the use of lexicon. We propose a solution based in the use of one interlingua as language pivot between source and target languages; we are considering Spanish and Arab languages in this case of application.*

***Keywords**: Controlled languages; Interlingua-based Machine Translation*

***ACM Classification Keywords**: I.2.7 Natural Language Processing*

***Conference**: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008*

Introduction

Most of translated publications belong to a technical, commercial or management field. This discussion only ratifies the challenge supposed for automating, as far as possible, this kind of document translation, because due to its nature, they only require translations that could be described as mechanic or routine. Literary texts, or generally the natural language, escape to the efforts of computer treatment because it is difficult to integrate the contextual knowledge that the speaker has, which is the only one that in many cases can solve the ambiguity that cannot be solved at a syntactic or semantic level. The quality of automatic translation remarkably improves when the advantages offered by specialty languages can be taken, regarding its precision, possibility of standardization, and vocabulary limitation. The automation of the translation requires, in this case, a laborious initial linguistic analysis of a corpus (texts of the domain that by its number and variety may be representative), but this initial effort is compensated by the time saved if the post-edition of produced documents is not needed.

Within the text translation pertaining to specific thematic fields (technical manuals, weather forecast, reports, legal texts, etc.), we must still distinguish between free or fixed format texts, understanding by fixed format a document structure divided into sections, with specific headings whose type content is known. The biggest structuring of the document is again an advantage when automating the translation, because it reduces the possible ambiguity of the terms to translate since there is information about the context where they appear.

The approach to the Automatic Translation (AT from now on) by means of the use of interlinguas consists of using an intermediate representation of the contents to translate, independent of source and target languages, from which the text is generated. One of the greater difficulties of this process lays in the definition of an interlingua that can work as an intermediate representation between any of the two languages. In fact, it is a new language requiring a definition of all of its components, with the additional challenge that being an artificial language it has to be as expressive as the natural languages. None of these systems have been successfully developed in text translation regarding domains with opened formats due to the difficulty in the interlingua design.

An advantage of these systems is the possibility of incorporating new languages without affecting the modules already developed for the other languages – it is necessary an encode module and a decode module (*to* and *from* the interlingua respectively) for each language. Figure 1 shows all the necessary modules to cover all possible pairs among A, B, and C languages.

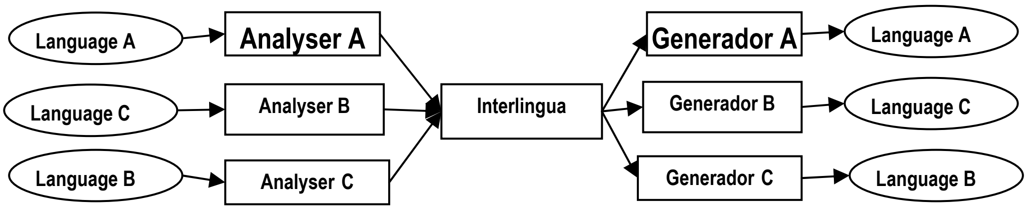


Figure 1. Modules for analysis and generation among 3 languages in an interlingua-base MT system

In this paper we are presenting a methodology devised to approach automation of analysis and language generation in a well defined and limited domain —pharmaceutical leaflet, with Spanish as source language and Arab as target language.

A pharmaceutical leaflet constitutes a text example pertaining to a closed domain with a fixed and standardized format. We are proposing the use of translation based in interlingua, because it is the only one guaranteeing a very precise coherence between the different language versions: the *Universal Networking Language* (UNL) a computing language developed as an essential element of UNL Project, an international project, promoted by the Institute of Advanced Studies of the United Nations (UNU/IAS).

It has software programs that allow introducing Spanish contents to UNL and from UNL to Arab. They work in interaction with linguistic resources typical of each one of the languages, stored in an electronic format: Grammar rules for language analysis and Spanish/UNL dictionary in the Spanish case; and Grammar rules for the language generation and UNL/Arab dictionary in the Arab case. The final mission is to adapt rules and dictionaries to the leaflet domain, so that the quality level of the produced translations may be acceptable, not requiring a post-edition process. This project is in phase of accomplishment, and the work plan — that we will expose in section 4 of this article—, has been born from the study of the difficulties posed by the dominion and characteristics of the source and target chosen languages.

Section 2 introduces UNL language and generally the translation process by means of an interlingua. In section 3 we give a brief explanation of software tools referred in the work plan. The most relevant contribution of this article, however, goes beyond the interlingua specifically chosen or the tools. The work plan could be carried on using other interlingua or other tools, because basically their functionality would have been the same one.

UNL Project: Aims and Components. Description of Translation Process

The UNL Project is born with the aspiration of developing an interlingua system to support a multitude of languages without any domain or lexical type restriction. It is an international project developed by the Institute of Advanced Studies of the United Nations together with research groups throughout the world, such as the Spanish Language Centre (CLE, `http://www.unl.fi.upm.es`), to which the authors of this paper belong. Their main objective is the dissemination, promotion, and formation in UNL technology with the aim to eliminate linguistic barriers in the Internet.

The UNL is composed of three main elements: Universal Words, Relations and Attributes. Formally, a UNL expression can be viewed as a semantic net, whose nodes are the Universal Words, linked by arcs labelled with the UNL relations. Universal Words are modified by the so-called attributes. UWs are a key element for UNL. A UW is intended to express a concept found in any natural language. To do that, UNL uses words and phrases taken from English but these English words are modified by semantic restrictions in order to eliminate ambiguity present in the vocabulary of natural languages. Thus, UWs are linked to the vocabulary of natural languages. The reason for choosing English is just practical: the inventory of the English vocabulary is rather well covered by many authoritative dictionaries and there are bilingual dictionaries of English to almost any other natural language.

Next we present an overview of the translation process by means of the UNL. For more detailed information see the references suggested in the bibliography (section 5).

The Analysis Process or Enconversion

This process consists in putting contents from a natural language to the UNL. The essential resources in this phase of the process are:

- A set of Grammar rules linked to the source language to transform the contents written in that language into UNL. Each language requires the development of a specific set of rules.
- A dictionary where each word of the source language having a semantic meaning is linked to a UW — again each language requires the development of such dictionary. The following pairs illustrate how a different UW is linked to each sense of the word "state" in order to disambiguate its meaning:

  Pair 1: <estado, state(icl>administrative_district>thing, equ>country)> denoting the territory occupied by a nation.

  Pair 2: <estado, state(icl>political_unit>thing)> denoting a politically organized body of people under a single government.
- The *Enconverter* is the component – a *parser* - that allows automatically passing the content from a source language to UNL. It interacts with the set of Grammar rules and the dictionary already mentioned in the previous section.

The UNL produced texts come to be part of the "Base Documental UNL", containing documents written in UNL language which are available to the user community of UNL System.

The Generation Process or Deconversion

This process consists in putting contents from UNL to a target language. The elements taking part in the process are:

- A set of Grammar rules for the generation, beginning from the UNL linked to a target language. Each language requires the development of its specific set of rules.
- A dictionary that links each UNL word to a word in the target language (same dictionary used in the process of *Analysis*).
- The *Deconverter* is the "opposite" component to the *Enconverter.* It is responsible from passing the UNL to a natural language. The set of programs forming the *Deconverter* works with the rules and dictionary already mentioned in the previous section.

For the Spanish language there are several thousand rules for the *Generation* process in Spanish language, that provide a rather acceptable cover up for the Spanish generation from contents written in UNL.

The Research Centres and Universities working together with the UNL/IAS Centre in the UNL Project are responsible for the development and permanent updating of the resources own to their respective languages — Grammar rules for the analysis and generation, and dictionaries — besides supervising the correct application of the UNL standards, maintaining the language servers for the testing tools, giving technical support to the users, content providers and builders, and in general to carry on all the necessary activities to promote the UNL system.

Software Tools of the Spanish Language Centre

The Spanish Language Centre represents the UNL Center for the Spanish language, and must help it in the programming, coordination, support, financing, research, and formation of the whole UNL System. It includes all the languages whose roots are related to the Spanish language, namely, all the Spanish-speaking countries and developments affecting the indigenous languages of Latin America. Most of the CLE members are also researchers of the Grupo de Validación y Aplicaciones Industriales (VAI, http://www.vai.dia.fi.upm.es) of the Facultad Informática, Universidad Politécnica de Madrid.

Next, the functionality of the software tools, already mentioned in our work plan, is briefly explained. The tools have been developed in the VAI laboratory.

Generator of Universal Words: it is used in the *Analysis* phase in particular in the identification and extraction of the words from the document in Spanish language that have a semantic content, and the construction of the corresponding universal word.

UNL Editor*:* it is a platform that integrates the components used in the *Analysis, Verification,* and *Generation* phase, allowing carrying on these processes in a comfortable and unified way from a unique program. Basically, the environment consists of a central module (UNL Editor) that controls and coordinates the rest of the system components; see Enconverter, Deconverter, rules and dictionaries. The functionalities of this editor are:

- Manual and automatic analysis of documents in a given language
- Text edition and graphic structures
- Validation and verification of UNL code
- Generation of a code in any language from UNL
- Access to word dictionaries

Work Plan: Pharmaceutical leaflets

Next, it is presented a list of tasks whose fulfillment will provide us the resources (programs and linguistic resources in an electronic format) needed to automate the translation of the pharmaceutical leaflets from Spanish to Arab.

Among the different tasks needed to approach, we emphasize two as the most important:

- **Exhaustive study of linguistic characteristics, characterizing the leaflets.** A deeper study will result in a higher lucidity in conclusions, more quality in translations and greater automation of translating process, since we will be able to anticipate every difficulty that may arise. This study will be carried on from a very numerous and varied number of leaflets —that we shall compile— in order to obtain general conclusions
- **Adaptation of linguistic resources managed by the programs in charge of translation, to the outcomes of the study.** We are considering sets of syntactic and morphologic rules, linked to Spanish and Arab languages, that interacting with the parsers and generators we have, will determine the translation from Spanish to UNL and from UNL to Arab respectively, and also bilingual dictionaries Spanish/UNL, and UNL/Arab.

A first examination of a reduced number of pharmaceutical leaflets in Spanish language shows us that there are several variations in their format, regarding their section division as well as their extension — reduced or extended versions — being the latter the most widely used at a commercial level. Differences are determined by issues as: type of symptom or disease for which medicine is prescribed, administration form — syrup, pills, etc. — and the manufacturing laboratory. Therefore, before compiling the set of leaflets that will constitute the corpus of the work, we will establish the quantitative criteria — how many will be enough? — and the qualitative criteria that they must fulfill in order to make them representative of the different types of leaflets found in the market. We must as far as possible include all the different groups of medicines regarding their pharmaceutical classification, and within each group, cover all administration forms, and also include medicines from the greater possible number of laboratories.

Once compiling and storage is finished, we will proceed to study the corpus. First, we pay attention to the documents structure, and we find a basic core of sections ("Indications", "Composition", "Administration", etc.) common to all of them. Variations come mainly from the manufacturing laboratory and type of product (symptom or fighting disease). The format study concludes relating each epigraph to the type of content it includes. It is important to identify the vocabulary, sentences etc., that appear linked to each epigraph, because it helps us to disambiguate meanings in the process of translation and to recognize new words (is it a “component”, a “secondary effect”, a medicine name?).

As to the linguistic study, due to the previous knowledge of the difficulties posed in that sense by automatic translation, and by detailed observation of the texts, we are synthesizing the most relevant aspects to be considered:

- **Lexical aspects:** as the degree of semantic ambiguity of the used words; estimation of specific vocabulary ratio compared to common vocabulary; or the guidelines in the words composition by frequently used suffix and prefix**.**
- **Syntactic aspects**: as the existence of syntactic constructions linked to the different sections and contents (type of sentence to express contraindications, or administrative instructions, etc.); and the observation of prepositional ambiguity.

The conclusions obtained in the phase of corpus study will be used for the adaptation of the syntactic and morphologic rules that we have in order to put contents from Spanish to UNL, and from UNL to Arab. Because of the type of vocabulary included in the leaflets, we shall emphasize the following aspects:

- Recognition of unknown words
- Recognition of proper names
- Recognition of morphologic composition processes.

We must also adapt the available Spanish/UNL and UNL/Arab dictionaries, incorporating new vocabulary own of the domain and disambiguating the translations of common words, which appear with certain frequency in the leaflet field. The updating of the dictionaries must be done according to the rules of general dictionaries and UNL specifications.

Passing to expose the work plan with more detail, we distinguished 5 great tasks whose content we will be refining according to the difficulties posed. Basically, these tasks must be successively developed, because the results of each task are used by the following ones. Figure 2 presents the dependencies diagram and priorities of the identified tasks and subtasks. After it, we are including an explanation table of the required objectives and resources in each task.

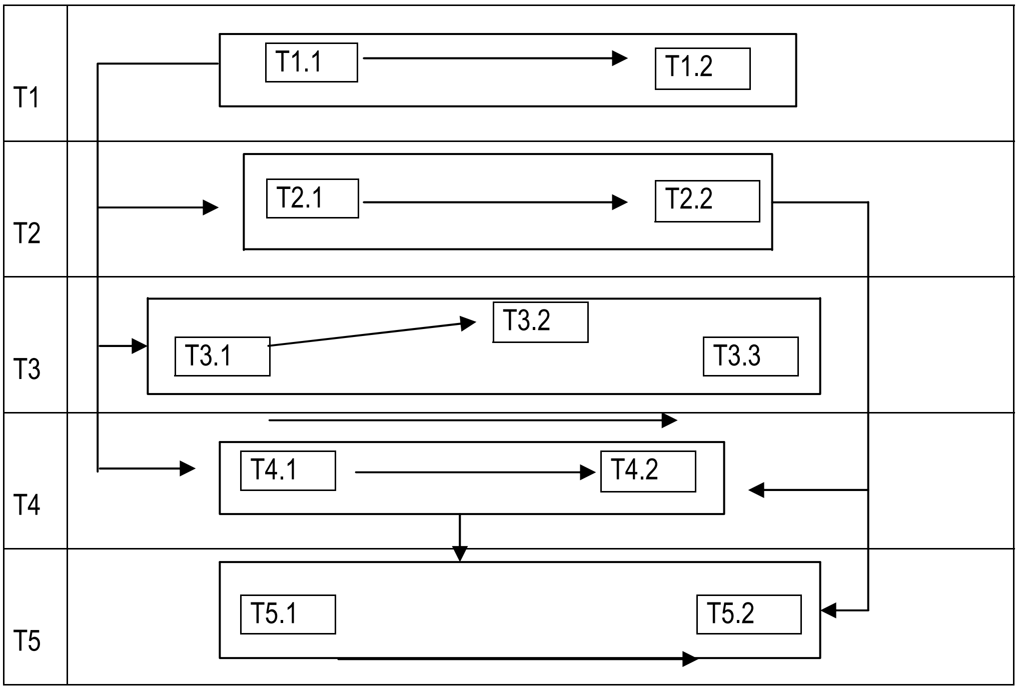


Figure 2. Diagrama de dependencias

Task identifiers stand for:

- T1.1: Criteria for the selection of texts
- T1.2: Obtaining and classification of texts
- T2.1: Analysis of text structure
- T2.2: Linguistic analysis of texts
- T3.1: UWs production
- T3.2: Update of Spanish/UNL dictionary
- T3.3: Update of Arab/UNL dictionary
- T4.1: Review of resources for Spanish language
- T4.2: Codification of texts into UNL
- T5.1: Review of resources for Arabic language
- T5.2: Generation into Arabic language

Detail of the tasks

TASK 1: CORPUS CREATION

Aims: compile a set of leaflets in Spanish language in order to constitute a representative corpus. To do that, it is necessary to establish how many are needed and the qualitative criteria that the chosen leaflets must fulfill (T1.1), and subsequently, proceed to their obtaining, classification, and storage (T2.2).

Resources: Access to leaflets in Spanish language (in electronic format).

TASK 2: CORPUS STUDY

Aims: study of the leaflets structure —section division, their content—- and the international rules and standards that may exist regarding this (T2.1). Also, the linguistic characteristics of the texts as to the lexical employed —polysemy, general or specific vocabulary, suffix, and prefix— also, syntactic construction own to each section, and the possible prepositional ambiguity will be studied (T2.2).

Resources: Access to the corpus of the selected texts.

TASK 3: PRODUCTION OF DICTIONARIES

Aims: To produce dictionaries adapted to the vocabulary employed in the leaflets. The first step (T3.1) is to automatically obtain the pair list [Spanish word/UW] for all those words having a semantic content that appear in the leaflets. The next step (T3.2) is to add the list of pairs to the general Spanish/UNL dictionary, adapting, if needed, the list of attributes describing each word to the domain. Finally (T3.3) a dictionary [UNL/target language] must be produced, from the list of the obtained UWs in T3.1. Again, any existing dictionary of general purpose, if there is one, may be reused, adapting the attributes to the leaflet field.

Resources: Generation Tool of UWs (T3.1), text editors to adapt the text format of the corpus to the one required by the Tool, dictionaries of general purpose pairing UNL with Spanish and Arab languages.

TASK 4: SPANISH ANALYSIS AND UNL ENCODING

Aims: to generate a UNL version of the leaflets in Spanish. In the first place, the base of Spanish analysis rules will be adapted according to the lexical studies and the syntactic phenomena, identified when making the corpus study (T4.1). The enhancements will be focused in the rules adaptation to the unknown words, proper names, processes of morphologic composition, and treatment of syntactic structures identified when making the linguistic study. If it was necessary, ad hoc rules and specific attributes of the domain will be incorporated (the latter will also require the modification of the attributes of the domain dictionary). After reviewing the rules and dictionary, the UNL encoding (T4.2) of the leaflets will be automatically obtained.

Resources: Enconverter, UNL Editor, texts' editors to adapt the texts' format to the corpus required by the Enconverter.

TASK 5: GENERATION OF LEAFLETS IN ARAB

Aims: to generate the Arab version. In the first place (T5.1) the Grammar rules must be adapted for the generation of Arab language and the [UNL/Arab] dictionary with the aim of allowing, as far as possible, producing understandable and correct texts. The last phase of the task and the global process – excluding final tests and evaluations – is to generate the Arab version in the selected leaflets (T5.2). This is a totally automated process, and there will not be any type of post-edition in the generators' outcome.

Resources: Deconverter, UNL Editor, texts' editors to adapt the texts format to the one required by the Deconverter.

We identified an additional task of tests that would consist in assessing all the generated resources along the process. We described in the summary diagram the type of tests that must be carried out and the software tools in order to accomplish them (see section 3 for a summary description of tools functionality).

TASK 6: TESTS AND ASSESSMENT OF RESULTS

Aims: To test and asses the adaptation of all the resources to the awaited results. The UNL Editor will help us to analyze the UNL generated code (T6.1). We will use the inference engine to verify that the corresponding set of rules and dictionaries allow adequately putting of contents (T6.2). Finally it will have to be verified the legibility and correctness of the translated version to the target language from the corpus texts (T6.3).

Resources: Editor UNL, Grammar rules of Spanish analysis, rules of Arab generation and dictionary, set of translated texts to Arab.

Conclusions

This paper describes a methodology of work to approach the translation based in interlingua among different pairs of languages in closed domains. The use of an interlingua allows us, on one hand, the total reuse of the analysis modules and the generation of language already existing before the inclusion of new pairs. On the other hand, to narrow down the problem to closed domains simplifies the analysis and generation tasks, producing results of better quality in comparison with those obtained from opened domains.

This methodology has been applied to the treatment of pharmaceutical leaflets; however, it would be equally valid for the treatment of any type of texts within any domain with a controlled language.

Bibliography

[Boguslavsky et al, 2005]. Boguslavsky, I., Cardeñosa J., Gallardo, C., and Iraola, L. The UNL Initiative: An Overview. Lecture Notes in Computer Science. Volume 3406/2005, pp 377-387. Springer Berlin / Heidelberg: 2005. ISBN 978-3-540-24523-0

[Fellbaum, 1998]. Fellbaum, C., (ed): WordNet: An Electronic Lexical Database. Language, Speech, and Communication Series, MIT Press (1998)

[Uchida et al, 2005] Universal Networking Language (UNL). Specifications Version 2005. Edition 2006. 30 August 2006. http://www.undl.org/unlsys/unl/unl2005-e2006/

Authors' Information

Jesús Cardeñosa – *Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: carde@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es*

Carolina Gallardo – *Group of Validation and Industrial Applications. Escuela Universitaria de Informática. Universidad Politécnica de Madrid. Cta de Valencia Km.7. 28041 Madrid; email: cgallardo@eui.upm.es. http://www.vai.dia.fi.upm.es.*

Adriana Toni *– Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: atoni@fi.upm.es. http://www.vai.dia.fi.upm.es*

COMPUTER-AIDED SYSTEM OF SEMANTIC TEXT ANALYSIS OF A TECHNICAL SPECIFICATION

Alla Zaboleeva-Zotova, Yulia Orlova

***Abstract**: The given work is devoted to development of the computer-aided system of semantic text analysis of a technical specification. The purpose of this work is to increase efficiency of software engineering based on automation of semantic text analysis of a technical specification. In work it is offered and investigated the model of the analysis of the text of the technical project is submitted, the attribute grammar of a technical specification, intended for formalization of limited Russian is constructed with the purpose of analysis of offers of text of a technical specification, style features of the technical project as class of documents are considered, recommendations on preparation of text of a technical specification for the automated processing are formulated. The computer-aided system of semantic text analysis of a technical specification is considered. This system consists of the following subsystems: preliminary text processing, the syntactic and semantic analysis and construction of software models, storage of documents and interface/*

***Keywords**: natural language, semantic text analysis, technical specification.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing*

***Conference**: The paper is selected from International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008*

Introduction

Now designing of the software represents the labor-intensive process demanding of the user deep knowledge of a subject domain and skills in designing.

Most known of the commercial software products used at designing of the software, basically are intended for visualization intermediate and end results of process of designing. Some of them allow fully automating last design stages: generation of a code, creation of the accounting and accompanying documentation, etc. Thus the problem of automation of the initial stage of designing - formations and the analysis of the text of the technical project remains open. It is connected to extraordinary complexity of a problem of synthesis and the analysis of semantics of the technical text for which decision it is necessary to use methods of an artificial intellect, applied linguistics, psychology, etc. However, it is possible to come nearer to achievement of the given purpose, having allocated some small subtasks quite accessible to the decision by known methods of translation.

Proceeding from the aforesaid, it is possible to draw a conclusion, that the problem of creation of means for automation of process of designing is actual [1].

On CAD-department of the Volgograd state technical university questions of automation of designing of software products with use of natural - language support for a number of years are investigated.

The main ideas of the developed direction are:

- realization of the unified procedures of the designing equally answering to requirements of the expert
- design the requirements to technology to modeling of software products.

Designing of the software at the initial stages with use of a natural language is based on the following main principles:

1. Performance of all design procedures is modeled in language of internal representation of system. Internal representation is the unified model of designing of the software, based on methodology of the theory of systems and technologies of natural language processing.

2. A number of representations of the project is generated. Translation of a condition of the project into the certain language which is distinct from language of internal representation refers to as representation. Programming languages, natural languages or artificial formal languages of modeling of processes of designing can be attributed to such languages (UML, IDEF-diagrams, model of diagrams of streams of the data). Different representations reflect only separate aspects of the project.
3. Thus due to use of uniform internal model consistency of representations is provided.
4. The software of process of the designing, guaranteeing an opportunity of conducting the project on any of languages of representations is developed.
5. The basic language of representation of the project for the person - the customer and the designer - is the natural language. Dialogue between the customer and the designer is traditionally conducted in a natural language - language of human dialogue, but, as a rule, are entered new formal structures - diagrams, circuits, schedules. According to the developed concept, natural - language representation of the project supplements formal and serves as the tool facilitating understanding of process of designing.

