# Fifth International Conference

# INFORMATION RESEARCH AND APPLICATIONS

**26-30 June 2007, Varna**

i.tech

# PROCEEDINGS

## Volume 1

**ITHEA**

**SOFIA, 2007**

Kr. Markov, Kr. Ivanova (Ed.)

Proceedings of the Fifth International Conference “Information Research and Applications” i.TECH 2007, Varna, Bulgaria

**Volume 1**

Sofia, Institute of Information Theories and Applications FOI ITHEA – 2007

First edition

Printed in Bulgaria by Institute of Information Theories and Applications FOI ITHEA
Sofia -1090, P.O. Box 775
e-mail: info@foibg.com
www.foibg.com

**ISSN 1313-1109 (paperback) ISSN 1313-115X (CD) ISSN 1313-1222 (online)**

FSBN Volume 1: 978-954-16-2001-4 (paperback) 978-954-16-2002-1 (CD) 978-954-16-2003-8 (Online)
FSBN Volume 2: 978-954-16-2004-5 (paperback) 978-954-16-2005-2 (CD) 978-954-16-2006-9 (Online)

# PREFACE

The Fifth International Conference "***Information Research and Applications***" (**i.TECH 2007**) is organized as a part of "ITA 2007 - Joint International Scientific Events on Informatics".

ITA 2007 as well as the i.TECH 2007 is supported by

International Journal on Information Theories and Applications (IJ ITA)
and
International Journal on Information Technologies and Knowledge (IJ ITK)

i.TECH 2007 is dedicated to:

- 60th Anniversary of the Institute of Mathematics and Informatics of Bulgarian Academy of Sciences;
- 15th Anniversary of the Association of Developers and Users of Intelligent Systems (Ukraine);
- 10th Anniversary of the Association for Development of the Information Society (Bulgaria).

The aim of the conference is to be one more possibility for contacts for scientists. The usual practice of IJ ITA and IJ ITK are to support several conferences at which the papers may be discussed before submitting them for referring and publishing in the journals. Because of this, such conferences usually are multilingual and bring together both papers of high quality and papers of young scientists, which need further processing and scientific support from senior researchers.

We would like to express our thanks to all who support the i.TECH 2007 and especially to the *Natural Computing Group* (NCG) (http://www.lpsi.eui.upm.es/nncg/) of the Technical University of Madrid, which is leaded by Prof. Juan Castellanos.

Let us thank the Program Committee of the conference for referring the submitted papers. Special thanks to prof. Viktor Gladun, prof. Alexey Voloshin, prof. Avram Eskenazi and prof. Luis Fernando de Mingo.

i.TECH 2007 Proceedings has been edited in the *Institute of Information Theories and Applications FOI ITHEA* in collaboration with *Institute of Cybernetics "V.M.Glushkov", NASU (Ukraine)*, *Kiev University "T.Shevchenko" (Ukraine)*, *Institute of Mathematics and Informatics, BAS (Bulgaria)*, *Institute of Information Technologies, BAS (Bulgaria)*, *University of Calgary (Canada)*, *VLSI Systems Centre, Ben-Gurion University (Israel)*.

The i.TECH 2007 Conference found the best support in the work of Organizing Committee Chairman Ilia Mitov.

To all participants of i.TECH 2007 we wish fruitful contacts during the conference days and efficient work for preparing the high quality papers to be published in the International Journal "Information Theories and Applications" or in the International Journal "Information Technologies and Knowledge".

Varna, June 2007 Kr Markov, Kr. Ivanova

**i.TECH 2007 has been organized by:**

- Institute of Information Theories and Applications FOI ITHEA (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)
- Technical University of Madrid (Spain)
- Taras Shevchenko National University of Kiev (Ukraine)
- Kharkiv National University of Radio Electronics (Ukraine)
- Association of Developers and Users of Intelligent Systems (Ukraine)
- Association for Development of the Information Society (Bulgaria)
- International Journal “Information Theories and Applications”
- International Journal "Information Technologies and Knowledge"

**Program Committee:**

Victor Gladun (Ukraine)

Alexey Voloshin (Ukraine)
Avram Eskenazi (Bulgaria)
Krassimir Markov (Bulgaria)
Luis Fernando de Mingo (Spain)

Adil Timofeev (Russia)
Alexander Gerov (Bulgaria)
Alexander Kuzemin (Ukraine)
Alfredo Milani (Italy)
Anna Kantcheva (Bulgaria)
Arkady Zakrevskij (Belarus)
Ilia Mitov (Bulgaria)
Ivan Popchev (Bulgaria)
Juan Castellanos (Spain)
Koen Vanhoof (Belgium)
Krassimir Manev (Bulgaria)
Krassimira Ivanova (Bulgaria)
Laura Ciocoiu (Romania)
Levon Aslanyan (Armenia)
Martin Mintchev (Canada)
Nelly Maneva (Bulgaria)
Orly Yadid-Pecht (Israel)
Peter Stanchev (USA)
Radoslav Pavlov (Bulgaria)
Rumiana Kirkova (Bulgaria)
Stanimir Stoyanov (Bulgaria)
Stefan Dodunekov (Bulgaria)
Stoyan Poryazov (Bulgaria)
Vladimir Ryazanov (Russia)

**Organizing Committee:**

Ilia Mitov
Emilia Saranova
Tsvetanka Kovacheva
Krassimira Ivanova
Rositsa Ovcharova
Valeria Dimitrova
Stoyan Poryazov
Todorka Kovacheva
Vera Markova

**Papers of i.TECH 2007 are collated in following sections:**

- Multimedia Semantics and Cultural Heritage
- Knowledge Discovery and Engineering
- Transition P Systems
- Neural Nets
- Decision Making
- Software Engineering
- Advanced Technologies
- Distributed and Telecommunication Systems
- Cyber Security

**Official languages of the conference are English and Russian.**

**General sponsor of the i.TECH 2007 is FOI BULGARIA ( www.foibg.com ).**

# TABLE OF CONTENTS – VOLUME 1

## Multimedia Semantics and Cultural Heritage

## Knowledge Discovery and Engineering

## Transition P Systems

## Neural Nets

# TABLE OF CONTENTS – VOLUME 2

## Decision Making

## Software Engineering

## Advanced Technologies

## Distributed and Telecommunication Systems

## Cyber Security

## About:

# INDEX OF AUTHORS

## Multimedia Semantics and Cultural Heritage

# MULTIMEDIA RETRIEVAL - STATE OF THE ART

Peter L. Stanchev

*(Keynote Speech)*

In this keynote speech we discuss the history, the state of art, and the future of multimedia retrieval [6]. Wide access to large information collections is of great importance in many aspects of everyday life. For this reason, significant effort has been spent in studying and developing techniques that support effective and efficient retrieval of multimedia data. Every day we are overwhelmed by information of many types: TV channels, news feeds, web sites, to name a few. Without an efficient and effective retrieval and filtering support, much time and effort is required in finding the information that we really need in this highly dynamic information age.

The process of image description consists of extracting the global image characteristics, recognizing the image-objects, and assigning semantics to these objects. The image data can be treated as physical image representation and its meaning as a logical image representation. The logical representation includes methods for describing the image and image-objects characteristics and the relationships among the image objects. Several visual descriptors exist for representing the physical content of an image, such as the MPEG-7 standard [5]. Historically, semantic retrieval was frequently based on computer vision. To reduce the semantic gap, the low-level content-based media features are frequently being converted to high-level concepts or terms.

The MPEG-7 descriptors can be classified as general visual descriptors and domain specific descriptors. The former include color, texture, shape, and motion features. The latter includes face recognition descriptors. Color is one of the most widely used image and video retrieval feature. The MPEG-7 standard includes five color descriptors, which represent different aspects of the color and include color distribution, spatial layout, and spatial structure of the color. The image texture is one of the most important image characteristic in both human and computer image analysis and object recognition. Visual texture is a property of a region in an image. There are two texture descriptors in MPEG-7: a homogeneous texture descriptor and edge histogram descriptor. Both of these descriptors support search and retrieval based on content descriptions. MPEG-7 supports region and contour shape descriptors. Object shape features are very powerful when used in similarity retrieval. Although distance functions are not part of the standard, we will present the most used distance functions.

A technique for improving the similarity search process of images in a Multimedia Content Management System is analyzed. The content based retrieval process integrates the search on different multimedia components, which are linked in XML structures. Depending on the specific characteristics of an image data set, some features can be more effective than others when performing similarity searches [2]. Based on this observation we propose a technique that predicts the effectiveness of the MPEG-7 image features that depends on a statistical analysis of the specific data sets in the Multimedia Content Management System. This technique is validated through extensive experiments with human subjects.

We illustrate several aspects of the fine art databases [4]. We showed that MPEG-7 descriptors can be used, but they give different results than applying on other type of images. The use of the Color Structure descriptor only produces sufficiently efficient results in the query search. The new generation Semantic Web languages, such as RDF(S) and OWL will play a major role. The integration of semantic understanding of pictures with personalized delivery raises new questions. The query language for this type of system is not yet scandalized but we hope that an emerging standard will come soon.

We discuss the problems witch arise working with Magnetic Resonance (MR) images. As an illustration of medical image processing tools we discuss MR brain segmentation problems. Functional analysis of different medical systems is made. We emphasize on the fact that working with medical images is different from working with other kind of images. As an illustration two systems are presented [3]. The first system is MEDIMAGE, which is a multimedia database for Alzheimer's disease patients. It contains MR images, text and voice data and it is used to find some correlations of brain atrophy in Alzheimer's patients with different demographic factors. The second system is Epilepsy system, which includes image data from MRI and SPECT, scans and EEG analysis results and it is used for patients with epilepsy.

We present a novel approach for efficient video stream filtering that is based on the use of the MPEG-7 descriptors and exploits the properties of metric spaces in order to reduce the computational load of the filtering receiver [1]. Among other types of information, Audio/Video can be considered today as a primary means of communication, due to its richness in informative content and to its appeal. This implies that the development of techniques supporting the retrieval of Audio/Video documents is of primary importance to enable the access for the general public as well as for professional users of significant asset of today's life. This process will gain more impetus from the adoption of standards to represent video content. The retrieval process is based on a simple schema: users specify their request needs (e.g. a set of keywords or a sample image) that are translated into a system query. The items in the archive are compared with the user's query, in order to determine if they are relevant for the user's request. To process this type of query, it is necessary to determine a set of properties of the objects stored in the archive (usually called features) and a similarity measure to compare queries and archive objects. Video features can be described by using the MPEG-7 standard. In case the similarity measure is metric, many possible approaches to create indexes can be adopted. These indexes allow improving the efficiency of the retrieval process, by comparing the query only with a limited number of objects in the archive. Our approach goes toward a solution of this problem, by proposing a novel approach to Audio/Video filtering that makes use of simple additional information sent together with the video. This allows avoiding the comparison of the filter with video features for many non-matching videos or video components that will not pass the filter in any case. The proposed approach requires that the measure of the similarity between the filter and the video representative is metric, and it is based on the use of the well known technique of pivots.

The following systems are discussed: Tamino, TV-Anytime, MILOS, PicToSeek, Marvel, imgSeek, IKONA, SIMPLIcity, ALIP, SIMBA, Viper, LCPD, Video Google, Cortina, Octagon, PicSOM, LIRE. Semantic retrieval in TREC is also presented.

The specific challenges in the field are highlighted. Conclusion remarks about the future of the multimedia retrieval are drowning.

## Bibliography

[1] Falchi F., Gennaro C., Savino P., Stanchev P., Efficient Video Stream Filtering, accepted for *ACM Multimedia,* 2007

[2] Stanchev P., Amato G., Falchi F., Gennaro C., Rabitti F., Savino P., "Selection of MPEG-7 Image Features for Improving Image Similarity Search on Specific Data Sets", *7-th IASTED International Conference on Computer Graphics and Imaging, CGIM 2004*, Kauai, Hawaii, 2004, 395-400.

[3] Stanchev P., Fotouhi F., Siadat M-R., and Soltanian-Zadeh H., Multimedia Mining- "A High Way to Intelligent Multimedia Documents", *book chapter 7- Medical Multimedia and Multimodality Databases,* D. Djeraba (Eds) - Kluwer Accademic Publishers, 2002, 139-160.

[4] Stanchev P., Green Jr D.., Dimitrov B., Semantic Notation and Retrieval in Art and Architecture Image Collections, *Journal of Digital Information Management, Vol. 3, No. 4, December 2005*, 218-221

[5] Stanchev P., Introduction to MPEG 7: Multimedia Content Description Interface, International *Conference on "Visualization, Imaging and Image Processing (VIIP 2004),* September 6-8, 2004, Marbella, Spain

[6] Stanchev P., Semantic Video and Image Retrieval - State of the Art and Challenges, ACMSE 2007, *The 45th ACM Southeast Conference*, Winston-Salem, North Carolina, USA, March 24, 2007, 512-513

## Author's Information

*Peter Stanchev is a Professor of Computer science at Kettering University, Flint, Michigan, USA and Professor and Department Chair at the Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria. He is also affiliated with the Institute of Information Science and Technologies, Italian National Research Council, Pisa, Italy. He has published two books, more than one hundred and fifty papers in monographs, journal, and peer-reviewed conferences that have more than 250 citations. His research interests are in the field of Image Processing, Multimedia Systems, Database Systems, and Expert Systems. He lectures courses in the areas of: Database Systems, Information Retrieval, Data Mining, Human-computer interaction, Computing and Algorithms, Web technology, Computer Architecture, Design of Information Systems, Computer Operating Systems, Image Databases, Fuzzy sets, Decision Support Systems, Multimedia Databases, Expert Systems, Image Processing, Computer Graphics and Game design. He is a Co-Chair of the annual multimedia semantics workshops.*

***Peter L. Stanchev** – e-mail: pstanche@kettering.edu; www.kettering.edu/~pstanche*
*Kettering University; Flint, MI, USA 48504;*
*Institute of Mathematics and Informatics, Sofia-1113, acad. G.Bontchev, bl.8, Bulgaria;*

# CREATION OF A DIGITAL CORPUS OF BULGARIAN DIALECTS

## Nikola Ikonomov, Milena Dobreva

***Abstract**: The paper presents our considerations related to the creation of a digital corpus of Bulgarian dialects.*

*The dialectological archive of Bulgarian language consists of more than 250 audio tapes. All tapes were recorded between 1955 and 1965 in the course of regular dialectological expeditions throughout the country. The records typically contain interviews with inhabitants of small villages in Bulgaria. The topics covered are usually related to such issues as birth, everyday life, marriage, family relationship, death, etc. Only a few tapes contain folk songs from different regions of the country.*

*Taking into account the progressive deterioration of the magnetic media and the realistic prospects of data loss, the Institute for Bulgarian Language at the Academy of Sciences launched in 1997 a project aiming at restoration* *and digital preservation of the dialectological archive. Within the framework of this project more than the half of the records was digitized, de-noised and stored on digital recording media. Since then restoration and digitization activities are done in the Institute on a regular basis. As a result a large collection of sound files has been gathered.*

*Our further efforts are aimed at the creation of a digital corpus of Bulgarian dialects, which will be made available for phonological and linguistic research. Such corpora typically include besides the sound files two basic elements: a transcription, aligned with the sound file, and a set of standardized metadata that defines the corpus. In our work we will present considerations on how these tasks could be realized in the case of the corpus of Bulgarian dialects. Our suggestions will be based on a comparative analysis of existing methods and techniques to build such corpora, and by selecting the ones that fit closer to the particular needs. Our experience can be used in similar institutions storing folklore archives, history related spoken records etc.*

***Keywords**: phonology, corpus, corpus linguistics, audio archive, digitization, restoration, metadata, alignment, transcription, phonetics*

## Definition of a Corpus

In the broad sense *corpus* means a collection of data, either written texts or a transcription of recorded speech which can be used for linguistic description or language studies. According to this definition we still cannot speak about a digital corpus of Bulgarian dialects, because all we have on hand at the moment are… a pile of magnetic tapes and numerous audio files stored on digital media, which represent the dialectological archive of the Institute for Bulgarian language.

## The Dialectological Archive

The dialectological archive of the Bulgarian language consists of over 250 audio tapes. All tapes were recorded between 1955 and 1965 in the course of regular dialectological expeditions throughout the country. Interviewers are dialectology researchers. The records typically contain interviews with aged people from small villages in Bulgaria. The topics covered are usually related to such issues as birth, everyday life, marriage, family relationships, death, etc. Only a few tapes contain folk songs from different region of the country.

## The First Project

In 1997 the Institute for Bulgarian Language started a 2-years project with the support of the British council the following basic aims:

- to secure the further preservation of the audio tapes;
- to start digitization of the records and their storage on a digital recording media

In order to secure the preservation of the audio archive we re-considered following basic issues, which are relevant when handling and storing sound recordings:

- that they be kept free of any foreign matter deposits;
- that they be kept free of any pressure that might cause deformations; and
- that they be stored in a stable, controlled environment.

Till 1997 none of these requirements has been met. The results of the inspection have forced us to take urgent measures to ensure suitable storage conditions for the tapes. At the same time we started to digitize the archive records and to store the digital content on CD's. In the framework of the Project more than 30% of the records have been digitized.

The workflow included the following steps:

- Digitization (sampling frequency of 44.1 KHz,16 Bits, Stereo, using a professional sound card equipped with a high end ADC);
- Digital restoration (elimination of most frequently encountered disturbances: impulsive disturbances, wideband noise, and harmonic disturbances);
- Recording on a CD-R.

## The current situation

Since 1999 digitization activities have been done in the Institute for Bulgarian Language on a regular basis Recently we changed the output format of the digital records from “wav” to “mp3” in order to save storage space. We also changed the recording media from CD-R to DVD for the same reasons. Due to the lack of financing the number of non-digitized tapes is still considerable (about 40%). The digitized records are not published electronically. Doing dialectological research under such circumstances is not easier than it was in the 50’s.

## What to do?

The solution is obvious – to create a digital dialectological corpus and make it available to the research community in a variety of formats:

- digitized sound (partly available);
- standard orthographic transcription;
- phonetic transcription;
- various levels of tagged text, all aligned.

Who will benefit from such an endeavor? First of all this will be the scientific community especially in such branches as:

- Arts and Humanities (cultural theory; history/geography and gender studies, linguistics, corpus linguistics, historical linguistics, speech recognition, text synthesis; dialectology);
- Sociology, social history and sociolinguistic research;
- Ethnography and cultural studies.

On the other hand, the experience which will be acquired throughout the project will serve the needs of various institutions with similar audio archives – folklore archives, history related records, etc. Last but not least the wide access to the data within the dialectological corpus will provide valuable information for lay persons, especially members of the local communities.

## Coding and coding standards

According to the basic requirements each speech corpus designed for phonological research must as a minimum consist of the following:

- A sound file;
- An orthographic transcription aligned with the sound file;
- A set of standardized metadata that defines the corpus.

## Sound files

For the completeness of the corpus all available tapes in the archive have to be further digitized and restored. It must be taken into account that the restoration and processing of the digital audio files are aimed only at making their content available; hence the playback quality is not a relevant parameter. In other words, cost reasons will specify the depth of the restoration efforts.

## The transcriptions

Transcription is the conversion into written form, of a spoken language source. There are different types of transcription:

- Orthographic transcription, which is done according to the basic orthographic rules of a corresponding language;
- Phonetic or phonemic transcriptions, which is the process of matching the sounds of human speech to special written symbols (IPA and its ASCII equivalent, SAMPA for example) using a set of exact rules, so that these sounds can be reproduced later. Phonetic transcriptions present three well known problems. They are hugely time-consuming and subjective in the sense that different transcribers typically produce different representations for a given speech segment. As the size of the corpus grows, so does the difficulty of maintaining consistency of practice across the transcription.

For cost as well as reliability reasons, the basic transcription of the sound file must be orthographic. But even if orthographic transcriptions are less costly and more reliable, defining standards for consistent transcription of speech by means of standard orthography is not trivial, and must be addressed. Transcription and sound must be aligned, so that the sound corresponding to a specific part of the transcription can be easily accessed.

Depending on the goals of a specific project, other types of transcriptions, such as phonetic or phonemic transcription, may be added, but they should not supplant the orthographic transcription. Different projects will have different needs for phonological tiers, depending on different kinds of use. The number of tiers is in principle limited, but a recommended list of relevant tiers might be useful.

A basic problem with all transcriptions is that they are products of interpretation. The result is that people do not trust each others transcriptions. Within a more long term perspective, the possibility of automatic transcriptions, which will make transcriptions at least more objective, (but not necessarily more correct), should be investigated.

## The Metadata

Researchers need standards for coding of metadata in order to be able to work on each other's databases. Two basic questions have to be answered:

- What are the relevant metadata?
- How should they be coded?