As illustration of process of designing ON with use of the offered concept the diagram «to be», resulted on figure 1 serves.

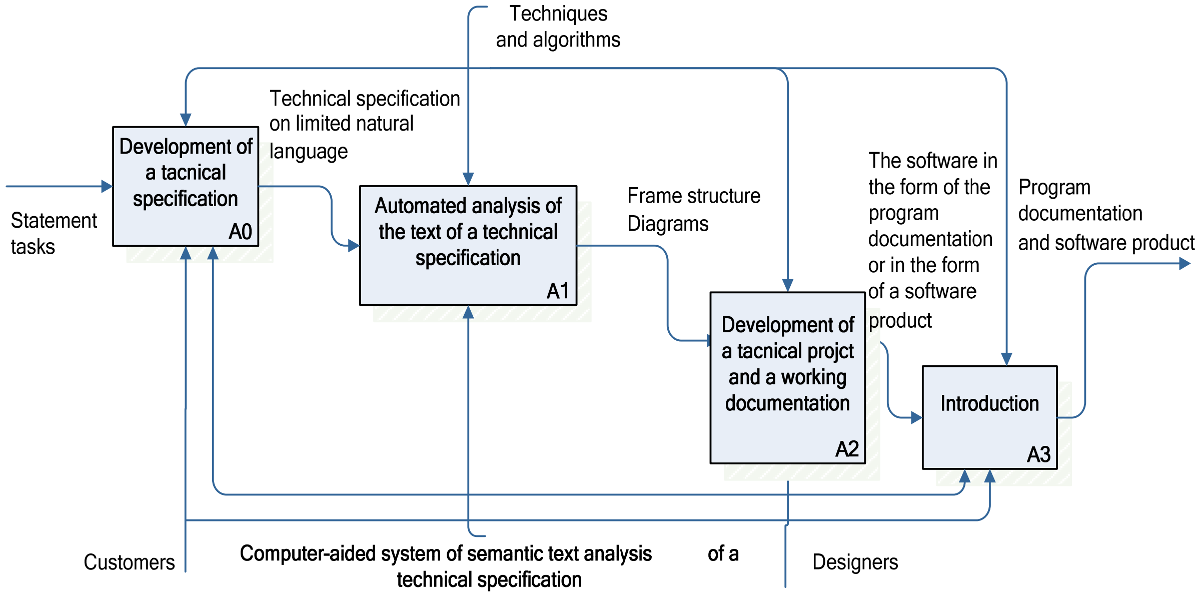


Figure 1: Diagram of process of designing «TO BE»

The given work is devoted to development of the computer-aided system of semantic text analysis of a technical specification.

The purpose of this work is to increase efficiency of software engineering based on automation of semantic text analysis of a technical specification.

To achieve this purpose it is necessary to solve the following tasks:

1. To carry out the analysis of software engineering process and models of semantic text analysis;
2. To develop and investigate model of the text analysis of a technical specification;
3. To develop a technique and analysis algorithms of text of a technical specification and construction of the software models;
4. To develop the computer-aided system of semantic text analysis of a technical specification;
5. To apply the system in software engineering process.

Model of the Text Analysis of a Technical Specification

In work it is offered and investigated the model of the analysis of the text of the technical project is submitted, the attribute grammar of a technical specification, intended for formalization of limited Russian is constructed with the purpose of analysis of offers of text of a technical specification, style features of the technical project as class of documents are considered, recommendations on preparation of text of a technical specification for the automated processing are formulated.

Model input is the requirement specification written in the limited natural language, its output is a set of the data flow diagrams, describing the program system (see Figure 2).

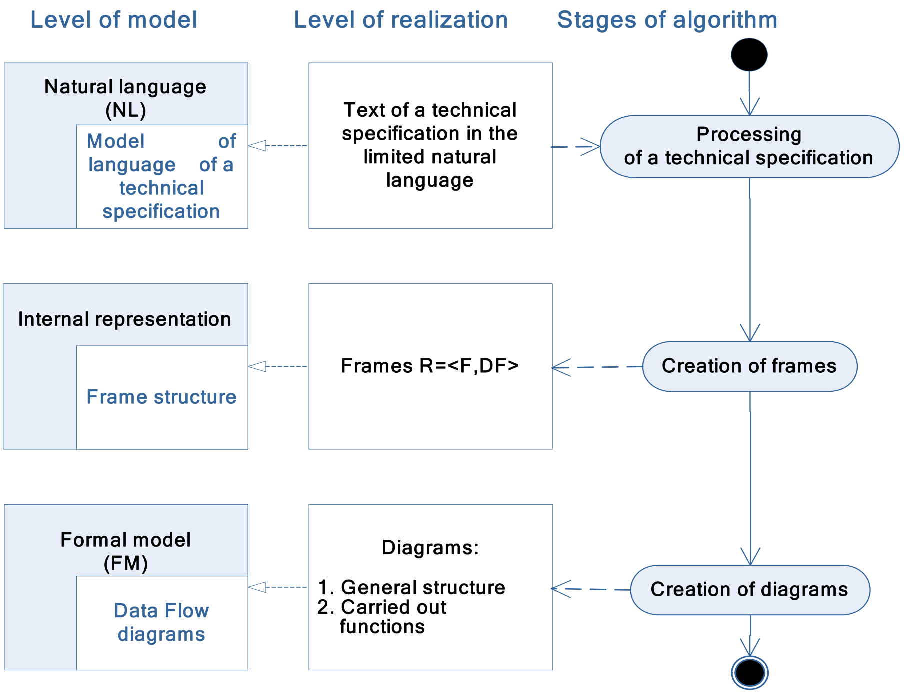


Figure 2: Model of the Text Analysis of a Technical Specification

The model consists of two levels:

1. Natural Language (NL), level of specification on the limited natural language.
2. Formal Model (FM), level of the description of limitations on the graphic Data Flow Diagrams language.

FM level includes the set of data flow diagrams:

1. Common system structure with the specification of incoming and outgoing data flows.
2. System functions and their incoming and outgoing data flows specifications.

According to the proposed model, the system can be considered as a black box. System behavior specification can be treated as a description of functions and data flows.

Structurally level NL can be divided into two parts: grammar of software specification on the limited natural language and frame model [3].

Grammar of software specification contains syntactic and semantic attributes (see Table1)

Table 1: Grammar of software specification

<list of incoming data flows >	*<incoming data flow name >* :: 'Name' *<incoming data flow description>* :: 'Contents' *< list of incoming data flows >*| ε
<incoming data flow description>	*The text containing "entrance" or “entrance data ”*:: 'Clause' *<incoming data flow>*::“Frame Data Flow=Creation ”, “Input=Giving”
<incoming data flow>	[*<Number of data units>*]:: “Slot AMOUNT OF DATA = Giving” [*<Type of data>*]:: “ SLOT TYPE OF DATA = Giving ” *<the Name of incoming data flow >*:: “ Slot NAME OF INCOMING DATA FLOW= Giving”
<function specification >	*< name of the functions liss >* :: 'Name'*< function description >*:: ”Frame FUNCTION = Creation ”*; < List of functions> | ε*
< function type >	«main» | «basic» | «additional»
< function description >	*< Name of function>*:: 'Name', “ Slot NAME OF FUNCTION = Giving ” *< List of incoming data flow > <List of out coming data flow>*

In connection with that the natural language is context-dependent for the description context-dependent grammars attributes are used. With their help it is possible to transfer the information from the left part of a generating rule in right and from the right part in left. Advantages attributive grammars that they can specify both context-free and context-dependent languages.

In rules of grammar there are syntactic and semantic attributes. For example, the syntactic attribute is underlined as follows: <*incoming data flow name* > :: 'Name', and semantic: < *function description* >:: ”Frame FUNCTION = Creation” - the name of attribute and action.

Actually grammar of a technical specification is used for splitting the initial text of the document into sections and processings of most important of them for our problem. It needs precise observance of structure of the document. Technical specification represents the structured text consisting of sequence of preset sections.

Level FM represents a set of Data flow diagrams: general structure of system with the indication of its entrance and target streams; the functions which are carried out by system with their entrance and target streams

Formally frame model can be described as R = <N_R, F_R, I_R, O_R>, where N_R is a name of system, F_R is system functions vector, I_R is incoming data flows vector, O_R is outgoing data flows vector F_R= <N_F, I_F, D_F, G_F, H_F, O_F, where N_F is function name, I_F is incoming data flows vector of F function, D_F is function action, G_F is subject of the function action, , H_F is limitations and restrictions for F function, O_F is outgoing data flows vector of F function.

Let’s denote the data flow by DF (Data Flow), then I_R, O_R, I_F, O_F are denoted by:

DF = <N_{DF}, D_{DF}, T_{DF}, C_{DF}>, where N_{DF} is data flow name, D_{DF} is data flow direction, T_{DF} is data type in flow, C_{DF} is data units per frame.

The model proposed is represented as a frame model with “a-kind-of” links (see Figure 3).

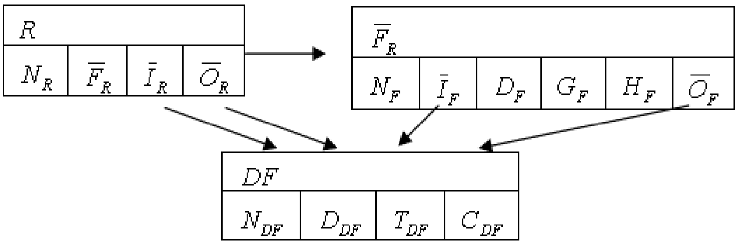


Figure 3: Frame network

Computer-Aided System of Semantic Text Analysis of a Technical Specification

The computer-aided system of semantic text analysis of a technical specification consists of the following subsystems: preliminary text processing, the syntactic and semantic analysis and construction of software models, storage of documents and interface (see Figure 4).

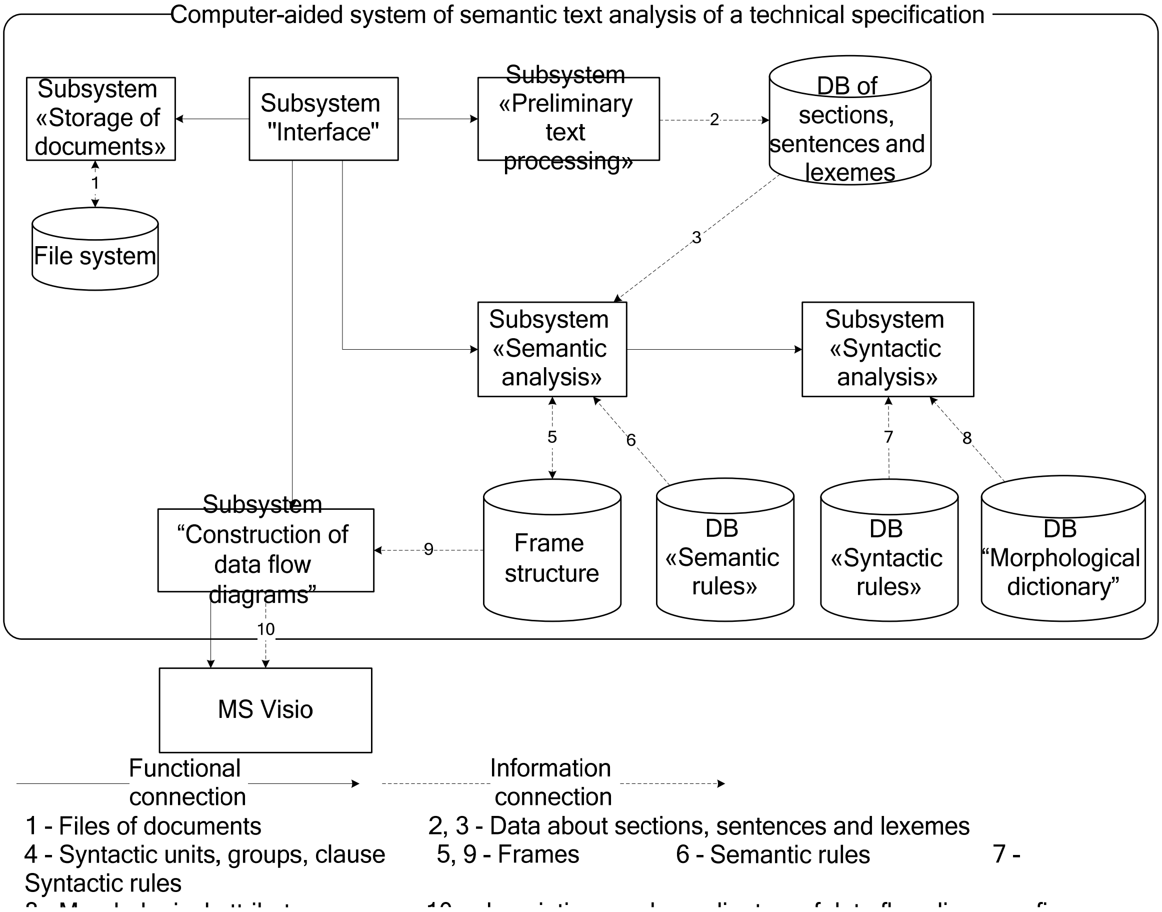

Computer-aided system of semantic text analysis of a technical specification
Subsystem «Storage of documents»
Subsystem "Interface"
Subsystem «Preliminary text processing»
DB of sections, sentences and lexemes
File system
Subsystem «Semantic analysis»
Subsystem «Syntactic analysis»
Subsystem "Construction of data flow diagrams"
Frame structure
DB «Semantic rules»
DB «Syntactic rules»
DB "Morphological dictionary"
MS Visio
1
2
3
5
6
7
8
9
10
Functional connection
Information connection
1 - Files of documents
2, 3 - Data about sections, sentences and lexemes
4 - Syntactic units, groups, clause
5, 9 - Frames
6 - Semantic rules
7 - Syntactic rules

Figure 4: Architecture of computer-aided system of semantic text analysis of a technical specification

Preliminary text processing is necessary to share of a technical specification on separate lexemes. The incoming information of a subsystem is the text of a technical specification in the limited natural language, the target information - tables of sections, sentences and lexemes of a considered technical specification. Results can be submitted both as corresponding tables, and as a tree of sections.

Already after the first stage work not with the text of a technical specification, but with its parts submitted on sections is made. On a course of work of a technical specification shares all over again on more and more fine sections, then on separate sentences (with preservation of sections structure) and lexemes with the instruction of an accessory to sentences.

Preliminary text processing is carried out with use of final automatic device. During the work of final automatic device the symbols acting on its entrance, collect in the buffer. In the certain conditions of final automatic device record of the current contents of buffer in one of tables then the buffer is devastated is carried out. Work of automatic device proceeds up to achievement of a final condition.

After that the received tables act on an entrance of a subsystem of syntactic and semantic analysis. The semantic analysis of a text is made on the basis of the developed grammar of a text of technical specification.

Rules of top level serve for analysis of sections of top level. Rules for analysis of sections consist of two parts: the first part serves for analysis of a section name; the second part serves for analysis of a text contents in section. Symbols of the given grammar possess syntactic attributes. In attributes of non-terminal symbols names of frames or names of slots in which the information received during the further analysis should be placed are specified. Syntactic attributes of text can be in addition specified in attributes of terminal symbols. Comparison of words at analysis is made in view of their morphology. During analysis the syntactic and morphological analysis are made only in the event that there is such necessity that time of performance of semantic analysis is considerably reduced.

Let's consider a fragment of the developed attribute grammar submitted in a xml-format:

```
… <global-rule id="Section42" comment = "Section 4.2. Requirements to functional
characteristics"> <rule><ruleref uri="#Section42Name"/><ruleref
uri="#Section42x"/></rule></global-rule>

<global-rule id="Section42Name" sectionPart="Name" comment= "Heading of the unit
4.2."><rule><clause clauseType="UNCERTAIN"/><rule type="or"><words contains="Functions"/>
<words contains= " functional characteristics "/> </rule></rule></global-rule>

<global-rule id="Section42x" frame= "FunctionFrame" frameSlot="Function"
comment="Function"><rule> <ruleref uri="#Section42xName" /><ruleref uri="#Section42xContent"
/> </rule></global-rule>

 <global-rule id="Section42xContent" sectionPart="Content" comment="Inputs and outputs of
function"><rule><ruleref uri= "#Section42xInputs" minOccurs="0"/><ruleref
uri="#Section42xOutputs" minOccurs="0"/></rule></global-rule>

<global-rule id="Section42xInputs" comment="Inputs of function">

<rule><sentence/><clause/><rule type="or"><words contains="Inputs"/> <words contains="entrance
data"/></rule><ruleref uri="#Input" maxOccurs="unbounded"/></rule></global-rule> …
```

The morphological and syntactic modules used in the program, are modules of the foreign developer. If in a rule of grammar there is a terminal having syntactic attribute the mechanism of syntactic analysis for current sentences is started [2].

After creation of a tree of analysis construction of frame description of a technical specification begins. For this purpose the information on frames and names of slots which contains in attributes of symbols of grammar is used.

The received frame structure contains the significant information about system: data about inputs and outputs of system, functions and restrictions. For each function inputs and outputs also are allocated. It allows receiving data flow diagrams of system which is described in a technical specification on the basis of frame structure.

The subsystem “Construction of data flow diagrams” carries out construction and ordering the column of data flows, and also creation the figures of data flow diagrams in Microsoft Office Visio.

For construction of data flows it is prospected of functions inputs conterminous to system inputs. Then functions on which all inputs data act, are located on the one level of diagram. Their inputs incorporate to system inputs. Further it is prospected functions which inputs coincide with outputs of functions received on the previous step. They are located on the following level, their inputs incorporate to outputs of the previous levels functions and with system inputs.

Work of algorithm proceeds until all functions will not be placed on the diagram. After that connection of function outputs with necessary system outputs is made.

The computer-aided system of semantic text analysis of a technical specification is developed on Microsoft .NET Framework 2.0 platform (language of development C#) using integrated development environment Visual Studio 2005.

Scientific Novelty

Scientific novelty consists in the following: the model of text analysis of a technical specification at the initial stages of software engineering, including semantic model of text of a technical specification, the technique of transformation matter of text into the frame structure and construction of the model of the software on its basis are developed.

Practical Value

Practical value of work is that as a result of development and introduction of a suggested technique quality of software engineering raises due to automation of routine work of the person on extraction of helpful information from standard documents and to displaying it as software models.

Conclusions and Future Work

Software designing differs from designing in other areas of a science and technics a little, therefore it is possible to expand results of the given work for application in other areas of human knowledge. Thus, opening prospects raise a urgency of the given work.

Bibliography

1. Kamsay, A. Computer-aided syntactic description of language systems/ A. Kamsay// Computational linguistics. An international handbook on computer-oriented language research and applications. Boston: Walter de Giuyter, 1989.- P.204-218
2. Reyle, U. Natural language parsing and linguistic theories/ U. Reyle. Berlin: Rohrer Dordrecht, 1998.- 625 p.
3. Tools Development For Computer Aided Software Engineering Based On Technical Specification's Text Analysis / A.Zaboleeva-Zotova, Y.Orlova // Interactive Systems And Technologies: The Problems Of Human-Computer Interaction: Proc. of the Int. Conf., Ulyanovsk.

Authors' Information

***Alla V. Zaboleeva-Zotova** – PhD, professor; CAD department, Volgograd State Technical University, Lenin av., 28, Volgograd, Russia; e-mail: zabzot@vstu.ru*

***Yulia A. Orlova** – PhD student; CAD department, Volgograd State Technical University, Lenin av., 28, Volgograd, Russia; e-mail: yulia.orlova@gmail.com*

SECOND ATTEMPT TO BUILD A MODEL OF THE TIC-TAC-TOE GAME [1]

Dimiter Dobrev

Abstract: *We want to make a program which can play any game or in other words we want to make AI. It is impossible to include in this program the rules of all games and that is why our program should be able to find these rules by itself. We cannot solve this problem in the general case. So, our first task will be to make a program which is able to find the rules of the Tic-Tac-Toe game. Even this task is too complicated. So, first we will try to find these rules manually and this will help us make a program which is able to find these rules automatically.*

Keywords: *Artificial Intelligence, Games, Automated reasoning.*

ACM Classification Keywords: *I.2.0 Artificial Intelligence - Philosophical foundations*

Conference: *The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Introduction

We are trying to build a formal model of one particular game. We need such a model in order to predict the future. For example, look at the third position shown on figure 1. At this position you see that if you play in the centre then you will win but if you do not play in the centre then probably you will lose. This means that you have in your head a model of the Tic-Tac-Toe game and you can predict the future and say what will be the consequences of your next move.

We try manually to find a model of the Tic-Tac-Toe game which will give us the possibility to play this game successfully. If we write a simple program which can recognise which move is wining and which one is losing then this program will be a model of the game. Anyway, we do not like this model because it is not easily discoverable. First reason for that is that the set of all programs is too huge. The second reason is that the program models are not easily checkable. This means that if you have one program which is a model of the game it is not easy to check this fact because for this you should play some time following the recommendations which this model gives and after that judge how good this model is on the basis of the results achieved in the testing period.

So, we are looking for an easily discoverable model. Do not forget that we need a model which can be found automatically.

There is one more reason why programs are not good candidates for such models. The programs are not easily modifiable. If you make a small random modification in one program then as a result you will receive a program which will work in totally different way and in most cases it will not work at all. In most optimisation tasks we try to find the best solution by making small modifications. Let's take for example the simplex algorithm which is for solving the linear programming problem, or the "go up" algorithm for finding the highest place, or the process of the evolution in nature where the child is a small modification of its parents.

How will our easily modifiable model look like? It will be a logic theory which consists of set of assumptions (axioms). A modification will be to add or to remove one assumption.

[1] This publication is partially supported by the Bulgarian Ministry of Education (contract БОЕ 4-02/2004 г.)

Formalisation of the game

Before finding a formal model of the game we have to formalise it. This means to represent the real game as a mathematical object. In this case this mathematical object will be the set of all possible sequences of inputs and outputs. Of course, the best representation of this set is the tree of all possible moves. Here when we say moves we mean our moves and the moves of our opponent.

You see that for the formalisation of the game we need formalisation of our opponent. Really, when you play a game you try to build a model both of the game and of your opponent. So, when you play you try to understand your opponent and to predict his behaviour.

From formal point of view, playing with different opponents is playing a different game. This is because different opponents play different moves and that is why the tree of all possible moves is different.

In order to formalise the game we need to fix the opponent. Let us assume that our opponent makes line and wins if he can do this on the next move. Otherwise he plays a random move choosing randomly from all possible correct moves with equal possibility for each of them to be chosen.

Note. If we have a game and a fixed opponent and if we try to make a model of both of them do we need to separate the model of the game from the model of the opponent? For example, if our opponent never starts with a move in the centre then should this fact be in the model of the game or should it be in the model of the opponent. The answer is that we do not care why the opponent has certain behaviour. Maybe this is part of the rules of the game or maybe not. In any case, if certain behaviour is fact then we can use this fact no matter what is the reason behind it.

There is one case when it is good to separate the model of the game from the model of the opponent. This is the case when we try to see the world through the eyes of our opponent. We will not try to do this because it is a difficult task. They say that children younger than three years cannot do this, so our program will not be able to do this either.

One more reason for this is that we will assume that each time we play the first move. So, the world is not symmetric for us and for our opponent and this makes it more difficult to look at the world through the eyes of our opponent.