In other words a specification of the relevant metadata is needed before we can decide how to code them. Here the question arises whether it is possible to define a set of metadata that is relevant for all projects, and whether project specific need to code additional metadata should be catered for by means of a set of general guidelines. Our research has shown that the IMDI[1] (ISLE[2] Meta Data Initiative) already offers a standard for different kinds of metadata.

How should metadata be coded? Up to now it has been a standard practice for corpus creators to define their own representational standards. The drawback of such an approach is that they are not easily portable. In response various standards have been proposed. The currently dominant one is the Text Encoding Initiative (TEI). However there are two basic problems with TEI for representation of phonetic /phonological corpora:

- The TEI recommendation for linguistics corpora is vestigial, and needs to be further developed if it is to be useful for any but the most basic representations.

[1] http://www.mpi.nl/IMDI/

[2] ISLE stands for International Standard for Language Engineering

- The overabundance of XML tags makes TEI-encoded corpora difficult to use directly, and requires development of XML-based analytical applications. Few of these exist currently.

Pending their appearance, we have accepted TEI as an archiving standard. We expect that TEI will be supplemented by provision of XSLT[1], tools which translate TEI representation into formats usable by existing non-XML-aware applications like relational databases.

As to the coding itself, XML should be recommended. It is flexible, and allows users to define their own tags. An important question is whether only standards for coding metadata should be recommended, or whether the coding standards should be extended to the linguistic content as well. The latter position implies that tags will reflect theoretical positions.

## Digital Corpus of Bulgarian Dialects (DCBD)

The content of the Digital Corpus of Bulgarian Dialects (DCBD) will be provided in several types of representation:

- Audio (partly available);
- orthographic transcription;
- part-of-speech tagged orthographic transcription;
- phonetic transcription.

The orthographic transcription of DCBD will contain a complete orthographic transcription of the audio recordings. The transcription process will consist of several (up to four) passes through the audio files. The first pass will produce a base text. The next passes (usually the second and the third) are correction passes aimed at improving the transcription accuracy. The last pass will be used for establishing uniformity of the transcription algorithm across the entire corpus. To avoid pre-judging discourse structure, capitalization and punctuation will not be used in the transcription. As a general principle, the DCBD will use the Standard Bulgarian orthography. In genuinely dialectal segments, it will use the Bulgarian dialect dictionary (in preparation).

The part-of-speech tagged transcription is a morphological - syntactic annotation. It represents the basic linguistic analysis. It will be done automatically using software tools called taggers. The tagger for Bulgarian texts is called "GrammLab“and is distributed freely by BACL (Bulgarian Association of Computer Linguistics).

The phonetic transcription is in fact discretization of the analog speech signal into phonetic segment sequences. DBCD will contain phonetic transcriptions of all the interviews. The process will include following basic steps:

- Selection of transcription scheme, that is, a set of symbols each of which represents a single phonetic segment (for example IPA)
- Partition of the linguistically-relevant parts of the analog audio stream such that each partition is assigned a phonetic symbol.
- The result will be a set of symbol strings each of which will represent the corresponding interview phonetically. These strings can then be compared and processed.

The usefulness of the DBCD would be enhanced by provision of an alignment mechanism to relate the representational types to one another, so that corresponding segments in the various types can be conveniently identified and simultaneously displayed. The first task in this process is the necessity to define how large the alignment segments should be - phonetic segment by phonetic segment, word-by-word, sentence by sentence, or utterance by utterance? The answer has to take into account two basic factors: the research utility, and the feasibility in terms of cost.

[1] eXtensible Stylesheet Language Transformations

All interviews consist of a sequence of “interviewer-question, interviewee-answer” pairs in which the utterance boundaries are generally clear-cut; rarely there is some degree of overlap on account of interruption and third-party intervention. The format of the interviews makes alignment at the granularity of utterance the natural choice.

In practice the alignment process has to take into account, that time is a meaningful parameter only for the audio level of representation in the corpus, and that text has no temporal dimension.

A time interval *t* is selected, and the audio level is partitioned into some number *n* of length-*t* audio segments *s*, $s(t \times 1)$, $s(t \times 2)$...$s(t \times n)$, '×' denotes multiplication.

Corresponding markers are inserted into the other levels of representation such that they demarcate substrings corresponding to the audio segments. In XML such marker could be the `<anchor>` tag), where, the `'id'` attribute will specify a real-time offset from the start of the audio file.

## Future cooperation

The future cooperation in the field will be achieved in the frames of the network *European Corpus Phonology Group (CorPho)*. It will assemble researchers and research teams interested in combining insights from theoretical phonology, both diachronic and synchronic, linguistic variation studies, phonetics, and corpus linguistics.

## Bibliography

[Clua, 2006] Clua Esteve; Lloret, Maria-Rosa “New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)”. In: Montreuil, Jean-Pierre (ed), *New Perspectives on Romance Linguistics. Vol. 2: Phonetics, phonology, and dialectology.* Amsterdam / Philadelphia: John Benjamins. (Available at: http://www.uv.es/foncat.)

[Ihalainen, 1990] Ihalainen, Ossi “A source of data for the study of English dialect syntax: The Helsinki Corpus.” In Aarts, Jan & Willem Meijs (eds) *Theory and Practice in Corpus Linguistics.* Amsterdam: Rodopi. 83-103.

[Lloret, 2002] Lloret Maria-Rosa; Perea, M. Pilar “A Report on Corpus Oral Dialectal del Català Actual (COD)”. *Dialectologia et Geolinguistica* 10: 1-18.

[Meurman-Solin, 2001] Meurman-Solin, Anneli. “Structured Text Corpora in the Study of Language Variation and Change”, *Literary and Linguistic Computing*, Vol. 16, No. 1, 5-27.

[Moisl, 2005] Moisl H., Jones V., “Cluster analysis of the Newcastle Electronic Corpus of Tyneside English: a comparison of methods”, Literary and Linguistic Computing 20, 125-46.

[Moisl, 2006] Moisl H., Maguire W, Allen W., “Phonetic variation in Tyneside: exploratory multivariate analysis of the Newcastle Electronic Corpus of Tyneside English”. In: F. Hinskens, ed. Language Variation. European Perspectives. Amsterdam: Meertens Institute.

***Web resources***

[NECTE] A linguistic time-capsule: the Newcastle electronic corpus of Tyneside English (available on http://www.ncl.ac.uk/necte/).

## Authors' Information

***Nikola Ikonomov**– Head of Laboratory on Phonetics and Speech Communication, Institute for Bulgarian Language, BAS, Shipchenski prohod 52, Sofia-1113, Bulgaria, Institute for Mathematics and Informatics, e-mail: nikonomov@ibl.bas.bg.*

***Milena Dobreva** – Head of Dept. on Digitization of Scientific Heritage, Institute of Mathematics and Informatics, BAS, Acad. G. Bonchev St., bl. 8, Sofia-1113, Bulgaria, e-mail: dobreva@math.bas.bg*

# KNOWLEDGE TECHNOLOGIES FOR DESCRIPTION OF THE SEMANTICS OF THE BULGARIAN FOLKLORE HERITAGE[*]

**Desislava Paneva, Konstantin Rangochev, Detelin Luchev**

***Abstract**: Preserving and presenting the Bulgarian folklore heritage is a long-term commitment of scholars and researchers working in many areas. This article presents ontological model of the Bulgarian folklore knowledge, exploring knowledge technologies for presenting the semantics of the phenomena of our traditional culture. This model is a step to the development of the digital library for the "Bulgarian Folklore Heritage" virtual exposition which is a part of the "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" project.*

***Keywords**: Knowledge Technologies, Ontology, Digital Libraries, Bulgarian Folklore, Ethnology.*

***ACM Classification Keywords:** I.2.4 Knowledge Representation Formalisms and Methods, H.3.7 Digital Libraries – Collection, Dissemination, System issues.*

## Introduction

In the first ICT Work Programme under the Seven Framework Programme of the European Community for Research and Technological Development (FP7), which defines the research priorities for 2007-2008, cultural heritage research is part of Challenge 4, named "Digital Libraries and Content". Its main objective is the development of "large-scale European-wide digital libraries of cultural and scientific multi-format and multi-source digital objects, assisting communities of practice in the creative use of content in multilingual and multidisciplinary contexts, and based on robust and scalable environments, cost-effective digitisation processes, semantic-based search facilities and tools for digital preservation" [ICT, '07].

The "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage" (FolkKnow) project follows this idea and aims to build a multimedia digital library with a set of objects/collections, selected from the fund of the Institute for Folklore of the Bulgarian Academy of Science, which corresponds to the European and world requirements for such activities, and is consistent with the specifics of the presented artefacts. The project will use the knowledge technologies and digital libraries as they are the most suitable tools for semantic description and virtual multimedia presentation of cultural historical artefacts.

This paper presents the first stage of the work done on module 3 of the project, named "Development of Digital Libraries and Information Portal with Virtual Exposition "Bulgarian Folklore Heritage"". It tracks out the creation of Bulgarian folklore ontology, describing the knowledge about Bulgarian folklore objects and their main features, technical data or context. This ontology is the backbone of the subsequent work of folklore digital library development. Section 2 of the paper is a short description of the main issues of the FolkKnow project. In section 3, a brief outlook of the project's digital library is included. Section 4 summarises the current state of ontology development. Sections 5 and 6 deal with different aspects of the Bulgarian folklore ontology including its scoping, conceptions, relations, and its implementation and utilization in the project.

[*] This work is supported by National Science Fund of the Bulgarian Ministry of Education and Science under grant No **IO-03-03/2006** "Development of Digital Libraries and Information Portal with Virtual Exposition "Bulgarian Folklore Heritage" from the project "Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage".

## Knowledge technologies for creation of digital presentation and significant repositories of folklore heritage

The aim of the project “Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage” is to build a multimedia digital library with a set of objects/collections, selected from the fund of the Institute for Folklore of the Bulgarian Academy of Science, which corresponds to the European and world requirements for such activities, and is consistent with the specifics of the presented artefacts. The complex structure and the multi-layer characters of the folklore objects require innovative approach for knowledge representation. The rich-in-content Web-presenting of the Bulgarian folklore knowledge defines the usage of modern methods and technologies for digital archive developing, which will be used not only for preservation and access to the information, but as a tool for scientific research analysis development. The main tasks is to create a digital library and information artery using knowledge-based technologies and Semantic web approach, in order to present in virtual form the valuable phenomena of the Bulgarian folklore heritage. The realization of the project gives possibility for wide social applications of the multimedia collections for the purposes of Interactive distance learning/self-learning, research activities in the field of Bulgarian traditional culture, and for the cultural and ethno-tourism in Bulgaria.

We assume that when Bulgarian folklore heritage is digitalized and presented virtually there will be a need of contemporary information technologies that allow complex multimedia presentations and descriptions, as well as broad and flexible access method. We believe that the digital libraries and Semantic web meet those requirements because they are powerful technologies for digitalization, semantic description, access provisioning, preservation, and virtual representation of cultural and historic values and especially of Bulgarian folklore heritage. The approach for building the module is formed as a result of the research experience of the team of Institute of Mathematics and Informatics and its know-how in multimedia applications gained in numerous European Information Society projects. It includes analytical research, choice and usage of suitable methods, tools and environments for digital representation and preservation of significant cultural and historical artefacts and their exposure into the global information space. This approach allows the integration of the idea of the traditional Bulgarian culture and folklore in the European culture space, while completely preserving its identity and diversity [Bogdanova et al. ‘06].

## Digital library of Bulgarian folklore heritage

Digital libraries are a contemporary conceptual solution for access to information archives. They contain diverse hypertext-organized collections of information (digital objects such as text, images, and media objects) that are organized thematically and are managed by complex specialized services such as semantic-based search, multi-layer and personalized search, context-based search, relevance feedback, resource and collection management, metadata management, indexing, semantic annotation of digital resources and collection, grouping and presenting digital information, extracted from a number of locations, services for digital information protection and preservation, *etc.* [Pavlov&Paneva, ‘06].

Besides that the flexibility, the automatic adaptation, the access anywhere and anytime, the decentralization, the wide variety of digital objects and collections, the information security, *etc.* are already key requirements for the advanced multimedia digital libraries [Pavlova-Draganova et al., ‘07] [Paneva et al., ‘05] [Pavlov et al., ‘06a] [Pavlov et al., ‘06b].

Information about the actual state of the research of the architecture of digital libraries, informational access to audio-visual and non-traditional objects and semantic interoperability is contained in the FP6 project DELOS “A Network of Excellence on Digital Libraries” (http://www.delos.info).

Having in mind this variety of useful properties and characteristics of the large-scale repositories of digitized knowledge their use for presentation of the valuable phenomena of the Bulgarian folklore is not casual. There are some national investigations and projects concerning the virtual existence and the digitalization of ethnographic and folklore artefacts, for example, experimental digital archive "Bulgarian Ethnographic Treasury" (http://mdl.cc.bas.bg/ethnography/) [Luchev, '05] [Luchev, '06], project "Yuper" (http://yuper.hit.bg/), project "Folklore Motives and Anthologies" (http://liternet.bg/), project "WebFolk Bulgaria" (http://musicart.imbm.bas.bg/EN/Default.htm), project "Living Human Treasures" (http://www.treasures.eubcc.bg/main.php), project "Virtual Encyclopaedia of Bulgarian Iconography" (http://mdl.cc.bas.bg/), *etc.*

The FolkKnow multimedia digital library can be similar valuable gallery of artefacts and knowledge for Bulgarian culture, art and folklore that will present a relatively limited number of specimens of different folklore narrative types (songs, rituals, faith, knowledge, proverbs, magic, *etc.*) and their audio-visual documentation. Until now, the Bulgarian folklore is always shown partially only with text, sound or image, but the authors' demand is for joint unities of words, music and motions. This possibility can be provided by contemporary multimedia environments. The ambitions of the authors are the demonstration of unique music dialects from different local folklore areas and advanced approaches for folklore content prescription representation through authentic sounds, videos, and photos of live rituals. Part of the Bulgarian folklore specimens will be presented from asynchronous point of view; other will be in their diachrony – unique materials, saved for years. Another task is the different record technique demonstration – inquiry, interview, inclusive observation, *etc.*

Multimedia digital library of Bulgarian folklore expects a wide range of potential users – professionals and scientists, non-professionals, connoisseurs and viewers, *etc.*

## Ontological presentation of folklore knowledge

Originally, the term ontology comes from philosophy where it is employed to describe the existence of beings in the world. In 1993, Gruber's definition becomes the most referenced on the knowledge technologies literature: "an ontology is a formal, explicit specification of a shared conceptualization" [Gruber, '93]. Conceptualization refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. Explicit means that the type of concepts used and the constraints on their use are explicitly defined. Formal refers to the fact that the ontology should be machine readable, which excludes natural language. Shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

Ontologies can be used for many different purposes. The literature on knowledge representation contains research on the use of ontologies for data-interchange, for data-integration, for data-querying or for data visualization. In general, visualization of information can be seen as a two-step process. In a first step, information is transformed into some intermediate semantic structure. This structure organizes the raw information into a meaningful structure. In a second step, this semantic structure is used as the basis for a formal visual representation. We will use this approach in our work on the Bulgarian folklore ontology development.

Tools for building ontologies usually provide a graphical user interface that allows ontologists to create ontologies without using directly a specific ontology specification language. Some tools have been created for merging and integrating ontologies [Fensel, '04].

Recently, many ontology languages have been developed in the context of the World Wide Web: Resource Description Language (RDF), RDF Schema, Simple HTML Ontology Extensions (SHOE), Ontology Exchange Language (XOL), Ontology Markup Language (OML), Web Ontology Language (OWL), Ontology Inference Layer (OIL), DAML+OIL, *etc.* Their syntax is based on XML, which has been widely adopted as a 'standard' language for exchanging information on the web [Fensel, '04].

To efficiently represent the folklore annotation framework and to integrate all the existing data representations into a standardized data specification, the folklore ontology need to be represented in a format (language) that not enforce semantic constraints on folklore data, but can also facilitate reasoning tasks on folklore data using semantic query algebra. This motivates the representation of Bulgarian folklore ontology model in Web Ontology Language (OWL). OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema by providing additional vocabulary along with a formal semantics. Knowledge captured from folklore data using OWL is classified in a rich hierarchy of concepts and their inter-relationships. OWL is compositional and dynamic, relying on notions of classification, reasoning, consistency, retrieval and querying. We investigated the use of OWL for making Bulgarian folklore ontology using Protégé OWL Plug-in.

## Ontology of Bulgarian folklore

Since one of the targets of the FolkKnow project is to present the valuable phenomena of the Bulgarian folklore in suitable virtual form using knowledge technologies, we have to observe and specify the experience that has been gained in the last 500 years in the area of traditional folklore *i.e.* to construct Bulgarian folklore domain ontology.

FolkKnow annotator/indexers using this ontology will semantically describe and index the raw audiovisual content in order to create and maintain reusable digital objects.

The ontology will be used also to realize semantic-based access to concrete digital objects, representing folklore objects, described by their main features, technical data or context. All this information is included within the Folklore Ontology Concept – the root concept for the ontology.

The process of building of the Bulgarian folklore ontology for the FolkKnow project is necessarily iterative. The first activity is the definition of the scope of the ontology. Scoping has been mainly based on several brainstorming sessions with folklorists and content providers. Having these brainstorming sessions allowed the production of most of the potentially relevant terms. At this stage, the terms alone represented the concept, thus concealing significant ambiguities and differences of opinion.

A clear issue that arose during these sessions was the difficulty in discovering of definite number of concepts and relations between these concepts. The concepts listed during the brainstorming sessions were grouped in areas of work corresponding naturally arising sub-groups. Most of the important concepts and many terms were identified. The main work of building the ontology was then to produce accurate definitions.

### Description of the conceptions

The scientific classification and documentation of folklore objects provide folklorists and content generators with a rich knowledge background with plenty of multidimensional data and metadata. There is a special relation among the metadata, which reveals all the knowledge concerning the folklore object obtained from the classification procedure.

The folklore object is related to three levels of knowledge, enriched with a set of sub-levels of the data classification. All these levels of knowledge or "thematic entities" in the ontology conception are supported by the scientific diagnosis results and the related documentation.

The entity "Identification and description" consists of general historical data, identifying aspects such as title, language, archival signature, period, current location of the folklore object, annotation, first level description, second level description, *etc.*,

The entity "Technical" includes technical information both revealing the technologies used for folklore object capturing and recording, record situation, record type, record place, record date, main participants in the process (record maker and informant), *etc.*

These main entities and their metadata are supported, documented and provided by the scientific diagnosis, which has been applied to the folklore objects.

### Ontological model

We will present the Bulgarian folklore ontological model using classes of concepts, organized in taxonomy and table with properties.

Taxonomies are used to organize ontological knowledge using generalization and specialization relationships through which simple and multiple inheritances could be applied. Properties are an interaction between individuals of the domain-classes and the range-classes.

Figure 1 depicts the main concepts and properties in the Bulgarian folklore ontological model.