In order to finish the formalisation of the game we have to say what is the information which we input and output on every move. Let us assume that on every move we input the game status (i.e. what we have on the board) and on every move we output the coordinates of the cell where we put a cross.

So, we want to represent the game as sequence of inputs and outputs like this:

$$a_0, b_0, a_1, b_1, a_2, b_2 \ldots$$

On figure 1 you can see an example of a game.

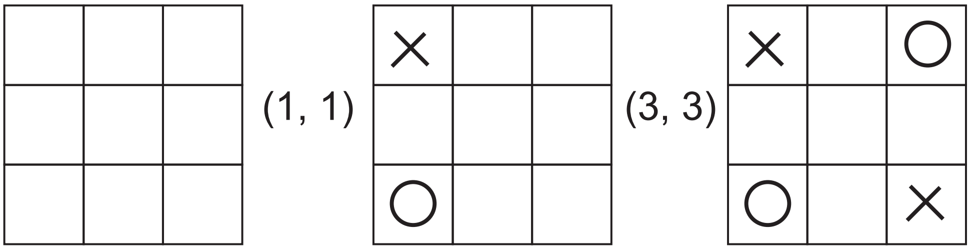


Fig. 1: Part of a game

The first attempt

On the basis of this formalisation we made our first attempt to build a model of the Tic-Tac-Toe game. Let's start with the size of the input and the size of the output according to this formalisation. The input includes the situation on the board. These are nine cells and to describe the situation we will use 18 bits (two bits per cell). Also we need three additional bits which we will call "victory", "loss" and "bad move". These three bits give us information about the result of our activity. For example, if we win the bit "victory" will be on and if the game is draw then both "victory" and "loss" bits will be on. The bit "bad move" will be on when we try to do something forbidden (like putting a cross in a cell which is not empty). These three additional bits give us the purpose of our program. It is not sufficient to have a model of the game because if you do not have a purpose you cannot distinguish good from bad and you cannot choose your next move even if you know perfectly well what will happen if you choose this move.

> **Note.** In the terms used in our definition of AI [1, 2, 3] the purpose of our program is called "the meaning of life". Here this purpose is clear. It is to achieve more victories and fewer losses. In this paper it is almost useless to explain what the purpose of our program is because it is obvious that when we play the Tic-Tac-Toe game our purpose is to win. Anyway, these explanations are not useless because in other games the purpose may be difficult for defining. Let's take for example the real life of a human being. In this case if we have a clear purpose (or a clear meaning of life) we can look at the real life as a game.

So, our input is 21 bits (18 for the state of the board plus three additional bits). Our output is 4 bits (two for the "x" and two for the "y" coordinate). Now we will try to make simple implications of the type: "If you see this and do that then on the next step you will see this." Here is an example of such simple implication:

$$p_{11} \,\&\, \neg\, out_{x1} \,\&\, \neg\, out_{x2} \,\&\, \neg\, out_{y1} \,\&\, \neg\, out_{y2} \Rightarrow bad\_move$$

The meaning of this simple implication is that if you have a cross at position (1, 1) and if you play at this position then on the next move you will see the bit "bad move" on.

How many are these simple implications? The maximum length of them is 46 (two times 21 plus 4). So, their number is 3 to the power of 46. Of course, we do not need all of them but only these which are true and even only small fraction of them which are essential. The number of these essential implications is millions and the first program which we have made in order to decide the problem cannot manage to proceed with so many implications. Anyway, we have made a more sophisticated program which keeps these implications in a tree structure, which allows it to proceed with sufficiently many implications in a reasonable time. You can find this program in [10].

The idea of the first attempt is that in the set of the simple implications which are true there is coded the information about the experience (about the first moves which will be used for our education). Of course, this is not all the information from the experience but this part of it which is essential. So, on the basis of this set of simple implications we can say which move is bad. For example, the implication which is described above, says us that if we have a cross at position (1, 1) we cannot play at this position.

Unfortunately, this first attempt was a complete disaster. Really, the essential implications were millions but this was only a technical problem, which was solved in [10]. The real problem was the number of steps which we have to make in order to collect enough experience. In [10] you will see that after 20,000 steps the program almost stops to make bad moves. This means that the time (the number of steps) for education is extremely big. Let's take an example connected with human beings. With people boys study slower than girls. Anyway, this is not a problem because they study just a little bit slower. In our case we have a serious problem because the time for education is so huge that it is practically infinite.

Where is the problem? This model is too stupid and here you cannot apply neither analogy nor something similar to analogy. Here the problem comes from the formalisation of the game. On one hand, the input is too big (21 bits) and on the other hand, we have made the wrong assumption that we see the full status of the game. For example, in real world this assumption is not true because we do not see the full status of the world (Actually, we cannot see behind our back.) The assumption that we see (receive as an input) the full status of the game is possible only with very simple games.

The conclusion is that the next attempt to build a model of the Tic-Tac-Toe game will start with a new formalisation of the game.

First published attempt

The first attempt was a complete disaster and that is why it is not published. The next attempt was much better and that is why it is published in [4, 5].

As we've said, this attempt starts with a new formalisation of the game. Here we reject the assumption that we see the full status of the game. Now we will assume that we see only one cell (the current cell). In figure 2 you will see the eye which pinpoints the current cell. In this case the input is only one cell which is two bits. Of course, we have also three additional bits as before. The output is also different. Before we had nine moves (to put cross in one of the cells). Now we have six moves and they are to move the eye in the four possible directions, to put a cross at the current cell and to start a new game.

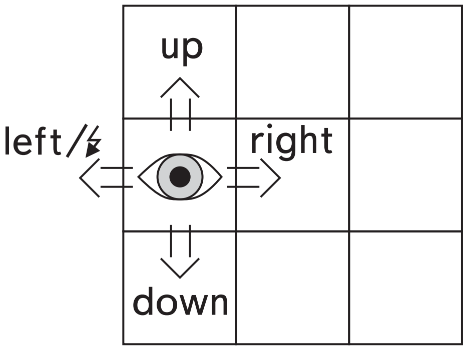


Fig. 2: The new formalisation

In this new formalisation the concept of the move is changed. Before the move was to put a cross somewhere but now it is to move through the board or to put a cross. The main change is in the concept of what we see. Before the assumption was that we see everything but now we see only a small part of the game status and we have to imagine what the status of the game is. So, now building of the model is a much more interesting task.

On the basis of this new formalisation in [4, 5] there was made an attempt to build a model of the Tic-Tac-Toe game. This model included three main parts: simple implications, FSMs (finite state machines) and first-order formulas. The connection between these three parts was not clear and the attempt from [4, 5] did not give us a working model of the game.

The second attempt

That is the reason why in this paper we will make a new attempt to build a model of the Tic-Tac-Toe game. The basis of this new attempt will be the first-order logic with types. From the theoretical point of view, there is no difference between this logic and the common first-order logic but types give us much bigger expressiveness.

What is the difference between the first-order logic and first-order logic with types? In the first case the universe is one not empty set but in the second case the universe is a union of several non-intersected sets. The second difference is that in the first case the relation and the function symbols have only valence but in the case with types every argument has a type.

In our first-order logic with types we will have one countable set ***T*** which will correspond to the time and several finite sets which will correspond to the states of some FSMs. In this paper we will mention only the sets ***X*** and ***Y*** which have three elements each and which correspond to the coordinates of the eye.

What will be the structure of the set ***T*** or what will be the structure of the time. We will have two function symbols "next" and "previous". These symbols will have one argument of type ***T*** and will return object of type ***T***. We have to decide whether to make ***T*** isomorphic to the natural numbers or to make it isomorphic to the integer numbers. In other words, to introduce one constant for the first moment or not. The better choice is to make the time isomorphic to integers because this model is simpler. Really, we have a first moment but we cannot use this moment in order to make conclusions. You cannot conclude something like "When I am born they give me milk" because you have not enough statistical information for such conclusion. Even if you have such rule you cannot use it because you will not be born again.

Let's see what one simple implication will look like. For example, "If you see a cross you cannot put a cross":

$$I_x(T)\ \&\ put\_cross(T) \Rightarrow bad\_move(next(T))$$

Now, let us take M_x. This is the FSM which was described in [4].

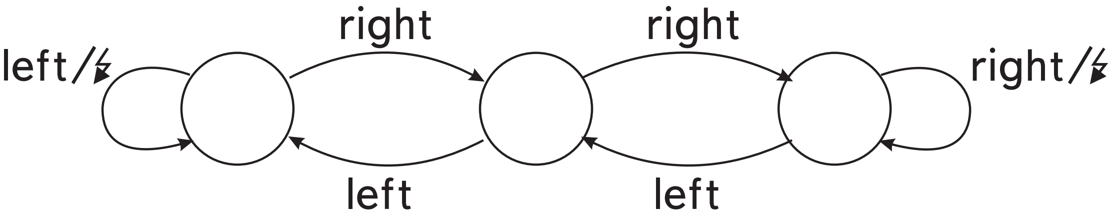


Fig. 3: The FSM which gives the "x" coordinate

M_x is essential part of our model because it corresponds to the "x" coordinate of the eye. M_x can be included in our model. Here is the description of one of its arcs:

$$x(T)=x_1\ \&\ right(T) \Rightarrow x(next(T))=x_2$$

Here ***x*** is a function symbol which has one argument of type time and which returns an object of type ***X***. The meaning is that it returns the current "x" coordinate of the eye or the current state of M_x. The constants x_1 and x_2 will correspond to two of the states of M_x. Of course, these constants will be of the type ***X***.

We will notice that the simple implications and these FSMs are easily discoverable. (This is important because we are looking for a model which can be found automatically.) In [10] you can see that the simple implications can be generated automatically and if the size of the input and output is not too big this can be done without combinatorial explosion. In [8, 9] you can see that M_x can be found through the method of suns. The idea of this method is that FSMs are too many but every FSM can be decomposed in simpler objects, which we call suns. You will receive such object if you observe only one letter in your FSM or observe only the arcs labelled with this letter. The result is one or several simple directed graphs with one cycle and several paths which flow in this cycle. Such graph looks like the picture of the sun which children use to draw. For example, if we take M_x and observe it only on the letter "right" then we will receive one loop (which is a cycle with length one) and one path with length two which flows in this cycle. In this way we will receive one "sun" which includes three arcs. This sun is easily discoverable because one of its arcs gives bad move each time and the other two give correct move each time.

As we said, for us it is important to make a model which is able to give us a mental picture of what we cannot see. The FSMs M_x and M_y give us the idea where we are at the moment (actually we cannot see directly the coordinates of the eye). The next information which we have to include in the model is what the situation on the board is. For this we need first-order formula like this:

$$p(x(T), y(T), T) \Leftrightarrow I_x(T)$$

Here we did not say anything essential. In our input we have five bits and two of them are nameless (we do not know nothing about them). It is normal to try to say something about these two bits. One of them is ***$I_x(T)$***. It is normal to assume that ***$I_x(T)$*** depends on something other than ***T*** or to assume that ***$I_x(T)$*** is a projection of some relation which has more arguments. Actually, ***p*** is this relation and it depends on the time and on the current coordinates. Existence of such relation would be not interesting at all if there was not the following formula:

$$\neg\, put\_cross(T)\ \&\ \neg\, new\_game(T) \Rightarrow\ (\forall X\ \forall Y\, (p(X, Y, T) \Leftrightarrow p(X, Y, next(T))\,))$$

This formula gives us the stability of the relation ***p***. This formula is not easily discoverable but we are looking for stable relations and this means that we are looking for formulas which describe stability.

At the end we will show what the formulas which describe the victory look like.

$$(\exists X \forall Y\, p(X, Y, T))\ \Rightarrow\ victory(T)$$

$$(\forall X \forall Y (d_1(X, Y) \Rightarrow p(X, Y, T)\,)) \Rightarrow victory(T)$$

The first of these formulas says that if we made a vertical line we won. The second formula says that if we made the first diagonal we won. The first diagonal is a relation between ***X*** and ***Y*** but we have 512 relations between ***X*** and ***Y***. Is this relation an easily discoverable one? Yes, because FSMs ***M_x*** and ***M_y*** are isomorphic and there are only two isomorphisms between them and these isomorphisms are the diagonals. So, this relation is easily discoverable because it is special.

Modification of the method of resolution

Even if we have a model of the game, we need a method for proving formulas in the first-order logic in order to make a program which can play successfully on the basis of this model. Of course, we have such a method and this is the method of resolution.

This method has one serious disadvantage, which we have to fix in order to make AI which is capable to work in acceptable time. In [6, 7] you can see that if we do not worry about the combinatorial explosion then it is easy to make AI (which is useless because it cannot work in acceptable time). If we want to make useful AI then we have to worry about its efficiency.

The problem, which is to be mainly blamed for the bad efficiency, is that the method of resolution starts every time from the beginning. In this way it is practically impossible to prove anything complicated by this method.

We would like to make such modification that allows constructing a database of proven disjuncts. Of course, only tautological disjuncts are true without any propositions but these disjuncts are not interesting. That is why we would like to build a semilattice of different logic theories which are to contain interesting implications. For example, if we want to prove $\varphi \Rightarrow \psi$ then we can find the logic theory where φ is a proposition and to check if ψ is an already proven implication in this theory.

Unfortunately, this works only for formulas which we can prove without skolemization. The main reason that the resolution starts every time from the beginning is that first we make the skolemization and on the next step we make the resolution. In order to improve the efficiency of the resolution we have to allow the skolemization and the resolution to work in parallel. This will be impossible if skolemization gives random names of the objects because if it gives a name to one object twice then it will give two different names to it but we would like the name on the second time to be the same.

So, what type of systematic names should our new skolemization use? For us the best choice is the system proposed by David Hilbert. For every formula $\varphi(x)$ he defined an object $\tau_x\varphi(x)$ which satisfies $\varphi(x)$ if there exists an object which satisfies $\varphi(x)$ (description of these Hilbert's terms you can find in [12]). An important fact is that if $\varphi(x) \Leftrightarrow \psi(x)$ then $\tau_x\varphi(x) = \tau_x\psi(x)$. This means that for one and the same object we give one and the same name.

For example, we want to prove $\forall x(\varphi(x) \Rightarrow \psi(x))$. If we do this in the old way the skolemization will give a random name to the object which satisfies $\varphi(x) \,\&\, \neg\, \psi(x)$. After that, the resolution will prove that the existence of such object leads to contradiction. It will be much faster if we have already developed the theory of $\exists x\ \varphi(x)$ and if in this theory $\psi(x)$ where x is $\tau_x\varphi(x)$ is already proven and this will be sufficient. Really, we want to prove that in this theory is true $\psi(x)$ where x is $\tau_x(\varphi(x) \,\&\, \neg\, \psi(x))$ but there is a connection between $\tau_x\varphi(x)$ and $\tau_x(\varphi(x) \,\&\, \neg\, \psi(x))$. For the second object we know more, so everything which we can say for the first object we can say also for the second one.

This modification of the method of resolution is only an idea and we need a lot of work in order to make a real system which is based on this idea.

Bibliography

[1] Dobrev D. D. AI - What is this. In: PC Magazine - Bulgaria, November'2000, pp.12-13 (in Bulgarian, also in [11] in English).

[2] Dobrev D. D. AI - How does it cope in an arbitrary world. In: PC Magazine - Bulgaria, February'2001, pp.12-13 (in Bulgarian, also in [11] in English).

[3] Dobrev D. D. A Definition of Artificial Intelligence. In: Mathematica Balkanica, New Series, Vol. 19, 2005, Fasc. 1-2, pp.67-74.

[4] Dobrev D. D. Testing AI in One Artificial World. In: Proceedings of XI-th International Conference KDS 2005, June, 2005 Varna, Bulgaria, pp.461-464.

[5] Dobrev D. D. AI in Arbitrary World. In: Proceedings of 5th Panhellenic Logic Symposium, July 2005, University of Athens, Athens, Greece, pp. 62-67.

[6] Dobrev D. D. Formal Definition of Artificial Intelligence. In: International Journal "Information Theories & Applications", vol.12, Number 3, 2005, pp.277-285.

[7] Dobrev D. D. Formal Definition of AI and an Algorithm which Satisfies this Definition. In: Proceedings of XII-th International Conference KDS 2006, June, 2006 Varna, Bulgaria, pp.230-237.

[8] Dobrev D. D. Two fundamental problems connected with AI, XII International Conference "Knowledge-Dialogue-Solution", June 2007, Varna, Bulgaria.

[9] Dobrev D. D. The "sunshine" Method for Finding Finite Automata, Trends in Mathematics and Informatics, July 2007, Sofia, Bulgaria.

[10] Dobrev D. D. Generator of simple implications, http://www.dobrev.com/AI/app4.html

[11] Dobrev D. D. AI Project, http://www.dobrev.com/AI/

[12] Bourbaki N. Theory of sets, Chapter 1.

Authors' Information

Dimiter Dobrev *– Institute of Mathematics and Informatics, BAS, Acad.G.Bonthev St., bl.8, Sofia-1113, Bulgaria; P.O.Box: 1274, Sofia-1000, Bulgaria; e-mail: d@dobrev.com*

СТРУКТУРИРОВАНИЕ ОНТОЛОГИИ АССОЦИАЦИЙ ДЛЯ КОНСПЕКТИРОВАНИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Виктор Гладун, Виталий Величко, Леонид Святогор

***Аннотация**: Рассмотрен подход к конспектированию ЕЯ текстов с использованием трехуровневой онтологии ассоциаций. Предложенная структура онтологии позволяет улучшить связность конспекта.*

***Ключевые слова**: тематический анализ текста, конспектирование текста, онтология ассоциаций.*

***ACM Classification Keywords**: I.2.7 Natural Language Processing - Text analysis*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

Процедуры репрезентации и получения новых знаний занимают в общей проблеме искусственного интеллекта ведущее место. В поисках универсальных средств конструктивного описания окружающего материального мира, адекватного человеческому мышлению и познанию, всё чаще обращаются к онтологиям. Онтология представляет собой совокупность концептов, отношений и функций интерпретации. К достоинствам онтологий можно отнести: а) глубокое взаимодействие объектов и явлений с контекстной средой; б) экономное хранение информации, требующее запоминания концептов и отношений, а не сцен; в) универсальный характер онтологии, допускающий использование её структуры в качестве инструмента для решения задач семантического анализа естественно-языковых (ЕЯ) текстов. Одной из таких задач является *конспектирование текста* как способ сжатого и релевантного представления содержания дискурса. В качестве примера программы выполняющей конспектирование русскоязычного текста можно привести программу КОНСПЕКТ [1], в которой для указания семантических отношений между словами используются ассоциативные связи, а множество ассоциативных признаков отражает некоторые категории внешнего мира (например, ассоциации по роду деятельности, по времени и другие). Однако слабым местом системы ассоциативных связей является упрощённая, одноуровневая структура, которая не учитывает глубинную иерархию понятий, описывающих реальный мир. Поле ассоциаций представляет собой просто *список концептов*, размещённых в алфавитном порядке.

Отсутствие структуры над полем ассоциаций приводит к потере глубоких ассоциаций и, кроме того, не позволяет в рамках выбранной схемы оперировать элементами смыслового сопровождения при раскрытии содержания текста через конспект.

Задача заключается в том, чтобы превратить систему ассоциативных связей в иерархическую *структуру репрезентации знаний*, адекватную иерархии понятийного аппарата человека, и использовать её затем для семантического анализа ЕЯ текстов.

Обоснование подхода и решаемые задачи

В настоящее время предложено много вариантов описания мира, которые опираются на тезаурусы и онтологии [2]. Для репрезентации знаний об Универсуме они широко используют философские категории и предназначены либо для систематизации лексического богатства естественного языка (словари Роже, Дорнзайфа, Идеографический словарь русского языка О.С.Баранова [3]), либо для структуризации знаний человека о мире (онтологии «Mikrokosmos», SUMO, Дж. Совы [4]). Существуют и проблемно-ориентированные онтологии, которые приспособлены, например, для построения языково-онтологических информационных систем [5, 6]. Для построения многоуровневых онтологий используются методы и математические модели, содержащие модели онтологии, знаний и действительности предметной области (Про) [7]. Предлагаемая здесь структурированная онтология ассоциаций (ОнтА) служит для решения более частной задачи – *тематического анализа текстов* [8].

При построении ОнтА авторы исходили из следующих предпосылок: а) научной методологией репрезентации окружающего мира должен быть *системный анализ;* б) терминология (категории и концепты онтологии) должны в основном базироваться на понятиях, которые установились в *естественных науках,* с привлечением, в необходимых случаях, философских категорий; в) для конструктивного практического использования структура онтологии должна быть иерархической и содержать верхний, средний и нижний уровень иерархии понятий.

Системный анализ. Понятийно-содержательный подход представления знаний, присущий системному анализу, в отличие от формально-математического, важен именно с позиций выявления семантических отношений. При анализе *системно-информационной картины мира* [9] рассматривают следующие основные *типы ресурсов* в природе и обществе: *Вещество* – субстанция, отображающая состояние материи; *Энергия* – характеристика движения материи; *Информация* – мера порядка и самоорганизации материи; *Человек* – уникальный ресурс общества, субъект осознания материи, мера интеллекта; *Организация* – форма упорядоченности ресурсов и существования системы; *Пространство* – мера протяжённости (распространения и распределения) материи; *Время* – мера существования состояния материи (вещества).

Эти ресурсы, на наш взгляд, обладают необходимой *содержательной строгостью*, поскольку являются объектами исследований в физических и социальных науках. Они могут быть использованы при синтезе онтологии ассоциаций в качестве категорий верхнего уровня. Тем самым намечено существенное отличие ОнтА от известных моделей описания мира, которые были упомянуты выше.

Принципы построения онтологии. Задачи исследования и цель

С учётом рекомендаций, которые необходимо выполнять при построении онтологии [6, 10, 11], в данной работе предлагается трёхуровневая онтология для решения задачи конспектирования ЕЯ текстов. В основу разработки положены следующие принципы.

1. *Принцип полноты.* Категории верхнего уровня должны исчерпывающим образом представлять Материю; за пределами этих категорий не должно существовать никаких проявлений сущего.

2. *Принцип естественнонаучности и проблемной ориентации*. Все категории и концепты онтологии должны быть выражены понятиями, которые установились в естественных и математических науках при изучении материального мира и являются общепринятыми. При этом часть онтологии должна быть представлена концептами, которые широко используются в междисциплинарных текстах (с нейтральной, общедоступной лексикой), а вторая часть онтологии структурируется под конкретную область знаний (ПрО). Первая часть имеет постоянный статус, а проблемно-ориентированная онтология формируется специалистом и носит переменный характер.

3. *Принцип взаимосвязанности уровней.* Категории онтологии верхнего уровня раскрываются наборами концептов среднего уровня. В свою очередь, концепты нижнего уровня должны служить определителями для терминов словаря ПрО. Связь между средним и нижним уровнями организуется с помощью именованных отношений вида: «быть частью», «принадлежать множеству», «совпадать с», «находиться в семантическом отношении с».

4. *Принцип ассоциативности*. Концепты онтологии нижнего уровня должны служить полем для индексирования терминов ПрО. При этом используются семантические отношения вида: «находиться в ассоциативной связи с».

5. *Принцип отражения антагонизмов*. Концепты, которые отражают свойства или понятия, имеющие свою противоположность или дополнительность по равному основанию, входят в онтологию парами или тройками полярных обозначений.