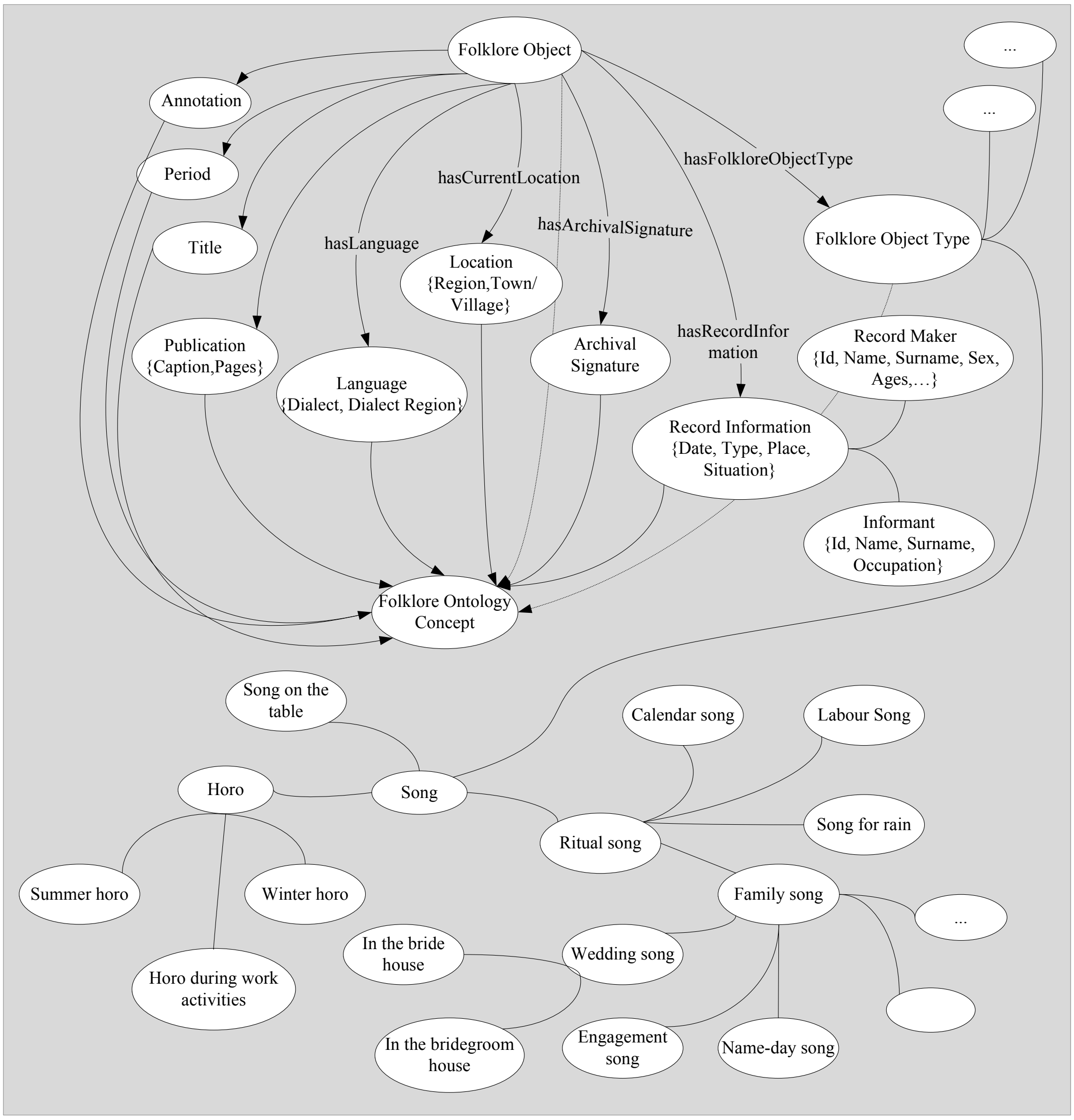


Figure 1: Part of the concepts and properties in Bulgarian folklore ontology

The most representative concepts have been defined first and then they have been specified appropriately in order to get a representation of the knowledge stored in the databases. The Bulgarian folklore ontology is composed of 70 concepts and 82 properties.

OWL properties represent semantic relationships between classes of objects. Below, a piece of the Bulgarian folklore ontology code, defining the property *isRecordPlaceOf* is presented. Here the declaration of the transitivity condition and the definition of property *hasRecordPlace* as its inverse can be seen.

```
<owl:ObjectProperty rdf:ID="isRecordPlaceOf">
  <rdfs:domain rdf:resource="#Record_Place"/>
  <owl:inverseOf>
    <owl:ObjectProperty rdf:ID="hasRecordPlace"/>
  </owl:inverseOf>
```

```
            <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
Represents that the element “Record_Place” is a record place of the folklore object.
            </rdfs:comment>
            <rdfs:range rdf:resource="#Folkore_Object"/>
         </owl:ObjectProperty>
         <owl:ObjectProperty rdf:about="#hasRecordPlace">
            <rdfs:range rdf:resource="#Record_Place"/>
            <rdfs:domain rdf:resource="#Folkore_Object"/>
            <owl:inverseOf rdf:resource="#isRecordPlaceOf"/>
            <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
Represents that the element “Folklore_Object” has a record place.
            </rdfs:comment>
         </owl:ObjectProperty>
```

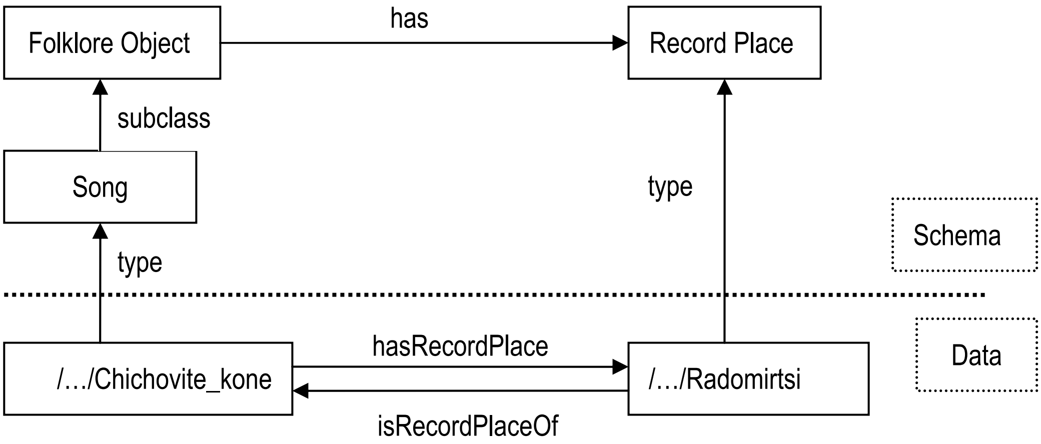


Figure 2: Scheme of relationships between classes of objects in Bulgarian folklore ontology

## Implementation of Bulgarian folklore ontology

Knowledge about Bulgarian folklore heritage and oral tradition is interesting not only for the wide audience of professionals (historians, philologists, psychologists, ethnologists etc), but also for non-professionals and institutions dealing with these problems. Folklore heritage specialists will reach to organized objects and semantically structured knowledge for their investigations. For example, the searching for an object “Ritual”, semantically connected with an object “Festival”, can give back not only the all rituals of the festival, but also the “Songs”, “Faith and Knowledge”, “Magic”, “Food”, *etc.*, semantically bound up with both “Ritual” and “Festival”. Cultural institutions and organizations, as well as non-professionals will be able to find information for semantically joined complexes of folklore objects on the base of starting points as “Location”, “Period”, “Language/Dialect”, *etc.*

The ontology gives the ability to describe the semantics of folklore content and to use new knowledge management services such as semantic search across aggregations of varied and complex sub-classes and objects in a robust, rich and user-friendly manner, personalized search, context-based search, multi-criteria search, metadata management, *etc.*

The semantically annotated objects can also be used as a base for eLearning courseware development; for example, folklore objects can be easily discovered and grouped in learning lessons, modules or parts of them.

## Bibliography

[Bogdanova et al. '06] Bogdanova G., Pavlov R., Todorov G., Mateeva V. (2006), Knowledge Technologies for Creation of Digital Presentation and Significant Repositories of Folklore Heritage, Advances in Bulgarian Science, pp. 7-15.

[Pavlov&Paneva, '06] Pavlov R., Paneva D. (2006), Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, Cybernetics and Information Technologies, vol. 6, № 3, pp. 51-62.

[Pavlova-Draganova et al., '07] Pavlova-Draganova L., Georgiev V., Draganov L. (2007), Virtual Encyclopeadia of Bulgarian Iconography, International Journal "Information Technologies and Knowledge", vol.1, №3, pp. 267-271.

[Paneva et al., '05] Paneva, D., Pavlova-Draganova L., Draganov L. (2005), Digital Libraries for Presentation and Preservation of East-Christian Heritage, Proceedings of the Second HUBUSKA Open Workshop "Generic Issues of Knowledge Technologies", Budapest, Hungary, pp. 75-83.

[Pavlov et al., '06a] Pavlov R., Pavlova-Draganova L., Draganov L., Paneva D. (2006), e-Presentation of East-Christian Icon Art, Proceedings of the Fourth HUBUSKA Open Workshop "Semantic Web and Knowledge Technologies Applications", Varna, Bulgaria, pp. 42-48.

[Pavlov et al., '06b] Pavlov, R., Paneva D., Pavlova-Draganova L., Draganov L. (2006), Digital Libraries with Multimedia Content and Applications in Bulgarian Cultural Heritage (Analytical study), State Agency for Information Technologies and Communication (SAITC), by contract № 8/21.07.2005 between IMI-BAS and SAITC, 2006, Sofia, Bulgaria, Available at: http://mdl.cc.bas.bg/Digital_libraries_with_multimedia_content_and_applications_in_Bulgarian_ cultural_heritage.pdf (accessed on April 10, 2007).

[ICT, '07] Information and Communication Technologies, Work Programme 2007-08, Available online: ftp://ftp.cordis.lu/pub/fp7/ict/docs/ict-wp-2007-08_en.pdf (accessed on April 10, 2007)

[Fensel, '04] Fensel D. (2004), Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Second edition.

[Gruber, '93] Gruber T. (1993), Towards Principles of the Design of Oncologist Used for Knowledge Sharing. International Journal of Human Computer Studies 43 907-928, Available at: http://www.itee.uq.edu.au/~infs3101/_Readings/OntoEng.pdf (accessed on April 10, 2007).

[Luchev, '06] Luchev, D. (2006), Experimental Digital Library - Bulgarian Ethnographic Treasury, Proceedings of the Modern (e-) Learning Conference, Varna, Bulgaria, pp. 106-111.

[Luchev, '05] Luchev, D. (2005), Experimental Digital Archive "Bulgarian Ethnographic Treasury" (Analytical study) by contract № 12/08.09.2005 between SAITC and EIM-BAS, IMI-BAS (subcontractor), Sofia, Bulgaria, Available at: http://mdl.cc.bas.bg/DigLibr_BulEthnoTreasure_Analytical_Study_Luchev.pdf (accessed on April 10, 2007).

## Authors' Information

***Desislava Paneva** - Institute of Mathematics and Informatics, BAS, Bulgaria, 8, Acad. G. Bonchev str.; e-mail: dessi@cc.bas.bg*

***Konstantin Rangochev** – Institute of Mathematics and Informatics, BAS, Bulgaria, 8, Acad. G. Bonchev str.; e-mail: krangochev@yahoo.com*

***Detelin Luchev** - Ethnographic Institute with Museum, BAS; Bulgaria, 1000 Sofia, Moskovska str. 6A; e-mail: luchev_detelin@abv.bg*

# ELECTRONIC PRESENTATION OF BULGARIAN EDUCATIONAL ARCHIVES: AN ONTOLOGY-BASED APPROACH

Anna Devreni–Koutsouki

***Abstract:*** *The paper presents an ongoing effort aimed at building an electronic archive of documents issued by the Bulgarian Ministry of Education in the 40ies and 50ies of the 20th century. These funds are stored in the Archive of the Ministry of the People's Education within the State Archival Fund of the General Department of Archives at the Council of Ministers of Bulgaria. Our basic concern is not the digitization process per se, but the subsequent organization of the archive in a clear and easily-searchable way which would allow various types of users to get access to the documents of interest to them. Here we present the variety of the documents which are stored in the archival collection, and suggestions on their electronic organization. We suggest using ontologies-based presentation of the archive. The basic benefit of this approach is the possibility to search the collection according to the stored content categories.*

***Keywords:*** *digitization, archives, history of education, ontologies, SWP.*

***ACM Classification Keywords:*** *H.5 Information Interfaces and Presentation, H.3.3 Information Search and Retrieval, H.3.7 Digital Libraries*

## Introduction

Digitization of cultural and scientific content in European countries is important field of work which results should contribute to the development of The European Library portal (TEL)[1]. Currently, there are numerous ongoing digitization projects and initiatives in libraries, archives and museums.