На основе сформулированных предпосылок становится ясной следующая перспектива действий. Необходимо выбрать категории, концепты и связи между ними для верхнего и среднего уровней ОнтА. Необходимо создать поле концептов нижнего уровня – внести в него термины ПрО. Согласовать поле концептов нижнего уровня с концептами среднего уровня. На заключительном этапе следует связать нижний уровень онтологии с базой данных естественного языка, для чего необходимо найти актуальные ассоциативные связи между словарём основ русского языка и терминами, выбираемыми из поля концептов нижнего уровня. Процесс установления ассоциативных связей называется индексацией словаря.

В результате выполнения этих действий онтологическая структура становится конструктивной для процедур семантического анализа и раскрытия темы. Она замкнёт слова русского языка (слова, взятые из текста) через их индексы (связи с концептами нижнего уровня) на *траектории* внутри сетевых структур нижнего, среднего и верхнего уровней. Это позволит, кроме составления самого конспекта, сопроводить его комментариями, которые будут активизированы на траекториях сетевых структур и тем самым улучшить семантическую компоненту конспекта.

Цель структурирования Онтологии ассоциаций заключается в том, чтобы создать трёхуровневую иерархическую онтологическую систему, которая в сжатом виде отражает актуальные знания о структуре внешнего мира, ориентирована на обработку корпуса текстов как общего (междисциплинарного), так и проблемно-ориентированного характера, и позволяет более глубоко раскрыть тему при конспектировании текста.

Выбор категорий верхнего уровня онтологии ассоциаций

Существует много подходов к дихотомии Мира. Нам представляется наиболее конструктивной идея, выдвинутая академиком Вернадским, который построил материалистическое мировоззрение как единство *Косного вещества, Биосферы и Ноосферы.* В качестве методологии онтологического синтеза, как указывалось выше, принят системно-аналитический подход.

Узловой точкой общей картины мира является философская категория *Материя*. Она может быть исчерпывающим образом представлена тремя категориями: *Вещество* (косное), *Энергия, Жизнь* (субстанция живого). Каждая из трех категорий представлена рядом подкатегорий, как показано на рис. 1. Используемые выше категории верхнего уровня образуют *кластеры понятий* для развития и усложнения онтологии. Приведенную онтологическую структуру необходимо снабдить некоторыми пояснениями.

Первое. Общепринятыми в теории познания являются такие важные философские понятия, как *Материя, Бытие, Сознание, Субстанция, Субстрат, Мера, Пространство, Время, Состояние, Свойство, Количество, Качество* и другие. В ОнтА почти все эти понятия, или эквивалентные им, перенесены на средний уровень, благодаря чему они освобождаются от чисто философского смысла и «работают» как термины естественных наук. Например, пространство и время присутствуют как конкретные признаки *локализации* объектов и явлений в четырёхмерном координатном пространстве. Термины количество и качество определяются при помощи *меры* и так далее.

Второе. На прагматическом уровне можно показать, что предлагаемая онтология обладает свойством *полноты.* Будем исходить из того, что всё, что мы знаем о свойствах *Материи*, заключено в следующих четырёх постулатах. *Материя*: а) *существует* как объективная реальность, б) *проявляет себя* в движении и развитии, в) *распределена* в пространственно-временном континууме и г) *отображается* разумом. В таком случае формами проявления *Материи* служат *Вещество, Жизнь и Энергия*. В свою очередь, *пространство* и *время* служат формами распределения *Материи*. Все пять форм проявления и распределения материи интегрируются в едином поглощающем понятии – *Бытие Материи*. Следовательно, онтология, замыкаясь на понятие *Бытия Материи*, исчерпывающим образом отображает все известные (или сущие) свойства данной субстанции, то есть – является *полной системой* верхнего уровня репрезентации знаний.

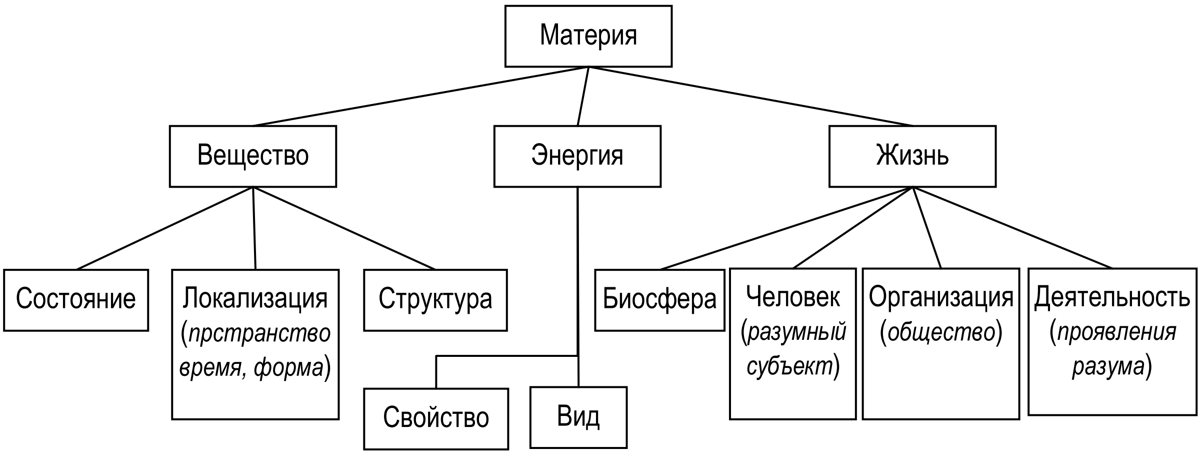

Материя
Вещество
Энергия
Жизнь
Состояние
Локализация (*пространство время, форма*)
Структура
Свойство
Вид
Биосфера
Человек (*разумный субъект*)
Организация (*общество*)
Деятельность (*проявления разума*)

Рис. 1. Структура онтологии ассоциаций верхнего уровня.

Третье. В онтологии ассоциаций важную роль играют такие понятия, как *мера* и *имя*: они пронизывают все уровни онтологии и характеризуют большинство концептов. Здесь понятие меры используется не в математическом и не в философском смысле (как связь между количеством и качеством), а как результат измерения; имя обозначает множество или кластер. Иногда можно проследить тесную связь между *мерой* и *именем*: если число служит точечной оценкой количества и часто несёт избыточную информацию, то имя задаёт сразу диапазон измерений, то есть – кластер. Например, на шкале температуры различают состояния: *жара, тепло, нормально, прохлада, холод, мороз;* все они обладают ясной и непосредственной семантикой. В онтологии ассоциаций перечислять все состояния системы (кроме случаев, когда они существенны для представления ПрО) нет необходимости; они будут обобщены понятиями: *мера тепла* или *имя состояния* или *свойство системы*. Хотя ОнтА оперирует с однословными концептами, в некоторых случаях применяются сложные концепты (например, *рождение-гибель*), когда пара или тройка связанных (по равному основанию и противоположных по смыслу) имён подчёркивает область определения сложного концепта. В общем случае имя служит идентификатором состояния системы, идентификатором множества или абстрактным понятием.

Уникальным свойством человеческого языка является передача смысла «по умолчанию». Эту функцию выполняет *мера*, когда она сопровождается контекстом. Например, слово «нормально» обозначает не только ситуацию, но и подразумевает *отклонения от нормы* в обе стороны, то есть – несёт значительную семантическую нагрузку. Авторы исследования разделяют гипотезу, что познание человеком бесконечно-разнообразной внешней среды возможно благодаря его умению классифицировать. Отсюда следует необходимость и универсальность семантической категории *имя*, которое может существовать только в общей языковой среде.

С учётом приведенных выше универсальных категорий, понятий и комментариев строится продолжение онтологической схемы.

Построение онтологии среднего уровня

Онтология среднего уровня (ОСУ) должна связывать категории верхнего уровня сложности с концептами, которые описывают конкретные свойства ПрО на нижнем уровне. Промежуточный уровень иерархии необходим для более глубокого и разветвлённого раскрытия общих связей и закономерностей, накопленных при изучении в разных дисциплинах. По-сути, он представляет собой *слой междисциплинарного человеческого знания* и обобщает коллективный опыт. Более того, проекция нижнего уровня онтологии на средний позволяет раскрыть содержательную компоненту ЕЯ текста и одновременно усилить её объяснительной компонентой онтологии ассоциаций. От удачного построения этой (средней) части онтологической структуры зависят в целом интерпретационные возможности системы семантического анализа.

ОСУ представляет собой конструкт, заполняемый один раз концептами общего назначения. Это не исключает его доработку специалистами разных областей знания. Инженер по знаниям имеет право согласовывать общий уровень онтологии ассоциаций с теми профессиональными знаниями ПрО, которые он будет детально формулировать на нижнем уровне иерархии. Фактически ОСУ выступает в качестве постоянной составляющей онтологии ассоциаций. В отличие от неё, онтология нижнего уровня является переменной составляющей.

Описание структуры ОСУ. Онтология среднего уровня представляет собой совокупность сетевых структур: именем каждой структуры служит категория верхнего уровня, узлами являются концепты среднего уровня, а внутренние связи раскрывают (характеризуют) основные свойства категории.

Заполнение узлов ОСУ произведено такими понятиями, которые являются общеупотребительными для обозначения элементов знания и имеют определённый смысл для специалистов разных, в том числе гуманитарных, областей. Однако задачей данного слоя не является полный охват этих элементов, наоборот: углубление в предметную область достигается средствами нижнего уровня. Ориентиром для выбора концептов ОСУ является семантический анализ дискурсов общетематической направленности.

Каждая структура ОСУ содержит как безусловные, так и сомнительные связи, которые могут быть скорректированы экспертом. Это не является недостатком онтологии ассоциаций, а придаёт ей динамический характер. Кроме того, эксперт по своему усмотрению может свободно выбирать и смешивать концепты и отношения разных типов (род-вид, часть-целое и т.д.); при этом он преследует цель выявления значимых семантических ассоциаций для выразительной репрезентации знаний о данной категории. Многие отношения в категориях среднего уровня представлены отглагольными существительными и могут повторяться в раскрытии структуры различных категорий. «Сила связи» между родовидовыми понятиями, понятиями типа часть-целое выше, чем между понятиями, связанными отглагольными существительными. Введение связей, представленных отглагольными существительными, позволяет расширить цепочки ассоциаций и повысить связность текста, отобранного в конспект.

Здесь, в целях экономии места, полные списки концептов ОСУ (от десяти до тридцати слов на категорию) вместе с их связями не приводятся, а концепты в скобках даны выборочно. В качестве примера представлен фрагмент графа связи понятий категории *Биосфера* (см. рис. 2).

Состояние = **{**устойчивое – изменчивое; покой – движение; конечное – промежуточное; твёрдое – жидкое – газообразное;…**}**. *Структура* = **{**система; однородность – неоднородность; форма; содержание;…**}**. *Локализация* = **{**пространство; время; распределение; начало – конец;…**}**. *Свойство энергии* = **{**движение – покой; процесс – акт; действие – противодействие; превращение – сохранение;…**}**. *Вид энергии* = **{**тепловая; кинетическая – потенциальная; созидание – разрушение;…**}**. *Биосфера* = **{**существо; организм; растение; животное; популяция; среда; экология; существование; выживание; эволюция; развитие-вырождение; размножение; потомство; рождение – гибель; опасность – безопасность; борьба;…**}**. *Человек* = **{**организм; эмоции; разум; характер; воля; занятие;…**}**. *Организация* = **{**социум; управление; семья; закон; свобода – необходимость; …**}**. *Деятельность* = **{**теория – практика; работа – занятие; познание; информация; тактика – стратегия; перемещение; …**}**.

Главное назначение ОСУ состоит в том, чтобы раскрывать категории верхнего уровня и одновременно интерпретировать в сжатом виде концепты нижнего уровня онтологии ассоциаций.

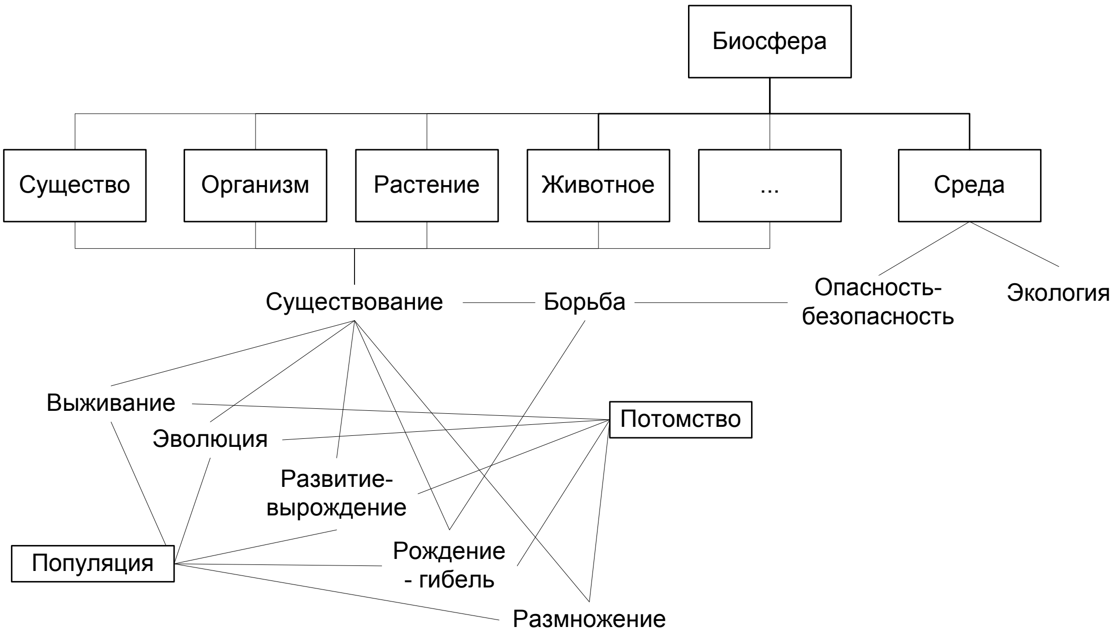


Рис. 2. Структура онтологии ассоциаций среднего уровня для категории *Биосфера*.

Онтология нижнего уровня.

Онтология нижнего уровня (ОНУ) предметной области представляет собой таблицу, в которой слова из словаря основ русского языка (из базы данных) напрямую связаны с ограниченным множеством концептов предметной области, причём концепты ПрО, в свою очередь, имеют восходящие связи к онтологии среднего уровня. Благодаря двусторонним связям в ОНУ база данных через таблицу включается в полную онтолого-ассоциативную структуру. Слова русского языка, которые помещены в словарь основ и употребляются в текстах, оказываются с помощью семантических отношений сопряженными с категориями и концептами всех уровней. Созданный конструкт позволяет решать задачи семантического анализа ЕЯ текста.

Проблема выбора поля концептов нижнего уровня здесь подробно не рассматривается, поскольку выбор зависит от специалиста, который решает определённую задачу. Специалист (эксперт) в своей области выбирает термины и определения из доступных ему источников: учебников, толковых словарей, монографий и т.д. и формирует поле ПрО. Это поле может быть структурировано. Важно, чтобы термины ПрО были некоторым образом связаны с ОСУ, то есть, чтобы ПрО не оказалась изолированной от верхних структур ОнтА.

Например, в определении семантической сети, взятом из толкового словаря по вычислительным системам [12], эксперт выбрал термины: *представление знаний, помеченный граф, вершина графа, понятие, концепт.* Именно эти термины будут соединены с концептом *семантическая сеть* ОНУ и использованы для отбора предложений из дискурса в конспект по теме «семантическая сеть». Концепт *семантическая сеть*, в свою очередь, может быть связан с концептами среднего уровня онтологии, например – *информация* или *наука* из категории *деятельность* ОВУ. Структура онтологии нижнего уровня создаётся в зависимости от требуемой детализации в выделении тематической направленности текста.

Технология конспектирования текста состоит в следующем. На вход системы семантического анализа поступает очередное значимое слово, которое выбрано из дискурса. Оно активизирует нужные связи онтологии ассоциаций и на каждом уровне иерархии возбуждает определённые фрагменты сети. Траектория возбуждения запоминается и используется затем либо для более глубокой интерпретации текста, либо как инструмент для нового раскрытия темы – с учётом уже найденных концептов.

Заключение

Разработанная трёхуровневая иерархическая структура онтологии ассоциаций является сетью, определяемой совокупностью связанных между собой категорий, понятий, концептов и основ русского языка. Её создание подчинено задаче тематического анализа текстов как общей природы, так и проблемно-ориентированных. Онтология верхнего уровня отражает системную картину мира и служит базой для развития онтологии. Онтология среднего уровня обслуживает, в основном, континуум и структуру междисциплинарных знаний. Проблемная область конструируется специалистом на нижнем уровне онтологии, где систематизируются терминологические и специальные знания, взятые из профессиональных источников. Все три уровня иерархии знаний замыкаются на базу данных естественного языка. Процедура концептуального тематического анализа текста состоит в том, что для очередного значимого слова, которое выделено в тексте, на всех уровнях иерархии активизируются соответствующие ему концепты, принадлежащие определённой семантической траектории, после чего данная траектория используется для более глубокого раскрытия темы при конспектировании текста.

Литература

1. Гладун В.П., Величко В.Ю., Святогор Л.А. Тематический анализ естественно языковых текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» Бекасово, 2006 г. – М.: Изд-во РГГУ.–2006. – с.115-118.
2. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология// Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. -Т.1. –Аксаково, 2001. – с.184-188.
3. Баранов О.С. Идеографический словарь русского языка. М.–2002, 1200 с.
4. John F. Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, ©2000. – 594p.
5. Палагін О.В., Петренко М.Г. Модель категоріального рівня мовно-онтологічної картини світу //Математичні машини і системи. –2006. – №3. – с.91-104.
6. Палагін О.В., Петренко М.Г. Розбудова абстрактної моделі мовно-онтологічної інформаційної системи //Математичні машини і системи. –2007. – №1. – с.42-50.
7. Артемьева И.Л. Многоуровневые модели предметных областей и методы их разработки // Десятая нац. конф. по искусственному интеллекту с междунар. участием, Обнинск, 25-28 сентября 2006: сб. тр. в 3-х томах. Москва: Физматлит. –2006. Т.1. – с.44-51.
8. Штерн І.Б. Вибрані топіки та лексикон сучасної лінгвістики. Енцикл. словник. – К.: “АртЕк”. –1998. – 336 с.
9. Казиев В.М. Введение в анализ, синтез и моделирование систем. – ИНТУИТ.ру, БИНОМ. Лаборатория знаний. – 2006. – 248с.
10. Соловьева Е.А. Естественная классификация: системологические основания. Под ред. М.Ф. Бондаренко. – Харьков. ХТУРЭ. –1999. – 222с.
11. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер. –2001. – 384с.
12. Толковый словарь по вычислительным системам. Под ред. В.Иллингуорта и др.: Пер. с англ. А.К. Белоцкого и др. Под ред. Е.К.Масловского и др.: –М.: Машиностроение. –1989. – 568с.

Информация об авторах

Гладун Виктор Поликарпович *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua*

Величко Виталий Юрьевич *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: vitaly@aduis.kiev.ua*

Святогор Леонид Александрович *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua*

К ВОПРОСУ ПРОЕКТИРОВАНИЯ ОНТОЛОГОУПРАВЛЯЕМОЙ ИС ОБРАБОТКИ ЕЯО

Александр Палагин, Николай Петренко

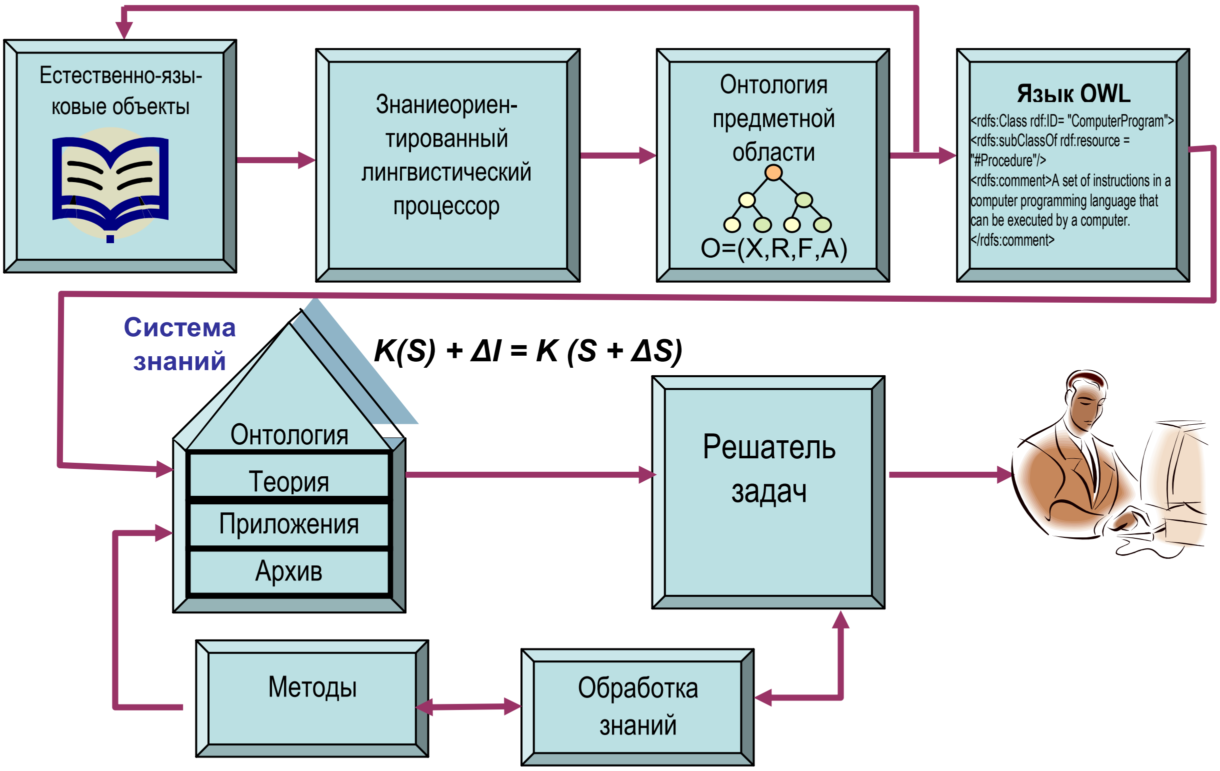
Аннотация*: В статье рассмотрен формальный подход и основное содержание методологии формализованного проектирования*

Ключевые слова*: онтолого-управляемой информационной системы обработки знаний, содержащихся в естественно-языковых объектах (ЕЯО).*

ACM Classification Keywords*: I.2.7 Natural Language Processing - Language parsing and understanding*

Conference*: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Постоянное увеличение объёмов научной информации в сети Интернет и других источников с естественным способом представления требует усовершенствования известных и разработки новых, более эффективных формальных подходов и моделей обработки, начиная от поиска релевантной информации, её анализа, устранения разного рода неоднозначностей, формально-логического представления, онтолого-семантического представления, информационно-кодового представления, представления на формальном языке описания знаний, адаптации известных и разработки новых процедур работы со знаниями и соответствующих им алгоритмов и до получения конкретных результатов пользователя.



```
<rdfs:Class rdf:ID= "ComputerProgram">
<rdfs:subClassOf rdf:resource =
"#Procedure"/>
<rdfs:comment>A set of instructions in a
computer programming language that
can be executed by a computer.
</rdfs:comment>
```

Рис.1. Компьютерная обработка знаний, содержащихся в естественно-языковых объектах

В более общем виде приведенная выше последовательность этапов обработки знаний, содержащихся в естественно-языковых объектах (ЕЯО) представляется цепочкой технологий Natural Language Processing (NLP) → Knowledge Representation (KR) → Knowledge Processing (KP) и показана на рис.1.

Проектирование компьютерных средств, реализующих объединение указанных технологий с учётом специфики входной информации (различного вида источники ЕЯО) можно представить как проектирование сложной системы, включающей языково-онтологическую информационную систему (ЯОИС), полную онтологию предметной области (ПрО) и систему обработки знаний в заданной ПрО. Архитектурно-структурная организация указанных подсистем, реализующая схему обработки знаний (рис.1), соответствует онтологоуправляемой ИС обработки знаний, содержащихся в ЕЯО [1, 2].

Особенности архитектуры и структуры современных "онтологизированных" знаниеориентированных систем (или онтологоуправляемых ИС) полностью определяют подход к их проектированию. Эти особенности можно разделить на системные и технологические.

Системные особенности вытекают из требования построения на базе онтологоуправляемой ИС инструментального комплекса обработки знаний в заданной предметной области с различными техническими характеристиками, в том числе с возможностью подключения аппаратных средств поддержки, спроектированных на базе современных ПЛИС-технологий, потребительскими свойствами и сводятся к созданию класса проблемно-ориентированных онтологоуправляемых ИС.

Технологические особенности определяются состоянием развития элементно-технологической базы современных компьютеров и связаны, прежде всего, с появлением сверхмощных программируемых логических интегральных схем (ПЛИС), вызвавших интенсивное развитие новых методов проектирования знаниеориентированных ИС. Это объясняется целым рядом преимуществ совместного использования программных и аппаратных средств, составляющих в целом информационную систему [3].

В качестве фундаментального механизма при разработке средств интерпретации ЕЯО предлагается модификация известного логико-информационного подхода [4], применительно к цепочке технологий NLP → KR → KP, названном *онтолого-инфологическим* (ОИП), сущность которого состоит в следующем.

В соответствии с логической концепцией онтолого-инфологической модели компонентные процессоры представляются известной композицией сетей операционных и управляющих автоматов. Последняя моделирует иерархическую систему управления современных ИС обработки знаний, и, в свою очередь, описывается композицией программируемых автоматов.

В соответствии с информационной концепцией компонентные процессоры рассматриваются как информационная система, вся информация в которой отнесена к трём "сферам" состояний: хранения, транспортировки и преобразования. Очевидно, что при определённых соотношениях между объектами информации в этих сферах можно получить оптимальные технические параметры онтологоуправляемой ИС.

Под уровнем управления $\tau_{\mathrm{i}}\left(\forall i = 0 \div h\right)$ понимается совокупность структурных и информационных (программных) средств представления и интерпретации операторов языка соответствующего уровня системы, а также формирования необходимой последовательности их выполнения в соответствии с заданным алгоритмом. Наиболее известными и распространёнными в архитектурах современных процессоров и ЭВМ являются микропрограммный, программный и алгоритмический уровни.

Таким образом, формальное описание модели можно представить в виде $\overline{\underset{\mathrm{i}=\overline{0,\mathrm{h}}}{\forall}}\tau_i = \left\langle A_i, \Lambda_i, R \right\rangle$, где A_i - множество алгоритмов, реализуемых и записанных в памяти на *i*-м уровне; Λ_i - множество операторов программирования *i*-го уровня, в терминах которых представлены эти алгоритмы; R - множество информационно-кодовых представлений операторов *i*-го уровня.

Оптимальной считается такая структурная реализация модели онтологоуправляемой ИС, для которой в соответствии с принятыми критериями найдены оптимальное количество уровней и оптимальные соотношения между определёнными характеристиками компонент на каждом уровне, а также соответствующими характеристиками компонент соседних уровней.

Кратко рассмотрим методологию проектирования указанной онтологоуправляемой ИС, которая в свою очередь представляется композицией методик формализованного проектирования подсистем обработки ЕЯО, полной онтологии ПрО и обработки знаний.

Подсистема обработки ЕЯО представляет собой ЯОИС, одной из отличительных особенностей которой от известных лингвистических процессоров является наличие языково-онтологической картины мира (ЯОКМ), которая, по сути, представляет собой онтологию для предметной области “компьютерная обработка ЕЯ”. Отметим, что подход к проектированию ЯОИС и ЯОКМ, их структурная и информационная модели рассмотрены в ряде работ [1, 5-9].

Подсистема полной онтологии ПрО представляет собой системную интеграцию онтологических знаний проблемной области и естественно-языкового обеспечения [10], что в общем случае представляется для первых – как композиция онтологии ПрО и одной из известных онтологий верхнего и среднего уровней (например, SUMO, онтология Дж. Совы или Mikrokosmos), а для вторых – композицией ЯОКМ и лингвистической онтологии для заданной ПрО. Приведём этапы и подэтапы проектирования полной онтологии ПрО, из названий которых следует смысл выполняемых работ на соответствующем этапе.

Этап составления технического задания (ТЗ) на проектирование полной онтологии ПрО включает: изучение и систематизацию начальных условий; определение перечня и краткое описание “активных” и “пассивных” процедур (эти понятия соответствуют введенным в [2] определениям “онтологоуправляемой” и “онтологознающей” ИС), что выполняются в блоке онтологии ПрО, в том числе процедуры системной интеграции в библиотеку онтологий; определение формальной теории представления онтологии ПрО на некотором языке *L*; определение схемы апробации разработанной онтологии ПрО.

Следующим, *основным этапом проектирования онтологии ПрО* является построение множеств $\{X\}$, $\{R\}$, $\{F\}$, $\{A\}$, представляющих кортеж известной схемы формальной модели онтологии ПрО. При этом систематизируются результаты предыдущего этапа, включающего, в том числе анализ, группирование и фиксацию знаний о предметной области *T*.

Обобщённая последовательность шагов алгоритма реализации основного этапа проектирования онтологии ПрО включает следующее.

Определение множеств $\{X\}=\{x_i\}, i=\overline{1,N}$, где *N* – множество концептов из *T*, и $\{A\}=\{a_j(d_l, r_{S_m})\}, j=\overline{1,J}, l=\overline{1,L'}, m=\overline{1,M}$, где $J, L'\ u\ \mathrm{M}$ - соответственно мощность множеств аксиом, определений и ограничений для *X*.

Определение множества $\{R\}=\{r_p\}, p=\overline{1,P}$, где P – мощность множества концептуальных отношений в *T* и их ранжирование по степени важности (т.е. построение кортежа $\langle r_1, r_2, \ldots, r_p \rangle$, где индекс 1 присвоен наиболее существенному концептуальному отношению в *T*, а p – наименее существенному).

Заметим, что формирование базовых наборов множеств $\{X\}$, $\{A\}$ и $\{R\}$ осуществляется оригинальными инструментальными средствами, названными соответственно “Concepts and Axioms Formation” и “Relations Formation”.

Разработка функционально-ориентированных компонентов онтологии ПрО.

3.1. Адаптация онтологии верхнего и среднего уровня к онтологии заданной ПрО.

3.2. Разработка лингвистической онтологии предметной области.

Определение множества $\{F\}=\{f_q\}, q=\overline{1,Q}$, где Q – мощность множества функций интерпретации, на кортеже $\langle X,R\rangle$, оптимальных для *T*.

Графическое проектирование компонентов онтологии у выбранной автоматизированной инструментальной среде и их системная интеграция в полную онтологию ПрО.

Создание файла формального описания разработанной онтологии ПрО.

Для данного этапа существенно не только разработать максимально полные множества, входящие в кортеж схемы формальной модели онтологии, но и обеспечить автоматическую проверку на целостность и непротиворечивость совокупности элементов упомянутых выше множеств.

Этап *апробации разработанной онтологии* полностью зависит от содержания подэтапа ТЗ, на котором разрабатывалась схема апробации для формальной модели. Теперь эту схему необходимо реализовать на реальном прототипе онтологии. В дополнение к указанной схеме необходимо выбрать из построенного множества функций интерпретации $\{F\}$ такие функции, которые позволили бы выполнить всестороннее тестирование. При этом особое значение имеют тестовые последовательности. По результатам тестирования разработчику необходимо рассмотреть вопрос о необходимости повторения некоторых шагов проектирования.

Подсистема обработки знаний (ПОЗ) в общем виде представляется структурой, представленной на рис.2.

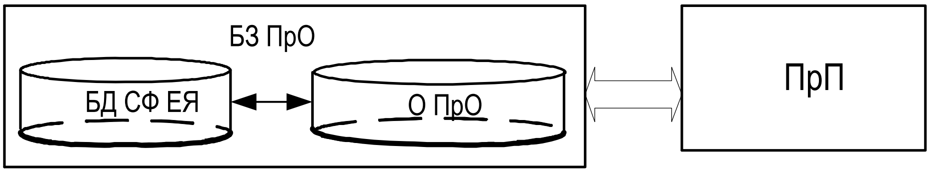


Рис.2. Обобщённое структурное представление ПОЗ

На рисунке приняты следующие обозначения:

БЗ ПрО – база знаний предметной области;

БД СФ ЕЯ – база данных суждений и фактов из заданной ПрО на естественном языке;

О ПрО – онтология ПрО;

ПрП – прикладной процессинг.

Особенности заданной ПрО полностью определяют подход к проектированию компонент подсистемы обработки знаний – БЗ ПрО и алгоритмов ПрП.

Для компоненты БЗ ПрО необходимо создать базу данных суждений и фактов и разработать алгоритмы её взаимодействия с онтологией ПрО, спроектированной в подсистеме полной онтологии ПрО. Указанное проектирование выполняется в соответствии с известными методами.

Проектирование компоненты ПрП в общем случае представляется разработкой совокупности алгоритмов, реализующих заданное множество задач пользователя, и их взаимодействия с БЗ ПрО и результатами распознанных знаний (полученными в подсистеме ЯОИС), содержащихся во входном ЕЯО. При этом аппаратная реализация указанных алгоритмов осуществляется в соответствии с методологией САПР ПЛИС, описанной в [3].

В заключение отметим, что полная онтология ПрО в процессе реализации цепочки технологий играет различную роль, а точнее – активны её различные компоненты.

При NLP–технологии основную роль играет внутриязыковое обеспечение ПрО – интеграция лингвистической онтологии из соответствующим фрагментом ЯОКМ. Оно поддерживает выполняемые ЯОИС процедуры анализа и распознавания знаний, содержащихся во входном ЕЯО.

При KR–технологии выполняется переход к формальному описанию знаний на некотором языке *L*, при котором активную роль играют онтологические знания ПрО, представленные онтографом. При этом формируются онтограф и интерпретационная структура, соответствующие знаниям, содержащимся во входном ЕЯО.

При KP–технологии активную роль играет база знаний ПрО, в состав которой входит формальное описание онтологических знаний и её аксиоматизация.

Литература

1. Палагин А.В., Яковлев Ю.С. Системная интеграция средств компьютерной техники. – Винница: «УНІВЕРСУМ-Вінниця», 2005. – 680 с.
2. Guarino N. Formal Ontology and Information Systems. In N. Guarino (ed.) Formal Ontology and Information Systems. //Proceedings of FOIS'98. - Trento, Italy. - 1998. - 6-8 June. – IOS Press, Amsterdam. – pp.3-15.
3. Реконфигурируемые вычислительные системы: Основы и приложения. / А.В. Палагин, В.Н. Опанасенко. – К.: Просвита, - 2006. – 280с.
4. Палагин А.В. К решению основной задачи эмуляции // УСиМ. – 1980. – №3. – С.24-28.
5. Палагин А.В. Организация и функции "языковой" картины мира в смысловой интерпретации ЕЯ - сообщений//Information Theories and Application. – 2000. – Vol. 7, №4. С.155-163.
6. Палагін О.В., Петренко М.Г. Модель категоріального рівня мовно-онтологічної картини світу //Математичні машини і системи. – 2006. - №3. - С.91-104.
7. Палагін О.В., Петренко М.Г. Архітектурно-онтологічні принципи розбудови інтелектуальних інформаційних систем //Математичні машини і системи. – 2006. - №4. - С.15-20.
8. Палагін О.В., Петренко М.Г. Розбудова абстрактної моделі мовно-онтологічної інформаційної системи //Математичні машини і системи. – 2007. - №1. - С.42-50.
9. Палагин А.В. Архитектура онтологоуправляемых компьютерных систем //Кибернетика и системный анализ. 2006 - №2. – С.111-124.
10. Палагин А.В., Петренко Н.Г. К вопросу системно-онтологической интеграции знаний предметной области //Математические машины и системы. – 2007. – №3,4. – С.63-75.

Информация об авторах

Палагин Александр Васильевич *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: petrng@ukr.net*

Петренко Николай Григорьевич *- Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: petrng@ukr.net*

МЕТОДЫ РЕШЕНИЯ ЛИНГВИСТИЧЕСКИХ ЗАДАЧ НА ОСНОВЕ ОНТОЛОГИЙ

Ольга Невзорова, Владимир Невзоров, Николай Пяткин

***Аннотация**: Онтолингвистические системы ориентированы на решение сложных задач обработки естественного языка, требующих семантических знаний. В основе проектирования онтолингвистических систем лежат процессы скоординированного взаимодействия онтологических и лингвистических моделей. В статье рассматриваются методы решения лингвистических задач на основе онтологий, разработанные при проектировании специализированной онтолингвистической системы «ЛоТА», предназначенной для анализа специальных технических текстов «Логика работы системы… ».*

***Ключевые слова**: онтолингвистические системы, онтологии, компьютерная лингвистика*

***ACM Classification Keywords**: I.2.7 Natural Language Processing*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

Повсеместная компьютеризация общества, становление и развитие сетевых информационных технологий способствовали переходу общества в новое качественное состояние глобальной информатизации. В этой связи активно стали развиваться информационные технологии обработки текстовой информации. Актуальными и востребованными являются технологии информационного поиска, извлечения знаний из текстов, автоматического реферирования, машинного перевода и др. В настоящее время исследования и разработки в области создания систем IE (Information Extraction) активно ведутся во всем мире [1-4]. Под извлечением информации (IE) понимается идентификация и семантическая классификация знаний, извлеченных из неструктурированных источников, таких как текст на ЕЯ, для задач информационных систем. В последние годы задача IE интегрируется в более крупные приложения, такие как выбор (поиск) информации для различных целевых задач, задача принятия решений. Широко развивается направление исследований, связанное с моделированием онтологической семантики. Особо следует отметить размеченные ролевыми дескрипторами онтологические ресурсы, такие как FrameNet, PropBank и VerbNet для английского языка; Salsa для немецкого языка; Spanish FrameNet для испанского языка. Все эти ресурсы являются необходимой базой для задач семантической классификации.

Важнейшая роль семантических знаний всегда подчеркивалась в когнитивных исследованиях, на основании которых можно утверждать, что семантический уровень является уровнем, связующим все языковые уровни, т.е. интегральным системообразующим свойством языковой системы.

Онтолингвистические системы ориентированы на решение сложных задач обработки естественного языка, требующих семантических знаний. В основе проектирования онтолингвистических систем лежат процессы скоординированного взаимодействия процессы скоординированного взаимодействия онтологических и лингвистических моделей. Целью создания онтолингвистических систем является обеспечение решения сложных задач обработки естественного языка путем организации системы взаимодействий различных языковых уровней, включая построение адекватной модели предметной области, исследование свойств объектов предметной области и зависимостей между объектами. Моделирование предметной области осуществляется на основе онтологического подхода, интегрирующего знания экспертов и лингвистические знания.

Класс онтолингвистических систем отличается объединением экстралингвистических (онтологических) и лингвистических знаний, эвристических и формальных методов обработки ЕЯ. При этом роль и значение эвристических методов возрастает по мере возрастания сложности рассматриваемых лингвистических моделей.

Модель решения прикладной задачи в онтолингвистической системе

При проектировании лингвистических приложений предлагается использовать новый подход, центральной идеей которого является построение решения прикладной задачи на основе организации взаимодействия полифункциональной онтологической системы: прикладной онтологии, онтологии свойств и онтологии задач. Основная идея нового подхода заключается в следующем. Объектами лингвистического анализа являются текстовые документы, обработка которых производится для определенной целевой задачи. Тексты описывают совокупность объектов прикладной области, обладающих определенным набором свойств, важных для конкретной целевой ситуации. Иные целевые ситуации могут потребовать задания объектов с другим набором свойств. Другими словами, в разных задачах (практических целевых ситуациях субъектов) ***объекты обладают разным набором свойств***, не только в разных проблемных областях, но и в одной и той же проблемной области, в которой решаются разные задачи.

Задача как некоторый ***тип*** практической ситуации субъектов в большей степени определяется их способом существования (структурой мышления и пр.), чем конкретной проблемной областью. Таким образом, можно предположить, что подмножество языка, маркирующее структуру событий, связанных с задачами, квазинезависимо от конкретной проблемной ситуации и соответствующей ей структуры свойств объектов, то есть выделяемо в отдельную **онтологию задач**. Тем самым, можно выделить некоторое универсальное (пополняемое) множество базовых задач (типовых элементарных ситуаций), на основе которых можно с помощью определенной логики последовательностей конструировать более сложные задачи. Таким образом, решение прикладной задачи может быть спроектировано как **система взаимодействий трех онтологий**: прикладной онтологии проблемной области, онтологии свойств и онтологии базовых задач. Для каждой онтологии формируются свои концепты, совокупность текстовых входов концептов и связи между концептами, базирующиеся на ключевых для данной онтологии отношениях. При этом взаимодействие онтологий реализуется в разметке концептов прикладной онтологии концептами-свойствами для конкретных концептов-задач.

Онтология задач на уровне файловых представлений должна быть унифицирована с онтологиями свойств и прикладной онтологии. Выделяются следующие типы концептов онтологии задач: *задачи*, *операции*, *данные* (входные/выходные). Метод построения спецификаций прикладной задачи должен быть реализован как процессор (интерпретатор) со всеми свойствами программируемой среды, который настраивается на конкретный концепт-задачу и последовательно реализует базовые операции этой задачи. Соответствующая инструментальная среда должна быть выстроена как набор специализированных и универсальных базовых операций, управляющих процессом решения. Таким образом, любая задача, решаемая процессором, представляет собой концепт онтологии задач, связанный с другими концептами связями "принадлежности-следования". Онтология задач может быть связана с онтологией свойств через механизмы конкретизации параметров концептов-данных и значений метрик отношений, определенных на онтологии.

При проектировании технологии взаимодействия полифункциональной системы онтологических моделей необходимо обеспечить решение следующих основных задач:

- реализацию операций разметки концептов прикладной онтологии концептами-свойствами для конкретных концептов-задач;
- разработку механизма взаимодействия компонентов онтологической системы;
- разработку механизмов контроля целостности онтологической системы.

Архитектура онтолингвистической системы

В структуре онтолингвистической системы выделяются две основные взаимодействующие компоненты: онтологическая и лингвистическая.

Онтологическая компонента позволяет поддерживать онтологическое моделирование предметной области. Разработка онтологической компоненты может опираться на существующие стандарты разработки онтологий и тезаурусных систем, а также иметь специфические методы.

Функционирование онтологической системы обеспечивает поддержку решения следующих задач:

- семантическая разметка исследуемых текстов элементами (концептами, отношениями) онтологической системы;
- извлечение информации из текстов (распознавание и интерпретация концептуальных структур);
- онтологическая поддержка задач лингвистического анализа:
  = разрешение грамматической и лексической многозначности;
  = сегментация внутри предложения (частичный синтаксический анализ);
  = разрешение референции и восстановление эллипсиса
- поддержка онтологических выводов.

Лингвистическая компонента обеспечивает поддержку решения следующих лингвистических задач:

- распознавание символов (графематический анализ);
- сегментация предложений;
- распознавание типов лингвистических объектов (словоформы, числа, дата, время, аббревиатура и т.п.);
- морфологических анализ словоформ;
- разрешение грамматической и лексической многозначности;
- синтаксический анализ и разрешение синтаксической многозначности;
- разрешение референции и восстановление эллипсиса.

Методы решения лингвистических задач

Рассматриваемый подход реализован в проектировании онтолингвистической системы «ЛоТА», предназначенной для анализа специализированных текстов типа "Логика …" [Невзорова&Федунов, 2001].

Основной задачей системы "ЛоТА" является извлечение из специализированного технического текста информационной модели схемы бортовых алгоритмов, решающих определенную задачу в определенной проблемной ситуации, и контроль структурной и информационной целостности выделенной алгоритмической схемы.

Решение основной задачи обеспечивается комплексом технологий обработки текстов:

- технологии морфосинтаксического анализа;
- технологии семантико-синтаксического анализа;
- технологии взаимодействия с прикладной онтологией.

Указанная сумма технологий формируется на основе центрального ядра – прикладной онтологии (авиаонтологии), обеспечивающей согласованное взаимодействие различных программных модулей. Авиаонтология концептуально представляет предметную область информационного (алгоритмического) обеспечения различных полетных режимов антропоцентрических систем [Добров и др., 2004].

Разрабатываемая программная система содержит типовой набор компонентов онтолингвистической системы. Основное внимание при данном подходе уделяется разработке механизмов совместного взаимодействия компонентов при решении конкретных задач обработки текста.

Программный комплекс состоит из двух взаимодействующих подсистем: подсистемы лингвистического анализа технических текстов "Анализатор", подсистемы управления и ведения онтологии "OntoIntegrator". Взаимодействие подсистем реализовано на базе технологии "клиент-сервер", причем в различных подзадачах подсистемы выступают в различных режимах (режим сервера или режим клиента) [Невзорова, 2006].

В рамках развиваемого подхода разработаны методы решения различных лингвистических задач:

- задача построения лингвистической оболочки онтологии;
- задача построения индексированной базы контекстов омонимов;
- задача разрешения многозначности;
- задача онтологической разметки текста;
- задача сегментации текста.

Метод построения лингвистической оболочки онтологии обеспечивает загрузку прикладной онтологии в специальную лингвистическую оболочку для последующего ее использования в задачах обработки текстов.

Интегрированная программная технология построения индекса базы контекстов омонимов различных типов (функциональных, лексических) включает модули создания и ведения индекса омонимов, модуль согласования индексной базы с основным лингвистическим ресурсом – грамматическим словарем, а механизмы выполнения внешних запросов по разрешению (поиску) типовых омонимических контекстов в текстовом корпусе на основе индекса омонимов.

Интегрированная программная технология разрешения многозначности является комплексной технологией, объединяющей три разработанные программные технологии. Первая технология - технология разрешения функциональной омонимии на основе контекстных правил. Метод разрешения многозначности на основе контекстных правил позволяет разрешать функциональную (грамматическую омонимию) на основе контекстных правил, которые формулируются как результат тщательной лингвистической экспертизы поведения омонима в современных корпусах русского языка. В настоящее время разработано свыше 40 обобщенных правил наиболее частотных типов функциональных омонимов, в том числе правила для сложных случаев типа разрешения (например, для омонимов *это, все/всё* и др.).

Вторая технология разрешения омонимии базируется на использовании индексируемой базы контекстов омонимов. Этот метод позволяет эффективно разрешать как функциональную, так и лексическую омонимию. Механизмы разрешения основаны на распознавании контекстов омонимов во входных предложениях. Модель контекста омонима имеет ряд распознаваемых параметров (грамматические характеристики компонентов коллокации, расстояние до разрешающей словоформы), при обнаружении которых выдается информация о типе омонима и его грамматических характеристиках.

Третья технология разрешения омонимии использует лингвистическую оболочку онтологии, т.е. грамматическую информацию об онтологических концептах и их текстовых (синонимических) формах.