Within this general picture, extensive work on digital capture and exposure of educational archives has not been undertaken so far, according to our research. In the educational field most attention is concentrated on the development of e-learning applications while historical documents of the educational institutions and the governmental bodies shaping the policy in education and research field are still not digitized on mass scale.

However, such documents could be of interest not only to the researchers who study the development of the educational system (in one country or on comparative basis). Educational archives contain documents which could be of interest to the local historians, and to the general citizen.

Therefore, we decided to undertake an effort which would present in the electronic space the documents from the archive of the Ministry of Education of Bulgaria. We decided to start this effort with practical work on the documentation from the 40ies and 50ies of the 20th century, since this was one significant period of reform of the educational system in Bulgaria.

The presentation of educational archives also imposes some challenges.

### 1. Digitisation and metadata.

This type of archive contains quite diverse documents - official documentation, letters, notes, photographs, various documents, newspapers. The text documents can be printed, typewritten or handwritten. On the one hand, the digitization requires the application of different workflows. On the other hand, the metadata for these various documents, if detailed, should follow different structures.

[1] http://www.theeuropeanlibrary.org/portal/index.html, date of last visit March 21, 2006.

### 2. Presentation and use

There has been a standing issue coming from the past – the problem related to the storage and access provision to already created materials, which were not designed for computer processing. We envisage the vast amount of documents, forms, protocols, letters/correspondence, photographs, maps, images and other objects which could be found in private or public museum collections or state, local or personal archives. The educational archive is a typical example of such a diverse collection. How should this collection be organized in the electronic space? If it just follows the traditional archival structure, the search of documents would be very difficult – one would have to browse everything, or search for the exact document. The general user does not necessarily have this information, neither should he (she) be knowledgeable about the metadata used. Thus our work is directed to looking for better and more user-friendly ways to provide access to the electronic collection.

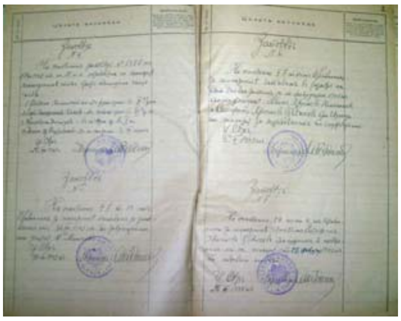
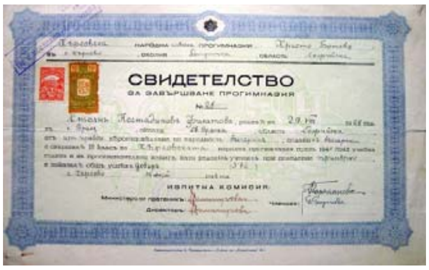
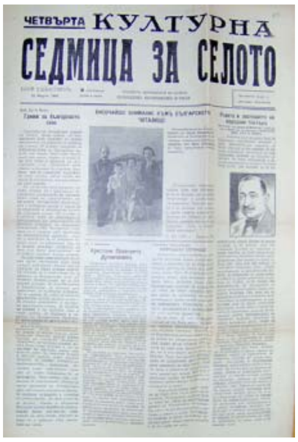
## The Archive and its Presentation

The idea for this effort was coined within a group of historians and education specialists from the University of Ioannina, Greece who work on comparative study of the Greek and the Bulgarian educational systems in the middle of the 20th century. Till now, the archives of the Bulgarian Ministry of Education (Ministry of the Enlightenment in the studied period) have been studied within 1940-1945. The sources are stored in funds 798k and 177k. of the Ministry of Peoples' Education 1879-1944.

Digital copies of several thousands of documents have been made. They are not sufficient for the purposes of the comparative study of the educational systems, but are sufficient for our purposes to suggest the organisation of the electronic collection and its use. The collected materials are interesting for the variety of types of sources they present. The next table summarizes the available document types which can be found as separate archival units. Here we do not discuss the issues of creating metadata on the whole inventory of documents, but rather describe the issues of describing the separate archival units.

Table 1. Samples of documents from the archives of the Bulgarian Ministry of Education, 1940-1945

| DOCUMENT TYPE | EXAMPLE | METADATA AND CONTENT PRESENTATION PROBLEMS |
|---|---|---|
| Handwritten texts (general purpose documents, orders, notes, etc.) | This type of documents is typical for all archival collections. | Metadata for describing archival units can be applied. If we aim full text presentation, we have to face a massive amount of hand text entry. Typical elements appearing in these documents are names (personal and place names), dates, affiliations. Such documents are interesting for study of the problems which circulated in the educational administration. |
| Typewritten documents (general purpose documents, orders, notes, etc.) in some cases with handwritten resolutions | This type of documents is also typical for all archival collections. We place it separately from the group above, because digitisation and processing of typewritten documents may involve OCR and the workflow would be different. | The same as above; OCR can be tested for text recognition. |

| | | |
|---|---|---|
| Handwritten documents presenting records related to the educational sector, in some cases with signatures and stamps | <br>Sample from Fund 798 k, inventory list 2, archival unit 98. Book of orders of the Seres High School | Here we can use again the general metadata. If we aim to present the full text we should re-create the structure. Additional issue is how to present structured data on stamps and signatures. |
| Typewritten documents presenting records related to the educational sector, in some cases with signatures and stamps | This type of documents is also typical for all archival collections. As with typewritten generic texts, we place these documents separately from the group above, because digitisation and processing of typewritten documents may involve OCR and the workflow would be different. | The same as above; OCR can be tested for text recognition. |
| Individual documents with signatures, postal stamps, state fee stamps | <br>СВИДЕТЕЛСТВО<br>ЗА ЗАВЪРШВАНЕ ПРОГИМНАЗИЯ<br>ИЗПИТНА КОМИСИЯ<br>Sample from Fund 798 k, inventory list 2, archival unit 114. Certificate for a completed educational degree, Seres High School | Here we can use again the general metadata. Again, an issue is how to present structured data on stamps and signatures. State stamps might be of interest, for example, to philatelists, i.e. in a very structured approach we should encode data on these objects too in order to make the information on them searchable. |
| Newspapers | <br>ЧЕТВЪРТА КУЛТУРНА<br>СЕДМИЦА ЗА СЕЛОТО<br>Sample from Fund 177 k, inventory list 2, archival unit 2251. Certificate for a completed educational degree, Seres High School<br>The newspaper contains orders of the Ministry of education, reports, letters of local administrations, materials about a cultural week of the village, etc. | Here we can use again the general metadata for the archival unit, but then we should decide how to present the contents of the newspaper. A highly structured approach would require to present the content in detail, and/or provide full text search capabilities. The photographs in the newspapers also should be considered as a separate object. |

| | | |
|---|---|---|
| Photographs | <br>Sample from Fund 177 k, inventory list 2, archival unit 2251. Certificate for a completed educational degree, Seres High School | The description of photographs differs from description of documents. Currently we study projects which deal with electronic presentation of historical photographs in order to suggest what metadata to use within the frameworks of our endeavour. |

This brief presentation illustrates some of the problems which we face:

- How detailed should be the presentations of the various types of documents? On the one hand, we might be tempted to provide full text for all documents, but is this effort justified?
- How exactly to present multimodal objects (as we see in the examples, we have special layouts in some cases; stamps; signatures; marginal notes, etc.).

We believe that one approach which makes such collections searchable even without the application of very detailed and fragments presentations is the proper use of ontologies. Below we will present briefly the concept of ontologies and then will present one possible practical solution, SWP.

## Ontologies

In philosophy, ontology (from the Greek ὄν, genitive ὄντος: of being (part. of εἶναι: to be) and -λογία: science, study, theory) is the study of being or existence. It seeks to describe or posit the basic categories and relationships of being or existence to define entities and types of entities within its framework. Ontology can be said to study conceptions of reality. It is often confused with epistemology, which is about knowledge and knowing.

According to recent artificial intelligence research "an ontology is a shared and common understanding of some domain that can be communicated across people and computers" [Gruber, 1993], [Guarino, 1995], [Borst, 1997] and [van Hejlst et al., 1997]. Ontologies can therefore be shared and reused among different applications [Farquhar et al., 1997]. "An ontology can be defined as a formal, explicit specification of a shared conceptualization" [Gruber, 1993], [Borst, 1997]. "Conceptualization" refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. "Explicit" means that the type of concepts used, and the constraints on their use are explicitly defined. "Formal" refers to the fact that the ontology should be machine-readable. "Shared" reflects the notion that ontology captures consensual knowledge, i.e. it is not private to some individual, but accepted by a group.

The concept of ontology is defined even narrower within the famous project *Ontolingua* of Stanford University [Ontolingua project]. It suggests that the ontology is an *explicit specification of some topic*. This approach presupposes formal and declarative presentation of a given topic, which includes the vocabulary (or names) for referring to the terms in that subject area and the logical statements that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other.

Ontology describes the subject matter using the notions of *concepts, instances, relations, functions,* and *axioms*. Table 2 presents the requirements to which ontology has to be compliant.

From the practical point of view, in the simplest case an ontology is „a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions))" [Noy, McGuiness]. If we take accept this as a rule of thumb, an ontology together with a set of individual instances of classes indeed can be seen a *knowledge base.* However, in reality, there is a fine line where ontology ends and the knowledge base begins – the latter can be more sophisticated presentation of a subject domain while ontology is always hierarchical and follows certain requirements as described above. From technological point of view, ontologies can be seen as knowledge bases of special kind, which can be "read" and understand, and could be shared between users and/or developers.