Интегральный метод разрешения омонимии реализует весь комплекс перечисленный выше технологий. Первоначально осуществляется поиск в базе контекстов омонимов, при отсутствии необходимой информации о разрешении омонимии запускаются процедуры разрешения на основе контекстных правил.

Метод онтологической разметки текста распознает в тексте онтологические концепты. Реализация метода базируется на специальных протоколах обмена между подсистемой "Analyzer" (клиент) и подсистемой "OntoIntegrator" (сервер). Распознавание линейных онтологических входов в тексте осуществляется на основе грамматических описаний, заданных в лингвистической оболочке онтологии. Новые результаты получены при разработке методов распознавания онтологических единиц, подвергшихся сочинительному сокращению в тексте. При анализе сочиненных синтаксических конструкций определенных типов решается обратная задача выделения потенциальных составляющих конструкции и их распознавание как самостоятельных онтологических единиц. На основе разработанных механизмов в тексте распознаются синтаксические конструкции с однородными членами, а также некоторые типы симметричных конструкций. Например, в синтаксической конструкции "*атаки пар и звеньев истребителей"* выделяются составляющие "*атаки пар истребителей*" и "*атаки звеньев истребителей*", которые распознаются как отдельные онтологические единицы.

Решение обратной задачи (выделение составляющих) не всегда является однозначным. Например, в сочинительной конструкции "*прикрытие бомбардировщиков и штурмовиков в районе боевых действий*" выделяются составляющие "*прикрытие бомбардировщиков в районе боевых действий*" и "*прикрытие штурмовиков в районе боевых действий*", однако в других случаях предложно-падежная группа (типа "*в районе боевых действий"*) может не являться общим элементом составляющих. Выделение составляющих из сочинительных конструкций производится на основе специальных правил, которые учитывают явление "семантической однородности". Семантическая однородность предполагает построение синтаксических конструкций с семантически однородными членами, т.е. члены однородных конструкций должны относится к одному семантическому классу. На этапе построения правил выделяются два основных семантических класса: класс предметных и непредметных имен. Семантическая однородность допускает построение синтаксических конструкций либо для предметных, либо для непредметных сущностей. Например, допустимыми являются конструкции типа "*самолеты и ракеты противника*" (предметная однородность), либо "*перехват и уничтожение противника*" (непредметная однородность).

Синтаксические конструкции с однородными определениями составляют другой тип синтаксического сокращения. В этом случае выделяется группа составляющих с одиночными определениями. Так, например, однородная синтаксическая группа типа "*естественные и искусственные помехи*" распознается, как состоящая из элементов "*естественные помехи*" и "*искусственные помехи*".

Все составляющие сложных синтаксических конструкций затем отождествляются как онтологические входы. С каждым распознанным в тексте онтологическим входом передается информация об онтологическом концепте и его семантическом классе (концепте верхнего уровня по иерархии). Метод распознает различные ситуации распределения онтологических входов в предложении. При вложении сегментов как результат передается сегмент максимальной длины, при перекрытии сегментов передаются все перекрывающиеся составляющие.

Задача сегментации текста на составляющие – сегменты является одной из ключевых в процессе анализа текста. Результатом сегментации предложения является иерархическая совокупность семантико-синтаксических сегментов. Выделенные сегменты являются "блоками", из которых собираются по тексту информационные модели, определяемые типом решаемой задачи. Сегмент имеет синтаксический и семантический тип. Выделяются два главных семантических типа сегментов: группа субъекта и группа предиката предложения. С главными семантическими типами связаны подмножества синтаксических типов. Синтаксический тип определяет синтаксическую модель сегмента. Выделенное подмножество синтаксических моделей исчисляет синтаксические шаблоны для главных семантических типов. Распределение главных семантических типов по синтаксическим моделям фактически задает различные синтаксические структуры предложений русского языка. Текущая версия модуля сегментации поддерживает сегментацию базовых классов моделей русского предложения, а именно полных (двусоставных) предложений с группой субъекта в форме N*им/Abb (сущ./местоименное сущ. в именит. падеже или аббревиатура) и глагольным предикатом. В этих ограничениях количественные оценки синтаксических моделей для группы субъекта – 4 подтипа, для глагольной группы – 11 подтипов (простой глагольный, осложненный частицей, составной глагольный, именной составной). Помимо главных семантических типов выделяются семантические сегменты с дополнительной семантической характеристикой – атрибутивные сегменты. Атрибутивные сегменты имеют собственное подмножество синтаксических моделей. Алгоритмы сегментации на основе синтаксических моделей выделяют главные семантические сегменты и их расширения в виде атрибутивных сегментов. Сегменты, не вошедшие в расширенные модели главные сегментов, интерпретируются как атрибуция предложения в целом (например, сегменты - локативы или сегменты - темпоративы).

Задача сегментации решается совместно с задачей онтологической разметки. Построение сегментов осуществляется в границах распознанных онтологических входов. Процесс организации взаимодействия подсистем при решении задач сегментации и онтологической разметки приведен на рис. 2.

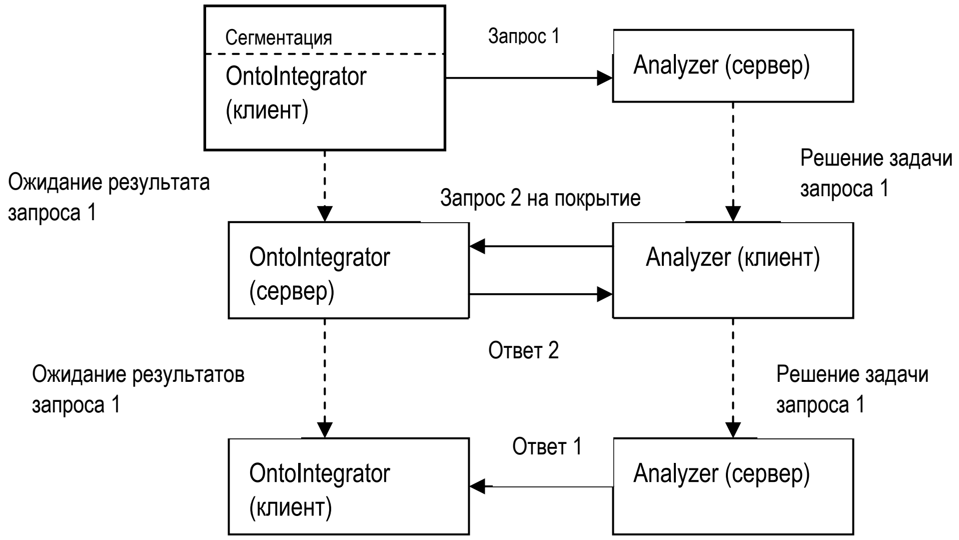
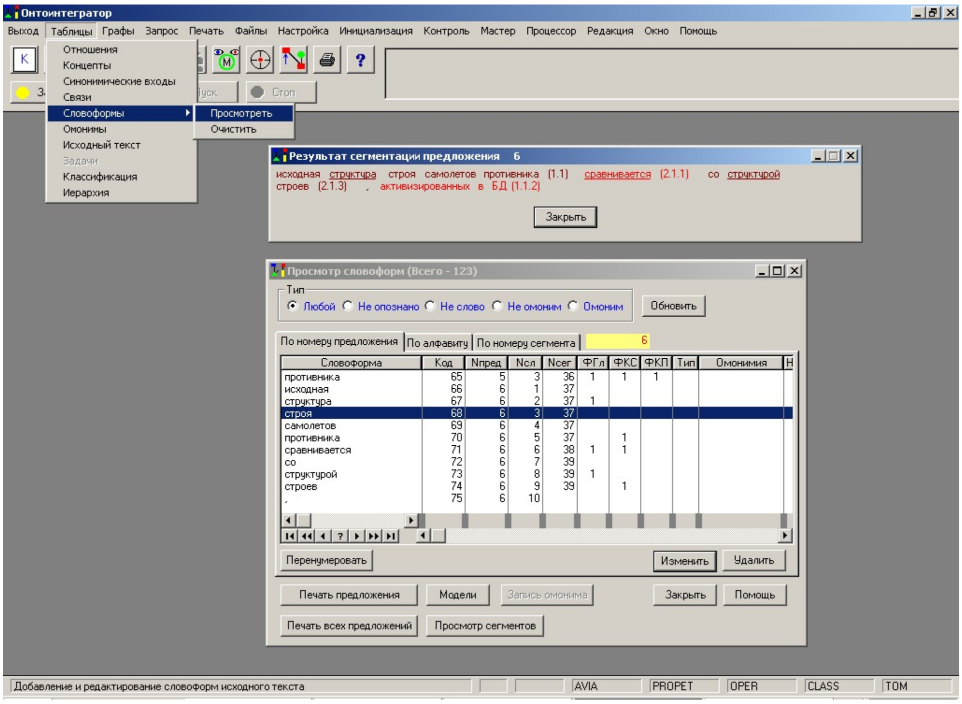


Рис.2. Взаимодействие подсистем при решении задачи сегментации

Запрос на решение задачи сегментации передается от подсистемы OntoIntegrator к системе Analyzer. Для решения этой задачи подсистема Analyzer запрашивает у подсистемы OntoIntegrator информацию об онтологической разметке текста. Полученная информация используется для уточнения границ сегментов.



Результат сегментации предложения 6

исходная структура строя самолетов противника (1.1) сравнивается (2.1.1) со структурой строев (2.1.3) , активизированных в БД (1.1.2)

Просмотр словоформ (Всего - 123)

Тип: ● Любой ○ Не опознано ○ Не слово ○ Не омоним ○ Омоним

Словоформа	Код	Nпред	Nсл	Nсег	ФГл	ФКС	ФКП	Тип	Омонимия	Н
противника	65	5	3	36	1	1	1			
исходная	66	6	1	37						
структура	67	6	2	37	1					
строя	68	6	3	37						
самолетов	69	6	4	37						
противника	70	6	5	37		1				
сравнивается	71	6	6	38	1	1				
со	72	6	7	39						
структурой	73	6	8	39	1					
строев	74	6	9	39		1				
,	75	6	10							

Добавление и редактирование словоформ исходного текста

Рис. 3. Вывод результатов сегментации предложения

На рис. 3. показан результат сегментации предложения "(*Исходная структура строя самолетов противника) (сравнивается) (со структурой строев),(активизированных в БД.)".* Построенные сегменты заключены в круглые скобки, каждому сегменту приписан семантический тип. Если сегмент содержит в своих границах онтологический концепт, то информация о семантическом классе онтологического концепта учитывается при определении семантического типа сегмента.

Заключение

Класс онтолингвистических систем отличается объединением экстралингвистических (онтологических) и лингвистических знаний, эвристических и формальных методов обработки ЕЯ. Ядром онтолингвистических систем являются знания различной природы, в том числе различные онтологии, представляющие прикладные знания, метазнания, в том числе знания о прикладных задачах и их свойствах.

Основные лингвистические задачи решаются через взаимодействия онтологической и лингвистической компонент онтолингвистической системы, при этом общая структура решаемой задачи может динамически меняться через специальные механизмы настройки типа решаемой задачи.

Благодарности

Данная работа выполнена при поддержке Российского Фонда Фундаментальных исследований, грант № 08-07-90407.

Литература

[Невзорова&Федунов, 2001] Невзорова О.А., Федунов Б.Е. Система анализа технических текстов "ЛоТА": основные концепции и проектные решения // Изв. РАН. Теория и системы управления. 2001. № 3. С. 138-149.

[Добров и др., 2004] Добров Б.В., Лукашевич Н.В., Невзорова О.А., Федунов Б.Е. Методы и средства автоматизированного проектирования прикладной онтологии // Известия РАН. Теория и системы управления. М.: 2004. № 2. С. 58-68.

[Невзорова, 2006] Невзорова О.А. Подход к разработке методов автоматизированного контроля информационной целостности технических текстов //Труды десятой национальной конференции по искусственному интеллекту КИИ-2006. Том 2. М.,Физматлит, 2006. С. 564-571.

Authors' Information

***Ольга Невзорова** – НИИММ им. Н.Г. Чеботарева, Татарский государственный гуманитарно-педагогический университет, Казань, Россия; e-mail: olga.nevzorova@ksu.ru*

***Владимир Невзоров** – Казанский государственный технический университет им. А.Н. Туполева, Россия; e-mail: nevzorov@mi.ru*

***Николай Пяткин** – НИИММ им. Н.Г. Чеботарева, Казань. Россия; e-mail: nikolaip@mail.ru*

СЕМАНТИЧЕСКАЯ ВИКИПЕДИЯ КАК ИСТОЧНИК ОНТОЛОГИЙ ДЛЯ ИНТЕЛЛЕКТУАЛЬНЫХ ПОИСКОВЫХ СИСТЕМ

Анатолий Гладун, Юлия Рогушина

***Аннотация**: Определены подходы к интеллектуальному поиску информации при помощи современных Web-технологий. Проанализированы источники онтологических описаний предметных областей поиска, в частности, семантическая Википедия. Предложены методы использования онтологий для повышения пертинентности информационного поиска.*

***Ключевые слова**: Semantic Web, Википедия, онтология, тезаурус, поиск информации.*

***ACM Classification Keywords**: I.2.4 Knowledge Representation Formalisms and Methods*

***Conference**: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

В информационном обществе приоритетным направлением является создание и эффективное использование знаний и информационных ресурсов (ИР). Превращение World Wide Web в универсальный источник информации и знаний с неизбежностью приводит к появлению новых технологий работы с контентом. Его дальнейшее развитие многие ученые связывают с интеллектуализацией и интеграцией всех существующих ИР на семантическом уровне. В современных информационно-поисковых системах (ИПС) реализованы лишь некоторые элементы интеллектуальности. Более удобный доступ к нужным ИР обеспечивают он-лайновые энциклопедии, созданные в процессе коллективного сбора и анализа информации, например, Википедия. Они не рассматриваются как альтернатива ИПС, но дополняют их.

Технология Wiki и ее основные характеристики

Wiki - это технология построения Web-сайта, позволяющая пользователям через Web-интерфейс активно участвовать в процессе редактирования его контента - исправлении ошибок и добавлении новых материалов. Wiki-технология не требует использования специальных программ, регистрации на сервере и знания HTML. Каждая страница обычно содержит большое количество гиперссылок на другие страницы.

Технология Wiki позволяет аккумулировать знания человечества, представляя их в электронной интероперабельной форме, обеспечить навигацию по этой базе знаний и средства ее актуализации. При этом использовать Wiki могут сообщества различного объема и тематической направленности, создавая при этом базы знаний различного масштаба - от глобальных Википедий и электронных энциклопедий крупных корпораций до легко обновляемых справочных систем небольших организаций, предприятий и учебных заведений. Основные характеристики Wiki:

1) Количество авторов соизмеримо с количеством пользователей Wiki-ресурсов;

2) Обеспечивается поддержка многопользовательской работы;

3) Имеется возможность многократного редактирования текста с помощью самой Wiki-среды (Web-сайта);

4) Появление изменений сразу же после их внесения;

5) Каждая статья Wiki имеет собственное уникальное имя;

6) Язык разметки достаточно прост и не требует специальных знаний;

7) Существует возможность возвратиться к предшествующей версии.

На базе Wiki-технологии в 2001 году была создана Википедия - многоязычная общедоступная свободно распространяемая энциклопедия, публикуемая в Интернете. Эта «общественная» энциклопедия является ярким примером стремительного роста и накопления знаний в распределенной среде Интернет.

Сегодня начала формироваться тенденция использования технологий Wiki для формирования корпоративных баз знаний. Это особенно актуально для новых и быстро развивающихся областей, где информационные ресурсы Интернет часто оказываются устаревшими или не согласованными в области терминологии. Если предприятие считает определенные разработки своей интеллектуальной собственностью либо еще недостаточно разработанными для широкой публикации, то можно использовать один из свободно распространяемых Wiki-движков и сформировать корпоративную энциклопедию (развернутый тезаурус), охватывающую определенную предметную область (ПрО). Пополнение такой энциклопедии станет правом и обязанностью сотрудников данного предприятия и других привлеченных ими лиц.

Проект Semantic Web и его влияние на развитие интеллектуальных Wiki-технологий

Целью Semantic Web является преобразование всей совокупности информационных ресурсов Web в единую базу знаний, пользоваться которой могут как люди, так и программы. Для этого необходимо снабдить каждый ИР описанием его семантики и предоставить средства для автоматизированной обработки этих описаний и представления знаний о них.

Автором этой концепции является Т.Бернес-Ли, который ранее задумал и разработал Web, а теперь возглавляет Консорциум W3. Целью Бернеса-Ли было сформировать информационное пространство, к которому каждый имеет непосредственный и интуитивный доступ, и не только для просмотра, но и для создания информации. Машины становятся способными анализировать все данные в Web – контент, связи и транзакции между людьми и компьютерами. Эта концепция была принята и продвигается Консорциумом W3 – лидер в развитии технологий для Web (многие из основополагающих технологий, таких, как XML и RDF [1], были разработаны именно W3C). Для её внедрения предполагается создание сети документов, содержащих метаданные о ресурсах Web. Тогда как сами ресурсы предназначены для восприятия человеком, метаданные используются машинами (поисковыми роботами и другими интеллектуальными агентами). Согласно определению, данному Meta Data Coalition (http: //www.mdcinfo.com) в документе "Open Information Model", метаданные — это описательная информация о структуре и смысле данных, а также приложений и процессов, которые манипулируют данными. Росту популярности и широкому внедрению технологий Semantic Web способствует стандартизация консорциумом W3C синтаксической и семантической разметки электронных документов, особенно технологий XML, RDF/RDFS и OWL, поддерживающих синтаксическую и семантическую совместимость.

В модели RDF документ рассматривается как частично упорядоченный набор абстрактных объектов, обладающие свойствами (атрибутами) и имеющими уникальный идентификатор. Любой объект при своем создании получает генерируемый системой уникальный идентификатор, который связан с объектом во все время его существования и не меняется при изменении состояния объекта. RDF позволяет определять произвольные объекты в документе. Атрибуты (имена и значения) должны выбираться из словарей, связанных с теми или иными предметными областями. Формально RDF не накладывает никаких ограничений на значения атрибутов объектов, перекладывая создание соответствующих словарей на заинтересованные организации. Основной словарь имен объектов системы создан на основе словарей стандартных схем метаданных.

Три технологии составляют основу Semantic Web: 1. программные агенты – для того, чтобы представлять реальные объекты и автоматизированную разрешающую способность задачи от имени их владельца [2], 2. онтологии – для семантического расширения информации, которой обмениваются и обрабатывают Web-приложения [3], 3. Web-сервисы как вычислительные средства, доступные через Интернет [4].

Центральным компонентом концепции является применение онтологий. Онтологии разрабатываются и могут быть использованы при решении различных задач, в том числе для совместного применения людьми или программными агентами, для возможности накопления и повторного использования знаний в предметной области, для создания моделей и программ, оперирующих онтологиями, а не жестко заданными структурами данных, для анализа знаний в предметной области [5]. Использование технологий Semantic Web при разработке программных систем позволяет существенно упростить проблему совместимости систем из смежных областей и является первым шагом к построению высокоинтеллектуальных компонентов и агентов.

Semantic Web открывает доступ к чётко структурированной информации для любых приложений, независимо от платформы и независимо от языков программирования. Программы смогут сами находить нужные ресурсы, обрабатывать информацию, классифицировать данные, выявлять логические связи, делать выводы и даже принимать решения на основе этих выводов. Однако сегодня не совсем понятно, какие программные продукты можно называть приложениями Semantic Web.

Характерные свойства приложений Semantic Web

В проекте "Semantic Web Challenge" рассматриваются перспективы применения технологий Semantic Web для разработки прикладных программ, которые обрабатывают информацию с учетом ее семантики.

Приложением Semantic Web называют программный продукт, который отвечает следующим минимальным требованиям [6]:

1. Семантика данных играет основную роль: 1.1. Семантика данных должна представляться с использованием формальных определений; 1.2. Данные должны обрабатываться нетривиальными способами с целью получения полезной информации; 1.3. Обработка семантической информации должна играть центральную роль в достижении результатов, которые нельзя получить другим путем

2. ИР, используемые в приложении, должны: 2.1. иметь различных владельцев (т.е отсутствует возможность контроля ее изменения); 2.3. быть гетерогенными (синтаксически, структурно и семантически); 2.3. содержать данные реального мира (а не быть игрушечными примерами).

3. Поиск осуществляется в реальном информационном пространстве Web

Необходимо, чтобы все приложения воспринимали открытую внешнюю среду, т.е. учитывали, что информация о ней никогда не бывает полной.

Предполагается, что приложения используют каким-либо образом RDF, RDF Schema или OWL, хотя это и не является обязательным условием. Наиболее важно, что, если используется семантическая технология, то она играет ключевую роль в достижении новых уровней функциональности или представления.

Кроме этих минимальных требований, можно сформулировать еще ряд пожеланий:

- Демонстрируются преимущества семантических технологий или полученных результатов
- Приложение должно быть масштабируемым (в терминах объема используемых данных и в терминах распределенных компонентов, работающих вместе)
- Функциональность приложения отличается от обычного информационного поиска
- Приложение имеет явный коммерческий потенциал
- Контекстная информация используется для упорядочения рейтинга или порядка
- Обрабатываются мультимедийные документы
- Используются динамические данные, возможно в комбинации со статичной информацией
- Поддерживаются различные языки

Семантическая Википедия

Перспективы развития Wiki-технологий связаны с их интеллектуализацией, т.е. с переходом от обычных гиперссылок к системе семантической разметки контента на основе метаданных – семантической Wiki. Проект Semantic Web оказал большое воздействие на развитие технологий Wiki. Семантическая Wiki - расширение технологии Wiki, использующее модель знаний, которая позволяет указывать тип ссылок между статьями, типы внутри статей, а также метаданные о страницах. Цель семантической Wiki - автоматизировать обработку сведений, содержащихся в Википедии, и генерировать выделение информации по запросам пользователей. В этом расширении для запросов используется язык SPARQL. Система Semantic MediaWiki написана с помощью механизма расширений MediaWiki. Построенное на машино-понятном языке, это расширение позволяет семантически обрабатывать Wiki-контент, предоставляя пользователю возможность для добавления семантической разметки информации. Для запросов используется язык SPARQL. Система Semantic MediaWiki написана с помощью механизма расширений MediaWiki. Это упрощает интеграцию в существующие приложения MediaWiki.

Семантическая Википедия предоставит следующие элементы для разметки статей: категории, типизированные ссылки и атрибуты — свойства содержимого статей. Категории классифицируют статьи семантической Википедии в соответствии с их контентом, как и в обычной Википедии. Например, статья «Статистика» относится к категории «Науки Категоризация — это процесс структурирования схожих статей, но наличие хотя бы одной категории в статье обязательно.

Типизированные ссылки представляют собой тройки RDF, состоящие из субъекта, отношения и объекта. Например, в тройке Киев:[[столица :: Украина]] субъект – имя страницы «Киев», отношение – «столица» и имя страницы «Украина» - объект. Типизированные ссылки позволяют выполнять прямой запрос («Какой город является столицей Украины?»), логический вывод («Киев — столица Украины») => («Киев находится в Украине») и агрегирование поисковых критериев в запросе («Киев — столица Украины», «Украина - государство в Европе) => («Киев — европейская столица»). Для создания триплетов вводится новое пространство имён Relation:, позволяющее получить список известных троек (по аналогии с пространством Категория: в Википедии). Обработка этих троек состоит из извлечения типизированных ссылок из текста статьи, их преобразования в RDF-тройки и обновления соответствующей БД.