Table 2. Properties of ontologies.

| **Necessary properties of an ontology** | **Typical but not mandatory properties** | **Desirable properties, but not mandatory nor typical** |
|---|---|---|
| Finite controlled (extensible) vocabulary<br>Unambiguous interpretation of classes and term relationships<br>Strict hierarchical subclass relationships between classes | Property specification on a per-class basis<br>Individual inclusion in the ontology<br>Value restriction specification on a per-class basis | Specification of disjoint classes<br>Specification of arbitrary logical relationships between terms<br>Distinguished relationships, such as inverse and part-whole |

The basic reasons to create ontologies are summarized in [Noy, McGuiness] as follows:

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge

The development of ontologies is still a difficult and challenging task, because so far there are no common platforms and verified methods which would prescribe what procedures should be followed in the process of creating ontology. As [Jones et al.] explains it, „at present the construction of ontologies is very much an art rather than a science". This situation needs to be changed, and will be changed only through an understanding of how to go about constructing ontologies. In short what is needed is a good methodology for developing ontologies.

While there is no common methodology for building ontologies, there are principles for design and implementation suggested in [Gruber 1995]:

- **Clarity** – the ontology should present the terms included efficiently and without ambiguities. The definitions should be objective as much as possible, although the motivation for adding a term might be driven by the circumstances and the requirements for computability. A clear formalism should be used, and it is recommended to present the definitions in the form of logical axioms.
- **Coherence** – the definitions should be logically disambiguous, and all statements derived from the ontology should not be in disacordance with the axioms.
- **Extendibility** – the ontology should be designed so that the dictionaries of terms could be enlarged without revision of concepts already defined.
- **Minimal encoding bias** – the conceptual abstraction implemented in the ontology should be developed on the concept level instead of the level of the symbolic representation.
- **Minimal ontological commitment** – the ontology should contain only the most essential assumptions on the modeled world, so that there is enough space for making it wider or narrower.

**How do ontologies relate to our archival presentation task?** We believe that the use of ontology could be a good solution which would allow users to make a variety of searches within the collection of electronic documents while these documents are still not available in searchable full text format. If we incorporate as an element of the data several relevant ontological references, based on the assumptions for typical requests for information, the results returned to a query would include all documents which metadata are matching the concept from the ontology.

Definitely, this requires extra human effort: first, to develop a subject domain ontology (covering *educational administrative documentation*) – to the best of our knowledge such ontology does not exist, and moreover it would be specific for the Bulgarian documentary system; and second, to add references to the concepts from the ontologies within the archival units metadata. Compared to the creation of full text and sophisticated search tools, we believe that this approach will lead to fast and reliable results and will implement it in the nearest future.

## A Possible Practical Solution Involving Ontologies: SWP

Over the past few years, various approaches have been proposed to effectively organise digital content on the Web. Traditionally, these have included techniques such as building keyword indices based on image content, embedding keyword-based labels into images, analyzing text immediately surrounding images on Web pages, etc. Nevertheless, current Web technology presents serious limitations to make information accessible for users in an efficient manner. The general problem to find information on the Web is summarized in [Ding, Fensel 2001]: „searches are imprecise, often yielding matches to many thousands of hits". Moreover, users face the task of reading the documents retrieved in order to extract the information desired. These limitations naturally appear in existing Web portals based on this technology, making information searching, accessing, extracting, interpreting and processing a difficult and time-consuming task.

More recently, there has been a research focus on the Semantic Web technologies in different domains. The purpose of the Semantic Web is to create a universal medium for the exchange of sharable and processable data by automated tools, such as software agents, as well as by the users. "The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [Berners-Lee et al., 2001].

One solution to these problems is the use of Semantic Web Portals (SWPs), known also under the names *Knowledge Portals* or *Community Web Portals*.

There are different views in the research works on SWPs and their definition. An earlier study defines them as „portals for the information needs of particular communities on the web" [Staab et al., 2000]. According to [Gorcho, 2006] "a Semantic Web Portal is a Web application that offers information and services related to a specific domain, and that has been developed with Semantic Web technology". The same author emphasizes that the primary difference with the traditional Web Portals "is based on technological aspects: traditional Web portals are based on standard Web technology (HTML, XML, servlets, JSPs, etc.); semantic portals are based on that technology plus the use of Semantic Web languages like RDF, RDF Schema and OWL".

The SWPs which are well developed and functioning are not too many; also they are prone to various limitations. In [Karvounarakis et al, 2000] they are defined as Web applications that "provide the means to select, classify and access, in a semantically meaningful and ubiquitous way, various information resources (e.g., sites, documents, data) for diverse target audiences (corporate, inter-enterprise, e-marketplace, etc.)."

[Lausen et al., 2005] and [Lara 2004] offer more strict definition, which states that SWP has the following characteristics:

- It is a web portal. A web portal is a web site that collects information for a group of users that have common interests
- It is a web portal for a community to share and exchange information
- It is a web portal developed based on semantic web technologies.

The briefest but clear explanation is to view SWPs as "portals that typically provide knowledge about a specific domain and rely on ontologies to structure and exchange this knowledge" [Hartmann, Sure 2004]. The accent here is on the most typical feature of SWPs – their application in specific subject domains and the use of one or more ontologies as a backbone of the application.

Currently „SWP are still at their very early stages" [Lausen et al., 2005]. The benefits of implementing these Semantic Web technologies can be easily identified or foreseen as Semantic Web technologies have the potential to increase the information consistency and the information processing quality of portals. On the other hand, Semantic Web technologies themselves are still under development and most of the theoretical issues are no easy to be employed into real world applications.

## Conclusion

The national strategy of many countries, including private institutions, which possess such collections and archives, is making them widely-spread and accessible. The common practice is the creation of repositories of images or digital copies which can already be accessed through the Web [Hyvönen et al., 2004]. The management of such resources aims to reach maximum effectiveness of search in the sea of various forms of the stored information. Many of such collections currently exist and users are increasingly faced with problems of finding a suitable (set of) image(s) for a particular purpose. Each collection usually has its own (semi-) structured indexing scheme that typically supports a keyword-type search. However, finding the right image is often still problematic [Hollink et al., 2003].

In this paper we present a brief analysis of the types of documents in one particular Bulgarian archive (educational documentation from the 40ies and 50ies of the 20th century). We also made a brief overview of ontologies and SWPs which could help in structuring the electronic surrogates of archival documents. In our future work we will suggest ontology designed especially to present the documents from this archival collection and the implementation via SWP of search tools for use of the archive.

This collection of documents in the archive presented is highly fragile – already now the documents are deteriorating as it could be seen from the illustrations in Table 1. We hope that our effort will help to preserve for the future these documents which could be of interest to various groups of users.

## Bibliography

[Berners-Lee et al., 2001] Tim Berners-Lee, James Hendler and Ora Lassila, The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, Scientific American, 2001 (Visited 02-03-07) http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[Borst, 1997] W. N. Borst. Construction of Engineering Ontologies. PhD thesis, University of Twente, Enschede, 1997.

[Ding, Fensel, 2001] Ding, Y.; Fensel, D.: Ontology Library Systems. The key to successful Ontology Re- Use. In: Proceedings of the First Semantic Web Working Symposium. California, USA: Stanford University 2001; S. 93-112.

[Farquhar et al., 1997] A. Farquhar, R. Fikes, and J. Rice. The ontolingua server: a tool for collaborative ontology construction. International Journal of Human-Computer Studies, 46(6):707–728, June 1997.

[Gorcho, 2006] Corcho, O.: A Platform for the Development of Semantic Web Portals In Proceedings of the 6th international conference on Web engineering, Palo Alto, California: ACM International Conference Proceeding Series Pages 2006; P 145 - 152

[Gruber 1995] Gruber, T., "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", International Journal of Human-Computer Studies, Vol. 43 (1995), pp. 907-928

[Gruber, 1993] T. R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5:199–220, 1993.

[Guarino, 1995] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. International Journal of Human-Computer Studies, 43(5/6):625–640, 1995. Special issue on The Role of Formal Ontology in the Information Technology.

[Hartmann, Sure 2004] Hartmann J., Y. Sure, "An Infrastructure for Scalable, Reliable Semantic Portals" IEEE Intelligent Systems 19 (3): 58-65. May 2004

[Hollink et al., 2003] Hollink, L., Schreiber, G., Wielemaker J., and Wielinga. B. Semantic Annotation of Image Collections in Knowledge Capture - Knowledge Markup & Semantic Annotation Workshop (2003)

[Hyvönen et al., 2004] Hyvönen, E., Mäkelä, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S., MuseumFinland—Finnish museums on the semantic web in 3rd International Semantic Web Conference (ISWC2004, Hiroshima, Japan 07-11 November 2004)

[Jones et al.] Jones, D., Bench-Capon, T., Visser, P., "Methodologies for ontology development", (Visited 06-03-2007) http://www.iet.com/Projects/RKF/SME/methodologies-for-ontology-development.pdf

[Karvounarakis et al, 2000] Karvounarakis G, Christophides V, Plexousakis D, Alexaki S (2000) Querying community web portals. Technical report, Institute of Computer Science, FORTH, Heraklion, Greece.

[Lara 2004] Lara R., Sung-Kook Han, Holger Lausen, Michael Stollberg, Ying Ding, Dieter Fensel, " An Evaluation of Semantic Web Portals", IADIS Applied Computing International Conference 2004, Lisbon, Portugal, March 23-26, 2004

[Lausen et al., 2005] Lausen H., Ying Ding, Michael Stollberg, Dieter Fensel, Rubén Lara, and Sung-Kook Han, "Semantic web portals: state-of-the-art survey", Journal of Knowledge Management, 2005, Volume: 9 Issue: 5 Page: 40 – 49

[Noy, McGuiness] Noy, N, McGuiness, D, Ontology Development 101: A Guide to Creating Your First Ontology, http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

[Ontolingua project] (Visited 06-03-2007) http://www-ksl-svc.stanford.edu:5915/doc/frame-editor/what-is-an-ontology.html

[Staab et al., 2000] Staab S., J. Angele, Stefan Decker, Michael Erdmann, Andreas Hotho, Alexander Maedhe, Hans-Peter Schnurr, Rudi Studer, York Sure , „Semantic Community Web Portals", In: Computer Networks (Special Issue: WWW9 - Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, May, 15-19, 2000), Elsevier.

[van Hejlst et al., 1997] G. van Heijst, A. T. Schreiber, and B. J. Wielinga. Using explicit ontologies in KBS development. International Journal of Human-Computer Studies, 46(2/3):183–292, 1997.