Атрибуты описывают свойства объекта статьи семантической Википедии. Например, можно указать численность населения Украины или ее площадь. При обработке значений атрибутов нужно распознавать используемые единицы измерения. Поэтому с переменными, используемыми в качестве атрибутов объекта, связывают не только определенный тип данных (например, «целое», «текстовая строка»), но и их семантику (скорость в км/ч или м/с). Это позволяет преобразовывать значения, использующие различные единицы измерения. Для этого в расширение Semantic Wikipedia встроен преобразователь для популярных единиц измерения. Для хранения значений атрибутов также используются RDF-тройки.

Семантическая разметка позволяет значительно упростить всю структуру Вики, помогает пользователям быстрее находить нужную информацию.

Основные преимущества Семантической Википедии

1. Значительная часть информации в Википедии может быть представлена в виде списков, исходными данными для которых являются различные статьи. При обновлении таких статей Семантическая Википедия автоматически обновляет все производные от нее списки и результирующие данные. Например, если появились новые сведения о населении какого-либо города, то автоматически обновится информация о населении страны, к которой относится этот город. Запросы, по которым строятся сами списки, должны формироваться вначале пользователями, но поддержание их в актуальном состоянии - проблема, решаемая самой Википедией. Такие списки всегда актуальны и их легко настроить для получения дополнительной информации.

2. Использование метаданных позволяет лучше структурировать информацию. Например, если некий объект, описываемый в википедии, классифицируется как "Человек", то можно описывать такие его атрибуты, как "Имя" и "Профессия", а затем использовать эти атрибуты и их значения для поиска и формирования новых списков, например, найти всех людей, упомянутых в Википедии, день рождения которых совпадает с текущей датой.

3. Метаданные позволяют установить связи между аналогичными статьями на различных языках, что позволит актуализировать данные и находить несоответствия между ними. Например, если мы введем последние сведения о населении Киева в украинскую Википедию, то автоматически обновятся соответствующие статьи во всех Википедиях, где упоминается это число.

Таким образом, Википедия превращается из простого хранилища данных в распределенную базу знаний.

Постановка задачи

Сегодня основная проблема, возникающая при поиске информации в Интернете, связана с фильтрацией полученных результатов и отбором тех информационных ресурсов, которые соответствуют реальным информационным потребностям пользователя. Для такого отбора необходимо формализовать представления пользователя об интересующей его ПрО и разработать средства автоматизированного формирования соответствующей базы знаний. Для описания ПрО широко применяют онтологический подход. Однако создание онтологий – сложный и трудоемкий процесс. Поэтому предлагается использовать в поисковых системах внешние онтологии, сформированные различными приложениями

Semantic Web, в частности, семантической Википедией. Их можно использовать для создания тезаурусов пользователей и ИР, а затем на основании этих тезаурусов переупорядочивать результаты поиска.

Интеллектуальный поиск в Интернет

Информационный поиск - совокупность операций, необходимых для нахождения информации, которая удовлетворяет потребностям пользователя, выраженным в виде запроса. Запрос к ИПС является попыткой пользователя (не всегда удачной) формализовать свою информационную потребность. Традиционные механизмы поиска в Интернет, как правило, рассматривают запросы пользователя на поиск информации изолировано друг от друга и не учитывают полученные ранее результаты. Эффективный поиск информации в Интернет по мере увеличения ее объема и рассредоточения ее источников становится все более сложным и трудоемким. При этом критичным является не столько время поиска, сколько отбор информации, релевантной запросу пользователя. Под термином «релевантность» (от англ. relevancy - уместность) понимается формальное соответствие полученной в результате поиска информации запросу.

Однако для пользователя важнее другой параметр оценки качества функционирования ИПС – пертинентность, т.е. неформальное соответствие полученных результатов реальным информационным потребностям пользователя. Для ее повышения надо использовать не только сведения, содержащиеся в запросе, но и дополнительную информацию о пользователе и его предпочтениях.

Значительно повысить пертинентность поиска позволяет его персонификация, т.е. использование такого механизма поиска, которому использует сведения о предыдущих запросах пользователя и сфере его интересов. Такой персонифицированный поисковый механизм может размещаться как на сервере, так и на клиенте. Например, серверный механизм поиска Google способен отслеживать предыдущие запросы пользователя и выбранные им документы, а затем на основе этой информации сделать вывод о сфере его интересов. Но из-за того, что затраты на работу полномасштабного механизма поиска очень высоки, полномасштабная персонификация на сервере сейчас обходится слишком дорого для основных механизмов поиска в Web. Чтобы определить пертинентность ответа на запрос, надо, чтобы ИПС каким-то образом моделировала ПрО, соответствующую информационным потребностям этого пользователя. Один из перспективных подходов к моделированию ПрО базируется на онтологическом анализе [7].

Онтология - система, состоящая из набора понятий и набора утверждений об этих понятиях, на основе которых можно строить классы, объекты, отношения, функции и теории. Онтология, как пример общего соглашения о семантике ПрО, способствует установлению корректных связей между значениями элементов этой ПрО, тем самым создавая условия для их совместного использования. При этом возникают два основных вопроса: 1. Откуда брать такие онтологии (или как их формировать и изменять); 2. Как (по каким алгоритмам) сравнивать онтологии и извлекать содержащиеся в них знания о ПрО поиска.

Для представления онтологий разработан язык OWL, а также создан ряд инструментальных программных продуктов, позволяющих оперировать с онтологиями, представленными на OWL (редактировать их, визуализировать, объединять и т.п.) [8]. Чтобы описать свои информационные интересы, пользователь может сослаться на интересующие его страницы Семантической Википедии (и соответствующие им онтологии). ИПС анализирует такие наборы ссылок и извлечь из них онтологическую информацию – например, одну или несколько онтологий в формате OWL, наборы терминов этих онтологий, метаописания ИР в формате RDF и т.д. – и на основе этих сведений сформировать тезаурус пользователя.

Методы использования онтологий для построения поисковых тезаурусов

Наряду с использованием онтологий представляется целесообразным использовать для моделирования знаний пользователя о ПрО поиска частный случай онтологии – тезаурус, построение которого относительно проще. До недавнего времени термины онтология и тезаурус использовались как синонимы, однако, теперь в ИТ тезаурус чаще применяют для описания лексики в проекции на семантику, а онтологию - для моделирования семантики и прагматики в проекции на язык представления [9].

Тезаурус – это $Ts = \langle T, R \rangle$, где T - множество терминов, а R – множество отношений между этими терминами. Множества T и R конечны. Для того, чтобы отфильтровать результаты работы внешней ИПС и получить только те ИР, которые пертинентны информационным потребностям пользователя, необходимо предварительно сформировать тезаурус ПрО, интересующей пользователя, и тезаурусы этих ИР, а затем сравнить эти тезаурусы. Построить тезаурус для ИР несложно – задача выполняется автоматически на основе лексического анализа соответствующего текста.

Сложнее построить тезаурус ПрО, интересующей пользователя. Будем считать, что тезаурус ПрО – это совокупность терминов, знакомых пользователю ИПС. Это термины, содержащиеся в ИР, которые были найдены ранее по запросам пользователя и были признаны им, как относящиеся к этой ПрО. Рассмотрим этапы построение поискового тезауруса для пользователя:

1. Формирование тезауруса ПрО, интересующей пользователя. Для этого можно применить методологию разработки онтологических моделей – стандарт IDEF5 семейства IDEF (www.idef.com/IDEF5.html).

2. Формирование тезауруса ИР. По перечню слов, используемых в ИР, строится словарь терминов, из которого отбрасываются стоп-слова. Алгоритм применяется только для тех ИР, которые не сопровождаются метаописаниями. Иначе из метаописаний (в формате RDF или OWL) извлекаются термины тезауруса и связи между ними, которые дополняют построенный по контенту ИР словарь.

3. Фильтрация результатов запроса пользователя к внешней ИПС Интернет детально рассмотрена в [10].

Алгоритм фильтрации:

1. Пользователь вводит запрос, в котором идентифицируя свою информационную потребность с помощью множества ключевых слов.
2. Запрос передается внешней ИПС, от которой получают в соответствии с запросом множество I – набор ссылок на найденные ИР и их краткие описания .
3. Пользователь формирует тезаурус интересующей его ПрО (или указывает на ранее сформированный тезаурус) и соответствующий ему словарь терминов этой ПрО .
4. Если I содержит больше одной ссылки на ИР, то для каждого ИР из I формируются тезаурусы и соответствующие им словари терминов.
5. Для каждого ИР из I высчитывается его коэффициент близости к ПрО пользователя

$$K_j = \sum_{m,l} f\left(t_{j_l}, t_m\right), m = \overline{1,q}, l = \overline{1,l_j} \quad \text{где} \quad f(t_1, t_2) = \begin{cases} 0, если\ t_1 \neq t_2 \\ 1, если\ t_1 = t_2 \end{cases} \qquad (1)$$

,

где t_1 - термин из тезауруса ПрО, интересующей пользователя, t_2 - термин из тезауруса ИР.

Найденные ИР упорядочиваются в зависимости от коэффициентов (1), и пользователю предъявляются в первую очередь те ИР, которые имеют наиболее высокий коэффициент близости к ПрО пользователя, т.е. информация о семантике ИР используется для их упорядочения.

Программная реализация.

Интеллектуальная информационно-поисковая система МАИПС (http://progproblems.gradsoft.ua/maips-2006/) ориентирована на пользователей с постоянными информационными потребностями в областях, где они являются профессионалами (например, на научных работников). Такие пользователи достаточно четко представляют себе структуру и взаимосвязи своей ПрО, владеют соответствующей терминологией и хорошо представляют себе ИР, являющиеся объектом поиска. Для них важно отслеживать появление новых ИР в достаточно узкой сфере на протяжении длительного периода (например, статей по определенной области знаний, являющихся развитием идей какого-либо исследователя). Пользователь должен выбрать онтологию, характеризующую интересующую его ПрО, отметить в ней множество терминов, имеющих отношение к его запросу, и сформировать тезаурус запроса в виде перечня терминов с определенными весами или в виде облака тегов. После формирования тезауруса пользователь строит поисковый запрос, который переадресуется внешним ИПС. Результаты поиска возвращаются МАИПС и упорядочиваются в соответствии с тезаурусом.

Выводы

За последние несколько лет одним из ведущих направлений в развитии автоматического поиска, сбора и обработки информации стала технология Semantic Web, продвигаемая Консорциумом W3. Среди последних значительных достижений этого проекта - спецификация языка OWL, позволяющего создавать интероперабельные онтологии, и RDF, позволяющий создавать метаописания документов. Различные приложения Semantic Web (в частности, семантическая Википедия) позволяют структурировать и аккумулировать знания, представленные в Web, и стимулировать создание онтологий для различных ПрО. Эти онтологии могут впоследствии использоваться другими приложениями для обработки информации на семантическом уровне. В работе предложены методы применения таких онтологий для интеллектуализации поиска информации в Web.

Литература

1. Resource Description Framework (RDF) Model and Syntax Specification. W3C Proposed Recommendation. - January 1999. - http://www.w3.org/TR/PR-rdf-syntax.
2. Gladun A., Rogushina J. Knowledge Management in the Clinical Multiagent Information System // Труды Межд. научной конф. AITTH'2005 "Современные информационные и телемедицинские технологии для здравоохранения", Минск, Беларусь, 2005.- С.212-225.
3. Клещев А. С., Артемьева И.Л. Отношения между онтологиями предметных областей. Ч. 1. Онтологии, представляющие одну и ту же концептуализацию. Упрощение онтологии // Информационный анализ, В.1, С.2, 2002. – С.4-9.
4. Рогушина Ю.В., Гладун А.Я. Формирование и применение онтологий предметных областей для поиска Web-сервисов на семантическом уровне. // Труды Межд. конф. „Знания-Онтологии-Теория" ЗОНТ-2007, т.2, РАН И-тут математики им. С.П.Соболева, Новосибирск, Россия, 2007.-С.177-186.
5. Rogushina J., Gladun A. "Semantic Search of Internet Information Resources on Base of Ontologies and Multilinguistic Thesauruses" // International Journal «Information Theories and Applications», vol.14, 2006.-P.117-129.
6. Semantic Web Challenge. - http://challenge.semanticweb.org/.
7. Gladun A. , Rogushina J. Ontological Approach to Domain Knowledge Representation for Informational Retrieval in Multiagent Systems // International Jornal "Information Theories & Applications", V.13, N.4, 2006. – P.354-362.
8. Рогушина Ю.В., Гришанова І.Ю. Средства интеллектуализации поиска информационных ресурсов в сети Интернет // Тез. VI Междунар. конф. "Интеллектуальный анализ информации ИАИ-2007", 2007. – С.322-331.
9. Гладун А.Я., Рогушина Ю.В. Формирование тезауруса предметной области как средства моделирования информационных потребностей пользователя при поиске в Интернете // Вестник компьютерных и информационных технологий, Москва, № 1, 2007.-С.56-68.
10. Рогушина Ю.В., Гладун А.Я. Онтологии и мультилингвистические тезаурусы как основа семантического поиска информационных ресурсов Интернет// Proceedings of the XII-th International Conference "Knowledge-Dialogue-Solution", KDS'2006, FOI-Commerce, Sofia, Bulgaria.-P.115-121.

Информация об авторах

Гладун Анатолий Ясонович – *Международный научно-учебный центр информационных технологий и систем НАНУ, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40, email: glanat@yahoo.com*

Рогушина Юлия Витальевна – *Институт программных систем НАНУ, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40, Киевский Славистический Университет, email: _jjj_@ukr.net*

СИНТЕЗ НЕЙРОННЫХ СЕТЕЙ НА ОСНОВЕ ИНФОРМАЦИОННЫХ ГРАНУЛ

Лариса Катеринич, Александр Провотар

Аннотация: *Предлагается функциональная модель коммутационного элемента нейронной сети (информационной гранулы).*

Ключевые слова: *искусственный интеллект, нейронные сети, обучение нейронных сетей, информационная гранула.*

ACM Classification Keywords: *F.1.1 Models of Computation – neural networks, C.1.3 Other Architecture Styles – Neural Nets.*

Conference*: The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

При реализации системы Гомеопат [Катеринич, 2007] возникло ряд проблем связанных с тем, что выбранная архитектура сети в некоторых случаях не может полностью удовлетворять решению поставленной перед ней задачи. А именно, возникает такая ситуация, когда решение задачи диагностики является ошибочным или не полным. Для устранения таких ошибок предлагается, как правило, решать задачи в сети с другой архитектурой и другим алгоритмом обучения соответственно.

Как известно [Хайкин, 2006]:

- структура нейронных сетей тесно связана с используемым алгоритмом обучения, причем разные алгоритмы обучения эффективны для решения определенных классов задач и проблем

- важным свойством нейронных сетей является их способность обучаться на основе данных окружающей среды и в результате обучения повышать свою производительность. Повышение производительности происходит со временем в соответствии с определенными правилами.

- не существует универсального алгоритма обучения, подходящего для всех архитектур нейронных сетей [Барский, 2004]. Существует лишь набор средств, представленный множеством алгоритмов обучения (алгоритмы обучения отличаются друг от друга способом настройки синоптических весов нейронов), каждый из которых имеет свои преимущества и недостатки.

- отличительной характеристикой обучаемой нейросети является способ ее связи с внешним миром.

Учитывая то, что существующие основные модели обучения: на основе коррекции ошибок (реализует метод оптимальной фильтрации), с использованием памяти (предполагает явное использование обучающих данных), Хеббовское и конкурентное обучение (основаны на нейробиологических принципах) и метод Больцмана (на основе идей статической механики) имеют ряд ограничений, при создании универсальных систем с использованием нейронных сетей обойтись одной из моделей обучения очень сложно. А если учесть еще ряд других вопросов и задач, возникающих в каждой конкретной предметной области, то построение универсальной системы на основе существующих моделей и соответствующего математического аппарата представляется достаточно сложной, если не разрешимой задачей.

Информационная гранула

Исходя из этого возникла идея создания так называемого коммутационного элемента, обладающего набором необходимых значений-параметров для взаимодействия сетей, оптимально решающих разные задачи из представленной предметной области (ПО), основываясь на разных методах обучения, архитектуре сетей и т.д. с возможностью построения универсальной системы для этой ПО (возможно и для смежных ПО). Важным и очень ценным свойством таких систем была бы возможность идентификации узкой проблемы или задачи, что позволяло бы системе оптимально подобрать метод обучения всей системы новыми данными с учетом ранее приобретенных знаний.

Коммутационным элементом может выступать так называемая информационная гранула. Информационная гранула в общем случае может быть интерпретирована как одна из многочисленных маленьких частиц, формирующих большую единицу. Они имеют, по крайней мере, три основных свойства: внутренние свойства, внешние свойства, контекстные свойства. Также гранулу рассматривают как объединение единичных элементов, охарактеризованные внутренними свойствами и как неразделимое целое, охарактеризованное внешними свойствами гранулы. Формально информационная гранула А может быть представлена в некотором пространстве X как отображение $A : X \rightarrow \delta(X)$. δ обозначает формальную структуру информационной гранулы того пространства в котором она рассматривается. [Bargiela, 2003]

Информационная гранула имеет свою семантику и синтаксис. Семантический аппарат обращается непосредственно к грануле, а именно к тому составляющему компоненту, который отвечает за накопление и отображение соответствующей информации. Синтаксический аппарат информационной гранулы четко отделим от семантического аппарата. Он имеет свою архитектуру и формальные методы обработки представленной информации. При четком разделении между компонентами гранулы существует постоянная связь. Она обеспечивает возможность контролирования потоков информации.

В нашем случае информационная гранула обладает следующими базовыми свойствами: возможностью классификации входной и выходной информации, подготовкой выходной информации с оптимальной подготовленной структурой данных для дальнейшей обработки.

Рассмотрим реализацию этого механизма на таком примере.

Алгоритм синтеза нейросетей.

1. *Первое, что необходимо сделать, это выделить те характеристики, по которым система будет классифицировать входные данные и в дальнейшем выбирать ту модель обучения, которая является оптимальной в этой ситуации.* Для решения такой задачи можно отталкиваться от следующего, то есть выполнять ряд проверок.

Например, если входные данные (представлены вектором) выбраны из двух линейно-разделимых классов, алгоритм персептрона разработанный Розенблатом [Хайкин, 2006] сходится и формирует поверхность решений в форме гиперплоскости, разделяющих эти два класса. Для большего количества классов требуется гораздо больше времени для обучения сети. Таким образом, однослойный персептрон не дает большой скорости обучения и поэтому для больших наборов данных является не оптимальным в использовании. В тоже время на меньших объемах входных данных он будет работать, и давать результаты гораздо быстрее и эффективнее чем более поздние его собратья. Данный подход позволяет использовать его для решения очень узкой задачи из представленной ПО.

Способность к обучению на собственном опыте обеспечивает вычислительную мощность многослойного персептрона. Однако же эти же качества являются причиной неполноты современных знаний о поведении такой среды [Хайкин, 2006].

2. *Следующим шагом есть обучения подсистемы на основе выбранной модели и передача полученных знаний в реестр коммутационного элемента.* Основной задачей последнего является определения важности полученных знаний с более узкой предметной области и передача их другой подсистеме для получения необходимых или недостающих знаний в другой подсистеме.

Подсистема – это функциональный элемент основной системы, реализованный с помощью нейронной сети с использованием метода обучения оптимальной для решения задачи в более узкой предметной области. Таких коммутационных элементов может быть n-1, где n количество подсистем.

3. *И последним этапом является выдача нужной информации пользователю системы.*

Такой подход позволяет решить больший круг задач, а именно создать более универсальную систему, функционирующую на основе объединения нейронных сетей умеющих оптимально решать узкий класс задач, используя оптимальный для решения существующий на данный момент алгоритм. Другими словами была предложена концепция построения так называемых объединенных нейронных сетей с использованием коммутационных элементов (информационных гранул).

Классификация данных в системе Гомеопат

В системе «Гомеопат» основными входными данными есть симптомы пациентов, а также вспомогательные (дополнительные) данные такие как фамилия, имя, возраст, и т.д. Набор вспомогательных данных фиксирован, набор симптомов возможных для ввода пациентом неограниченна. Таким образом мы имеем множество G_{inp}={S, P}, где S={s1,s2,…sn} – множество симптомов пациента, P={p1,p2,…pn} – множество вспомогательной информации, G_{imp}- множество всей допустимых входных данных.

Рассмотрим пример задачи классификации. В упрощенной модели представления данных задача состоит в выработке правил классификации симптомов для передачи корректных данных на вход соответствующей сети в зависимости, например, от их корректности.

Для удобства изложения ограничимся двумя НС сетями решающие разные классы задач. В контексте системы ГОМЕОПАТ это может быть, например, классификация корректности ввода симптомов. Например, классификация симптомов заболевания сердца и дыхательных путей. Для решения проблемы разделения образцов на классы потребуется несколько гиперплоскостей (проблема называется нелинейной.). Функция выбора решения смоделирована с помощью нейронной сети.

Обучение и наполнение информационных гранул в системе Гомеопат

Обучение системы происходит по одной из моделей обучения. Передача данных происходит напрямую. Все действия по обработке данных выполняет коммутационный элемент. Коммутационный элемент реализован с помощью НС. Задача такого элемента есть дальнейшая классификация входной в него информации с дальнейшим распределением по сетям оптимально решающие задачи конкретной предметной области. Входной информацией на первом данном этапе будет корректные симптомы заболеваний.

Как известно проектирование, обучение и работа в рабочем режиме сети накладывает ряд ограничений. Сеть должна обучиться (обучающая выборка), а уже в дальнейшем, в рабочем режиме при подаче на вход НС сходной информации правильно реагировать. Также НС направляет представленную ей входную информацию на входы НС оптимально решающую этот класс задач. Обучающей выборкой в системе ГОМЕОПАТ, на первом этапе, служит правильно классифицированная информация, а на дальнейших этапах результаты работы НС решающие оптимально задачи ПО с разной архитектурой и методами обучения. Входная выборка BX1…BXN – входная информация. Пример функциональной структуры информационной гранулы представлен в таблице 1.

Таблица 1. Функциональная структура информационной гранулы

BX1	BX2	BX3	…	BXN	Класс задач, алгоритм обучения
Управляемое обучение. НС обучается классифицировать образцы в соответствии инструкциями. Целевой входной образец дает информацию сети, о том, к какому классу следует научиться относить входной образец.					Классификация образов

Обучение без управления. Сеть выполняет разделение на группы (кластеризацию) самостоятельно. Все образцы одного кластера должны иметь общую суть – тогда они буду оцениваться как подобные.	Кластеризация образцов
Суть заключается в выборе нужного образца из памяти, даже при отсутствии всей необходимой информации для начала поиска сохраненного образца.	Ассоциация образов
Обработка структуры входных данных как последовательности. Последовательность(П) – это цепочка образцов, имеющих отношение к одному и тому же объекту. П имеют разную длину.	Рекуррентные сети

Заключение

Представленный подход позволяет, используя как коммутационный элемент - информационную гранулу, построить систему, состоящую из разных сетей с разной архитектурой и методом обучения, что в свою очередь позволяет ускорить в целом реакцию всей системы на представленную информацию. Коммутационный элемент в силу своих функциональных свойств позволяет классифицировать представленную на вход информацию и на выходе предоставлять информацию с той структурой данных, которая наиболее оптимально подходит для нейронной сети решающую этот класс задач. Также следует отметить, что одним из ключевых функций является возможность обработки входной информации с дальнейшем выбором алгоритма и архитектуры сети.

Библиография

[Катеринич, 2007] Л. Катеринич, А. Провотар. Диагностирование на нейронных сетях в системе Гомеопат // XIII-th International Conference: Knowledge Dialogue Solution. - Sofia, 2007. - V1. – P.64-68.

[Хайкин, 2006] С. Хайкин. Нейронные сети: полный курс, 2-е изд. – М.: ООО «И.Д. Вильямс», 2006. – 220с.

[Bargiela, 2003] Bargiela, Andrzeyj and Pedrycz, Witold. Granular Computing: An introduction. – Kluwer Academic Publishers, 2003. – 5p.

[Барский, 2004] А.Б. Барский. Нейронные сети: распознавание, управление, принятие решений.- М.: «Финансы и статистика», 2004. - 398с.

[Оссовский, 2002] С.Оссовский. Нейронные сети для обработки информации. – М: «Финансы и статистика», 2002. – 365с.

[Терехов, 2002] В.А. Терехов, Д.В.Уфимов, И.Ю. Тюкин. Нейросетевые системы управления.- М.: «Радиотехника», 2002. – 467с.

Информация об авторах

***Катеринич Лариса Александровна** – ассистент факультета кибернетики Киевского национального университета имени Тараса Шевченка, katerinich@rambler.ru*

***Провотар Александр Иванович** – доктор физико-математических наук, профессор, заведующий кафедрою информационных систем факультета кибернетики Киевского национального университета имени Тараса Шевченка, aprowata@unicyb.kiev.ua*

КОМБИНИРОВАННЫЙ ПОДХОД К ПРЕДСТАВЛЕНИЮ СОДЕРЖАНИЯ И ТЕКСТОВОГО ОПИСАНИЯ МУЛЬТИМЕДИА

Дмитрий Ночевнов

Аннотация: *В статье рассмотрена проблема семантической разницы между содержимым мультимедиа и его текстовым описанием, определяемым вручную. Предложен комбинированный подход к представлению семантики мультимедиа, основанный на объединении близких по содержанию и текстовому описанию мультимедиа в классы, содержащие обобщённые описания объектов, связей между ними и ключевых слов текстовых метаданных из некоторого тезауруса. Для формирования этих классов используются операции иерархической кластеризации и машинного обучения. Данный подход позволяет расширить область поиска и навигации мультимедиа благодаря привлечению медиа-данных, имеющих схожее содержание и текстовое описание.*

Ключевые слова: *multimedia data mining, content based image retrieval, text-based image retrieval.*

ACM Classification Keywords: *H.3.1 Content Analysis and Indexing - Abstracting methods. H.5.1 Multimedia Information Systems - Evaluation/methodology.*

Conference: *The paper is selected from XIVth International Conference "Knowledge-Dialogue-Solution" KDS 2008, Varna, Bulgaria, June-July 2008*

Введение

Интерес к области обработки мультимедиа обусловлен быстром ростом сети WWW и возникающей необходимости быстрой индексации и поиска медиа данных среди большого и разнородного массива информации. Производство и потребление мультимедийного содержания и документов стали обычной практикой благодаря существованию эффективных инструментов создания метаданных в машиночитаемом текстовом формате или же в форме визуальных и аудио дескрипторов (например, в формате MPEG-7), и индексации мультимедиа. Это облегчает их обработку поисковыми машинами и интеллектуальными агентами [Stamou, 2005]. Однако остаётся нерешённой проблема семантического барьера между такими низкоуровневыми характеристиками мультимедиа, такими как цвет, текстура, сцена и т.п., и высокоуровневыми концепциями типа «горный ландшафт», «аномальное поведение человека», используемыми человеком для описания мультимедиа. Преодоление этого барьера является одной из задач Multimedia Data Mining [Stamou, 2005], [Petrushin, 2007].

Традиционный поиск мультимедиа принято разделять на два основных вида [Вихровский, 2006], [Goodrum, 2000] :

1. Поиск на основе текстового описания (Text-Based Retrieval). Данный поиск использует высокоуровневую информацию, опираясь на ключевые слова некоторого тезауруса или свободную текстовую аннотацию.

2. Контентно-зависимый поиск (Content-Based Retrieval), который опирается на использование низкоуровневой информации и визуальных данных в качестве запроса.

Основной проблемой Text-Based Retrieval является сложность точного и полного текстового описания мультимедиа, которое разные пользователи могут идентифицировать по разному исходя из собственного опыта и знаний [Goodrum, 2000]. Одним из решений этой проблемы является Content-Based Retrieval на основе шаблона искомого мультимедиа. Однако в некоторых случаях только на основе низкоуровневого описания сложно автоматически определить интересующую искателя информацию, например создателя мультимедиа, точное время суток, объекты «за кадром», трёхмерные отношения между изображёнными объектами и т.д. Кроме этого сложно обеспечить точность и полноту такого поиска из-за разнообразия отображения схожих по текстовому описанию мультимедиа в разных предметных областях.

Исходя из этого целесообразно использование комбинированного подхода к индексации и поиску мультимедиа, объединяющего низкоуровневое описание мультимедиа и более высокоуровневое его текстовое описание [Goodrum, 2000]. Это должно способствовать повышению эффективности обработки и поиска мультимедиа благодаря привлечению в процесс поиска сведений о мультимедиа, имеющих схожее содержание и текстовое описание.

1. Комбинированная модель представления знаний мультимедиа

В большинстве своём медиа-данные фактически содержат визуальные или звуковые следы физических объектов, записанные с помощью сенсоров [Stamou, 2005]. Если таких объектов несколько, то сохраняется также информация об их пространственных (в случае изображений и видео данных) или временных (в случае аудио и видео данных) взаимосвязях. Для обозначения этих объектов можно использовать существующие онтологии и тезаурусы, содержащие названия объектов мультимедиа, например Getty's Art and Architecture Thesaurus [AAT], состоящий из более чем 120000 терминов для описания искусства, архитектуры и других культурных объектов, или же Library of Congress Thesaurus of Graphic Materials [LCTGM], и хранить вместе с названием соответствующие метаданные в формате MPEG-7, характеризующие данный объект. Подобный подход к представлению знаний мультимедиа можно найти в работе [Petridis, 2004], в которой для хранения описания MPEG-7 мультимедиа используется онтология в формате Semantic Web. Для формализации связей между объектами можно использовать язык предикатов первого уровня.

Всё это позволит автоматизировать выделение объектов и связей между ними во время индексации мультимедиа, а также расширить область поиска путём привлечения близких по смыслу объектов.

1.1. Модель содержания мультимедиа

Одним из подходов к представлению мультимедиа является формализм семантических сетей [Petridis, 2004], [Dance, 1996]. Следуя ему, для формализации объектов и связей, отображаемых в мультимедиа **m**, и его текстового описания, будем использовать графа, содержащий:

1) множество вершин $v_i \in V_m$, представляющих объекты, отображённые в мультимедиа **m**; каждой из вершин ставится в соответствие некоторое ключевое слово $d(v_i)$ из тезауруса наименований объектов, такое что значения MPEG-7 дескрипторов этих объектов близки к значениям дескрипторов объектов тезауруса;

2) множество связей между двумя вершинами $a_{jk} \in A_m$, соответствующие реальным связям между объектами, отображёнными в мультимедиа; каждой связи ставится в соответствие предикат $p(a_{jk})$;

3) текстовое описание мультимедиа T_m, представляемое в виде множества ключевых слов, и определяемое вручную, или автоматически путём индексации текстового описания мультимедиа.

Одному ключевому слову тезауруса *d* может соответствовать несколько вершин моделей одной или нескольких мультимедиа в случае повторения одного и того же типа объекта. Один и тот же тип связи, обозначенный предикатом *p*, также может быть представлен в моделях одной или нескольких мультимедиа.

1.2. Модель класса мультимедиа

Семантически близкие мультимедиа $m_n \in M_c$ предлагаем объединять в классы ***c***, содержащие обобщённую информацию о типичных объектах и связях, отображаемых в мультимедиа, а также ключевые слова их текстовых описаний.

Для его формализации также используем граф, содержащий:

1) множество взвешенных вершин $v_i \in V_c$, описывающих представленные во множестве объединяемых мультимедиа $m_n \in M_c$ объекты; каждой из вершин ставится в соответствие некоторое ключевое слово $d(v_i)$ из тезауруса наименований объектов, а также вес $w(v_i)$, обозначающий количество повторений данного объекта во множестве M_c;

2) множество взвешенных связей между двумя вершинами $a_{jk} \in A_c$, описывающих представленные во множестве объединяемых мультимедиа $m_n \in M_c$ связи между объектами; каждой из связей ставится в соответствие некоторый предикат $p(a_{jk})$, а также вес $w(a_{jk})$, обозначающий количество повторений данной связи между объектами во множестве M_c;

3) текстовое описание класса мультимедиа T_c, определяемое индексацией текстовых описаний t_m множества мультимедиа $m_n \in M_c$; представляет собой множество взвешенных ключевых слов из текстовых описаний t_m с весами $w(t_i)$, обозначающими количество повторений данного ключевого слова среди t_m.

2. Последовательность обработки мультимедиа

Можно предложить следующий алгоритм обработки мультимедиа:

1. Индексация нового мультимедиа и составление модели содержания мультимедиа m:

а) автоматическое выделение объектов, отражённых в мультимедиа и поиск наиболее близких по значениям MPEG-7 дескрипторов ключевых слов d тезауруса объектов;

б) автоматическое выделение связей между объектами и поиск соответствующих этим связям предикатов p, и составление модели содержания мультимедиа;

в) определение текстового описания t_m.

2. Определение семантически наиболее близкого к мультимедиа класса c_i и обновление его модели данными о новом мультимедиа, или же добавление нового класса мультимедиа.

3. Проверка качества разбиения мультимедиа на классы и при необходимости кластеризация семантически однородных классов.

4. Использование во время поиска сведений о близких по содержанию и описанию мультимедиа из модели класса мультимедиа.

Семантически наиболее близкий класс c_i или множество классов С могут быть определены пользователем при навигации в базе знаний мультимедиа, или же вычислены автоматически путём анализа расстояния между мультимедиа m и существующими классами по формуле:

$$c = c_k \text{, при } h(m, c_k) = \min(h(m, c_k)) \le h_{\max}, i = \overline{1, N_c}\ , \qquad (1)$$

где N_c – общее кол-во классов мультимедиа.

3. Формализация расстояния между классами и экземплярами мультимедиа

Определим метод вычисления расстояния между классами и экземплярами мультимедиа. Для предложенных в предыдущем разделе моделей наиболее подходят модифицированные способы вычисления сходства, основанные на описании характеристик объекта в виде вектора ключевых слов и вычислении совпадения элементов векторов и их весов [Озкархан, 1989].

Согласно определению из [Дюран, 1977]:

Определение 1. Неотрицательная вещественная функция $z(x, y)$ называется *функцией близости*, если:

1) $0 \le z(x, y) < 1$ для $x \neq y$,

2) $z(x, x) = 1$,

3) $z(x, y) = z(y, x)$.

Определение 2. Неотрицательная вещественная функция $h(y, x) = 1 - z(y, x)$ называется *функцией расстояния*.

При вычислении семантически близкого класса мультимедиа необходимо определять расстояние между классом и экземпляром мультимедиа $h(m,c)$. Для его расчёта будем учитывать совпадение объектов моделей, связей между ними, ключевых слов в текстовых описаниях и их весов:

$$h(m,c) = 1 - z(m,c) = 1 - \frac{\lambda_v \cdot z_v(m,c) + \lambda_a \cdot z_a(m,c) + \lambda_t \cdot z_t(m,c)}{\lambda_v + \lambda_a + \lambda_t}, \qquad (2)$$

где $\lambda_v, \lambda_a, \lambda_t \in [0,1]$- коэффициенты влияния, определяемые опытным путём,

$$z_v(m,c) = \left(1 - \frac{\sum\limits_{\forall v \in (V_m \cap V_c)} n[w_c(v)]}{card(V_m \cap V_c)}\right) \cdot \frac{card(V_m \cap V_c)}{card(V_m \cup V_c)}$$ – степень близости по составу объектов,

$$z_a(m,c) = \left(1 - \frac{\sum\limits_{\forall a \in (A_m \cap A_c)} n[w_c(a)]}{card(A_m \cap A_c)}\right) \cdot \frac{card(A_m \cap A_c)}{card(A_m \cup A_c)}$$ – степень близости по составу связей между объектами,

$$z_t(m,c) = \left(1 - \frac{\sum\limits_{\forall t \in (T_m \cap T_c)} n[w_c(t)]}{card(T_m \cap T_c)}\right) \cdot \frac{card(T_m \cap T_c)}{card(T_m \cup T_c)}$$ – степень близости текстовых описаний мультимедиа,

$n(x) = \left|0.1 \cdot \lg(1+x)\right|_{\mathrm{mod}\,1}$ - функция нормализации весов модели класса мультимедиа,

$\lambda_v = 0$, если $card(V_m \cup V_c) = 0$; $\lambda_a = 0$, если $card(A_m \cup A_c) = 0$; $\lambda_t = 0$, если $card(T_m \cup T_c) = 0$.

При объединении множества близких по содержанию мультимедиа в класс возникает потребность в определении расстояния между отдельными экземплярами мультимедиа $h(m_k,m_l)$. Для его расчёта будем учитывать совпадение объектов и связей, отображаемых в мультимедиа m_k и m_l, а также совпадение ключевых слов их текстовых описаний:

$$h(m_k,m_l) = 1 - z(m_k,m_l) = 1 - \frac{\lambda_v \cdot z_v(m_k,m_l) + \lambda_a \cdot z_a(m_k,m_l) + \lambda_t \cdot z_t(m_k,m_l)}{\lambda_v + \lambda_a + \lambda_t}, \qquad (3)$$

где $z_v(m_k,m_l) = \frac{card(V_{m_k} \cap V_{m_l})}{card(V_{m_k} \cap V_{m_l})}$ – степень близости по составу объектов,

$z_a(m_k,m_l)=\dfrac{card(A_{m_k}\cap A_{m_l})}{card(A_{m_k}\cap A_{m_l})}$ – степень близости по составу связей между объектами,

$z_t(m_k,m_l)=\dfrac{card(T_{m_k}\cap T_{m_l})}{card(T_{m_k}\cap T_{m_l})}$ – степень близости текстовых описаний мультимедиа,

$\lambda_v=0,\text{если}\, card(V_{m_k}\cup V_{m_l})=0$; $\lambda_a=0,\text{если}\, card(A_{m_k}\cup A_{m_l})=0$; $\lambda_t=0,\text{если}\, card(T_{m_k}\cup T_{m_l})=0$.

Для определения качества разбиения множества мультимедиа на классы следует вычислять расстояние между классами мультимедиа $h(c_i,c_j)$. Будем определять его с учётом совпадение объектов классов, связей между ними, ключевых слов в текстовых описаниях и их весов:

$$h(c_i,c_j)=1-z(c_i,c_j)=1-\frac{\lambda_v\cdot z_v(c_i,c_j)+\lambda_a\cdot z_a(c_i,c_j)+\lambda_t\cdot z_t(c_i,c_j)}{\lambda_v+\lambda_a+\lambda_t}, \qquad (4)$$

где $z_v(c_i,c_j)=\left(1-\dfrac{\sum\limits_{\forall v\in(V_{c_i}\cap V_{c_j})} n(\left|w_{c_i}(v)-w_{c_j}(v)\right|)}{card(V_{c_i}\cap V_{c_j})}\right)\cdot\dfrac{card(V_{c_i}\cap V_{c_j})}{card(V_{c_i}\cup V_{c_j})}$ – степень близости по составу объектов,

$z_a(c_i,c_j)=\left(1-\dfrac{\sum\limits_{\forall a\in(A_{c_i}\cap A_{c_j})} n(\left|w_{c_i}(a)-w_{c_j}(a)\right|)}{card(A_{c_i}\cap A_{c_j})}\right)\cdot\dfrac{card(A_{c_i}\cap A_{c_j})}{card(A_{c_i}\cup A_{c_j})}$ – степень близости по составу связей,

$z_t(c_i,c_j)=\left(1-\dfrac{\sum\limits_{\forall t\in(T_{c_i}\cap T_{c_j})} n(\left|w_{c_i}(t)-w_{c_j}(t)\right|)}{card(T_{c_i}\cap T_{c_j})}\right)\cdot\dfrac{card(T_{c_i}\cap T_{c_j})}{card(T_{c_i}\cup T_{c_j})}$ – степень близости текстовых описаний,

$\lambda_v=0,\text{если}\, card(V_{c_i}\cup V_{c_j})=0$; $\lambda_a=0,\text{если}\, card(A_{c_i}\cup A_{c_j})=0$; $\lambda_t=0,\text{если}\, card(T_{c_i}\cup T_{c_j})=0$.

4. Обновление класса мультимедиа

Обновление класса мультимедиа сведениями о новом мультимедиа предлагаем делать в режиме обучения, увеличивая веса только тех объектов, связей между ними и ключевых слов текстового описания, которые повторяются в новом мультимедиа. В соответствии с этим правилом во время обучения класса ***c*** информацией о мультимедиа ***m*** последовательно выполняется:

1) модификация значений весов вершин модели класса $w_c(v_i)$ по формуле:

$$\forall v_i\in V'=V_m\cap V_c\rightarrow w_c(v_i)=w_c(v_i)+1;$$

2) модификация значений весов связей объектов $w_c(a_i)$ по формуле:

$$\forall a_i\in A'=A_m\cap A_c\rightarrow w_c(a_i)=w_c(a_i)+1;$$

3) модификация значений весов ключевых слов текстового описания класса $w_c(t_i)$ по формуле:

$$\forall t_i\in T'=T_m\cap T_c\rightarrow w_c(t_i)=w_c(t_i)+1;$$

4) дополнение класса ***c*** недостающими объектами, связями и ключевыми словами текстового описания с весами, равными 1:

$$c = c + (m \setminus c) = \{V_c = V_c + V_m \setminus V_c, A_c = A_c + A_m \setminus A_c, T_c = T_c + T_m \setminus T_c\}\ .$$

В случае, если не будет найден семантически близкий класс, информация о новом мультимедиа может быть добавлена в базу знаний в виде нового класса с единичными значениями весов объектов, связей и ключевых слов *t*.

5. Автоматическая классификация мультимедиа

Для автоматической классификации мультимедиа и разбиения их на классы предлагаем использовать *собирающий метод иерархической кластеризации* [Pedrycz, 2005]. В соответствии с ним кластеризация будет начинаться с единственных кластеров для каждого мультимедиа, которые затем объединяются в кластеры, формируя двухуровневую иерархическую структура, в которой:

1-ый уровень– кластеры C_i с минимальным расстоянием между кластерами $h_{\min}$=1;

2-ой уровень – подкластеры C_{ij} с минимальным расстоянием между кластерами $h_{\min}$<1.

Разбиение мультимедиа на подкластеры второго уровня предлагаем выполнять пошагово с изменением значения h_{min} от 0 до 1, пока не будет достигнуто приемлемое качество разбиения. Для его оценки предлагаем использовать меру внутренней однородности кластера η_o и меру разнородности кластеров η_m, вычисляемые следующим образом:

$$\eta_{o_i} = \begin{cases} \dfrac{2}{N_i(N_i - 1)} \sum\limits_{j=1}^{N_i} \sum\limits_{k=j+1}^{N_i} h(m_j, m_k), \text{если}\, N_i > 1 \\ 1, \text{если}\, N_i = 1 \end{cases}$$, где N_i – кол-во мультимедиа, объединённых в кластер C_i,

$$\eta_m = \begin{cases} \dfrac{2}{N_C(N_C - 1)} \sum\limits_{j=1}^{N_C} \sum\limits_{k=j+1}^{N_C} h(m_ц^{\langle C_j \rangle}, m_ц^{\langle C_k \rangle}), \text{если}\, N_C > 1 \\ 0\, , \text{если}\, N_C = 1 \end{cases},$$

где $m_ц^{\langle C_i \rangle}$ – центральное мультимедиа кластера C_i, для которого выполняется условие $\sum\limits_j h(m_j, m_ц) = \min, \forall m_j \in C_i$,

N_C – кол-во анализируемых кластеров.

По результатам кластеризации выполняется группировка мультимедиа $m_i \in C_j$ в класс c_i. При группировке можно использовать операцию обновления класса сведениями о мультимедиа, описанную в пункте 4 данной статьи.

Выводы

Хотя на сегодняшний день производство и потребление мультимедийного содержания и документов стали обычной практикой, остаётся нерешённой проблема семантической разницы между содержимым мультимедиа и его текстовым описанием, определяемым пользователями. В статье обоснована целесообразность использования комбинированного подхода к индексации и поиску мультимедиа, объединяющего низкоуровневое описание мультимедиа и более высокоуровневое его текстовое описание. Предложен комбинированный подход к представлению семантики мультимедиа, основанный на объединении близких по содержанию и текстовому описанию мультимедиа в классы, содержащие обобщённые описания объектов, связей между ними и ключевых слов текстовых метаданных из некоторого тезауруса. Для формирования этих классов используются операции иерархической кластеризации и машинного обучения.

Описанный подход даёт возможность расширить область поиска и навигации мультимедиа благодаря привлечению медиа-данных, имеющих схожее содержание и текстовое описание.

Литература

[AAT] Getty's Art and Architecture Thesaurus. - http://www.getty.edu/research/conducting_research/vocabularies/aat/

[Dance, 1996] Dance Sandy, Caelli Terry, Liu Zhi-Qiang. Picture Interpretation: A Symbolic Approach//World Scientific Series In Machine Perception And Artificial Intelligence; Vol. 20 – 1996.

[Goodrum, 2000] Goodrum A. A. Image information retrieval: An overview of current research // Informing Science, 3(2):P.63-66, February 2000.

[LCTGM] Thesaurus for Graphic Materials // Library of Congress. - http://www.loc.gov/rr/print/tgm1/

[MPEG-7] MPEG-7 Overview. http://www.mpeg-7.com (Industry Focus Group)

[Pedrycz, 2005] Pedrycz Witold. Knowledge-based clustering: From Data to Information Granules. - John Wiley & Sons, 2005. - 316p.

[Petridis, 2004] Petridis K., Kompatsiaris I., Strintzis M.G., Bloehdorn S., Handschuh S., Staab S., Simou N., Tzouvaras V., Avrithis Y. Knowledge representation for semantic multimedia. Content analysis and reasoning. – EWIMT, 2004.

[Petrushin, 2007] Petrushin Valery A. and Khan Latifur (Eds). Multimedia Data Mining and Knowledge Discovery. - Springer-Verlag London Limited, 2007. – 521p.

[Stamou, 2005] Stamou Giorgos and Kollias Stefanos (Eds). Multimedia Content and the Semantic Web. - John Wiley & Sons Ltd, 2005 - 392 p..

[Вихровский, 2006] Вихровский Кирилл, Игнатенко Алексей. Применение MPEG-7 для классификации и поиска визуальных данных // Сетевой журнал "Графика и Мультимедиа". - 23.12.2006 http://cgm.graphicon.ru/content/view/161/61/

[Дюран, 1977] Дюран Б., Оделл П. Кластерный анализ. – М.: Статистика, 1977. – 128 с.

[Озкархан, 1989] Озкархан Э. Машины баз данных и управление базами данных: Пер. с англ. – М.: Мир, 1989. – 696с.

Authors' Information

***Дмитрий Ночевнов** – доцент кафедры информационных технологий проектирования Черкасского государственного технологического университета.*

Адрес: Кафедра информационных технологий проектирования, Черкасский государственный технологический университет, 18006, г.Черкассы, бул.Шевченка, 460; e-mail: dmitry.ndp@gmail.com