## Author's Information

***Anna Devreni–Koutsouki** – PhD student, Sofia University, Bulgaria; e-mail: annadevreni@hotmail.com*

# TOWARDS CONTENT-SENSITIVE ACCESS TO THE ARTEFACTS OF THE BULGARIAN ICONOGRAPHY

**Desislava Paneva, Lilia Pavlova-Draganova, Lubomil Draganov**

***Abstract**:* This paper presents an ontological model of the knowledge about Bulgarian iconographical artefacts. It also describes content-sensitive services for access, browse, search and group iconographical objects, based on the presented ontology that will be implemented in the multimedia digital library "Virtual encyclopedia of Bulgarian iconography".

***Keywords**: Ontology, Bulgarian Iconography, Digital Libraries, Content-sensitive services.*

***ACM Classification Keywords:** I.2.4 Knowledge Representation Formalisms and Methods, H.3.7 Digital Libraries – Collection, Dissemination, System issues.*

## Introduction

East-Christian icon art is recognised as one of the most significant areas of the art of painting. Regrettably, it is neglected in the digital documentation and the registry of the art of painting. This tendency is suspended by the team from the Institute of Mathematics and Informatics (Bulgarian Academy of Sciences) with the development of the multimedia digital library "Virtual encyclopedia of the Bulgarian iconography" (http://mdl.cc.bas.bg/). This valuable galleria of knowledge and specimens of East-Christian culture and art is created during the project "Digital libraries with multimedia content and its applications in Bulgarian cultural heritage" [Pavlov et al., '06b]

[Pavlova-Draganova et al., '07] and includes several hundred specimens of Bulgarian icons and artefacts from different artists, historical periods, and schools [Pavlov&Paneva, '06] [Paneva et al., '05].

The impending development of the digital library points to the investigation and implementation of new techniques and methods for description of the semantics of iconographical artefacts and collections in order their valuable knowledge to be easy accessed and found. Knowledge technologies and Semantic web can provide these opportunities. Adopting them we make the first development steps of the "Bulgarian iconographical artefacts" ontology. This article presents the process of its consideration, scoping, and conceptualization, the ideas for its implementation in the "Virtual encyclopedia of the Bulgarian iconography". The interpretations of the iconographical knowledge do not have to be considered isolated from the standards and specifications in the field of cultural information representation. Therefore, section 2 summarizes the most important standard in the field of cultural heritage representation - CIDOC object-oriented Conceptual Reference Model (CIDOC CRM) and its use. Sections 3 deals with different aspects of the ontology development, based on CIDOC CRM concepts and properties. Section 4 discusses content-sensitive services for access, browse, search and group of iconographical objects, based on the presented ontology that are in a process of implementation in the multimedia digital library "Virtual encyclopedia of Bulgarian iconography".

## Ontological presentation of iconographical knowledge

One of the targets of the multimedia digital library "Virtual encyclopedia of the Bulgarian iconography" is to create rich context-based virtual presentation of the Bulgarian icon art and culture. Therefore, we observed and specified the experience that has been gained in the last 1000 years in the area of iconography to develop "a formal, explicit specification of a shared conceptualization" [Gruber, '93] about the iconography world - the ontology "Bulgarian iconographical artefacts". The annotator/indexers using this ontology will semantically describe and index the raw audiovisual iconographical content in order to create and maintain digital objects.

The interpretations of the iconographical knowledge do not have to be considered isolated from the standards and specifications in the field of cultural information representation because the goal is to maximize the reusability and portability of the designed ontological model. The most significant new development is the CIDOC Conceptual Reference Model, "object-oriented domain ontology" for expressing the implicit and explicit concepts in the documentation of cultural heritage. Since 9/12/2006 it is official standard ISO 21127:2006. It is the culmination of more than a decade of standards development work by the International Committee for Documentation of the International Council of Museums. Its role is to enable information exchange and integration between heterogeneous sources of cultural heritage information. CRM aims at providing the semantic definitions and clarifications needed to transform disparate, localised information sources into a coherent global resource. More specifically, it defines and is restricted to the underlying semantics of database schemata and document structures used in cultural heritage and museum documentation in terms of a formal ontology. It explains the logic of what they actually currently document, and thereby enables semantic interoperability. It intends to provide an optimal analysis of the intellectual structure of cultural documentation in logical terms.

The CRM is domain ontology in the sense used in knowledge technologies. It has been expressed as an object-oriented semantic model that can be readily converted to machine-readable formats such as RDF Schema, KIF, DAML + OIL, OWL, STEP, *etc*. It can also be implemented in any relational or object-oriented schema.

Real ontologies for concrete worlds of art objects are often developed as (conceptual at least) specializations of the CIDOC CRM ontology. During the creation of the "Bulgarian iconographical artefacts" ontology we observe the concepts and properties of CIDOC ontology and part of them we use in our ontology, other part we transform in order to fit for the iconography domain and several concepts don't belong to the CIDOC CRM ontology.

A juxtaposing example is shown in Table 1 for the concept 'Dimension' and its properties in the "Bulgarian iconographical artefacts" ontology and the respective chain of CIDOC CRM concepts and properties. The iconographical object can be adopted as a subset of the E22 Man-Made Object class. Ontology's concept 'dimension' is the same as CIDOC CRM E54 Dimension concept. The relationship between the iconographical object and its dimension is indicated by P43 has dimension property. In our ontology we adopted the 'Dimension→was observed in→Unit of Measurement' chain that is similar in CRM – 'E54→P40→E16'. In our ontology we split the 'E54→P90→E60' in three layers for the width, height and length of the iconographical objects [Pavlova-Draganova et al., '07].

| **Concepts and properties in "Bulgarian iconography artefacts" ontology** | **CIDOC CRM chains** |
|---|---|
| ***Dimension***<br>Iconographical object → has dimension → Dimension<br>Dimension → was observed in → Unit of Measurement<br>Dimension → has width value → Number<br>Dimension → has height value → Number<br>Dimension → has length value → Number | E70 Thing (E22 Man-Made Object) → P43 has dimension (is dimension of) → E54 Dimension<br>E54 Dimension → P40 was observed in → E16 Measurement (Unit of measurement of the dimension in our ontology)<br>E54 Dimension → P90 has value → E60 Number (value of the dimensions of our ontology) |

Table 1: The concept Dimension and its properties in the "Bulgarian iconographical artefacts" ontology and the respective chain of CIDOC CRM concepts and properties

## Description of the semantics of the Bulgarian iconographical artefacts

Bulgarian iconographical domain contains a rich knowledge base that has to be semantically described. For this aim we observe and specify the experience in the area of iconography and start the development of the ontology presenting iconographical knowledge and artefacts.

The first activity in the process of the development of "Bulgarian iconographical artefacts" ontology is the definition of the scope of the ontology. Scoping has been mainly based on several brainstorming sessions with artists and content providers. It depends on the future implementation of the ontology in the multimedia digital libraries "Virtual encyclopedia of the Bulgarian iconography". These brainstorming sessions allowed the production of most of the potentially relevant terms. At this stage, we also juxtaposed these concepts to the available concepts in CIDOC CRM, thus concealing significant ambiguities and differences of opinion. A clear issue that arose during these sessions was the difficulty in discovering of definite number of concepts and relations between these concepts. The concepts listed during the brainstorming sessions were grouped in areas of work corresponding naturally arising sub-groups. Most of the important concepts and many terms were identified. The main work of building the ontology was then to produce accurate definitions.

The iconographical object is related to three levels of knowledge, enriched with a set of sub-levels of the data classification. All these levels of knowledge or "thematic entities" in the ontology conception are supported by the scientific diagnosis results and the related documentation [Pavlova-Draganova et al., '07].

- The entity "Identification" consists of general historical data, identifying aspects such as title, type, author, clan, iconographic school, period, dimensions, current location, description of the iconographical object/collection,
- The entity "Description" consists of information concerning the descriptive details of the theme and forms of representation, providing a better understanding of the context, such as characters and scenes, participation of characters in scenes, *etc.*
- The entity "Technical" includes technical information both revealing the techniques and the base materials used in the creation of the iconographical object/collection, and also concerning examinations of the condition, such as diagnosis or conservation treatments history.

These main entities and their metadata are supported, documented and provided by the scientific diagnosis, which has been applied to the iconographical objects and collections.

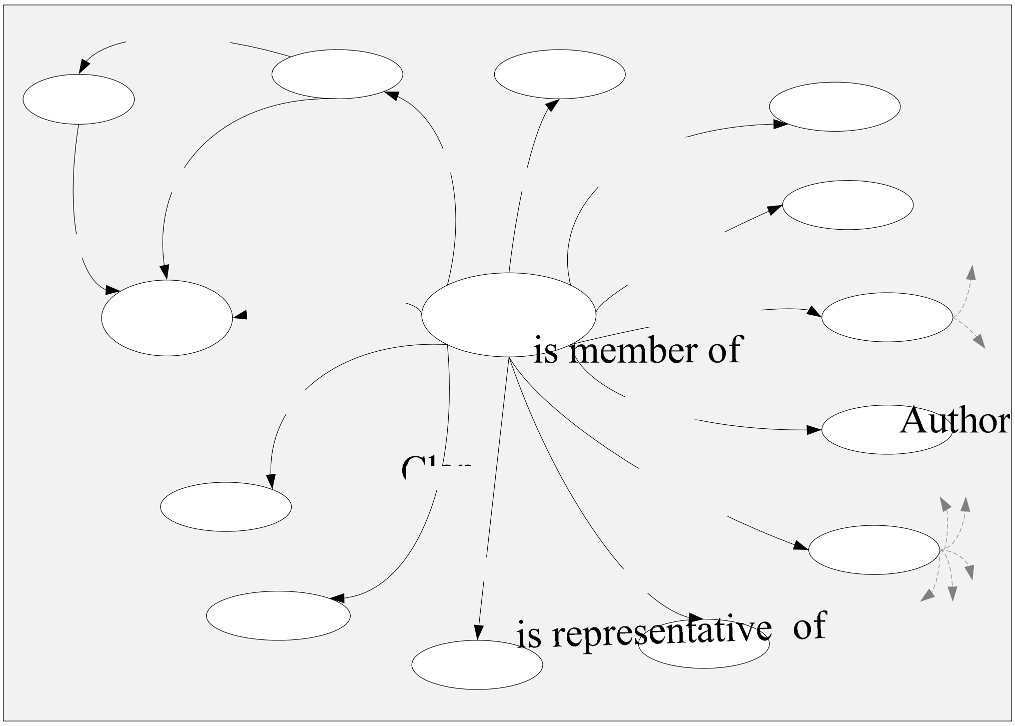


Figure 1: Main classes and relations related to the ‘Iconographical Object’ in the “Bulgarian iconographical artefacts” ontology

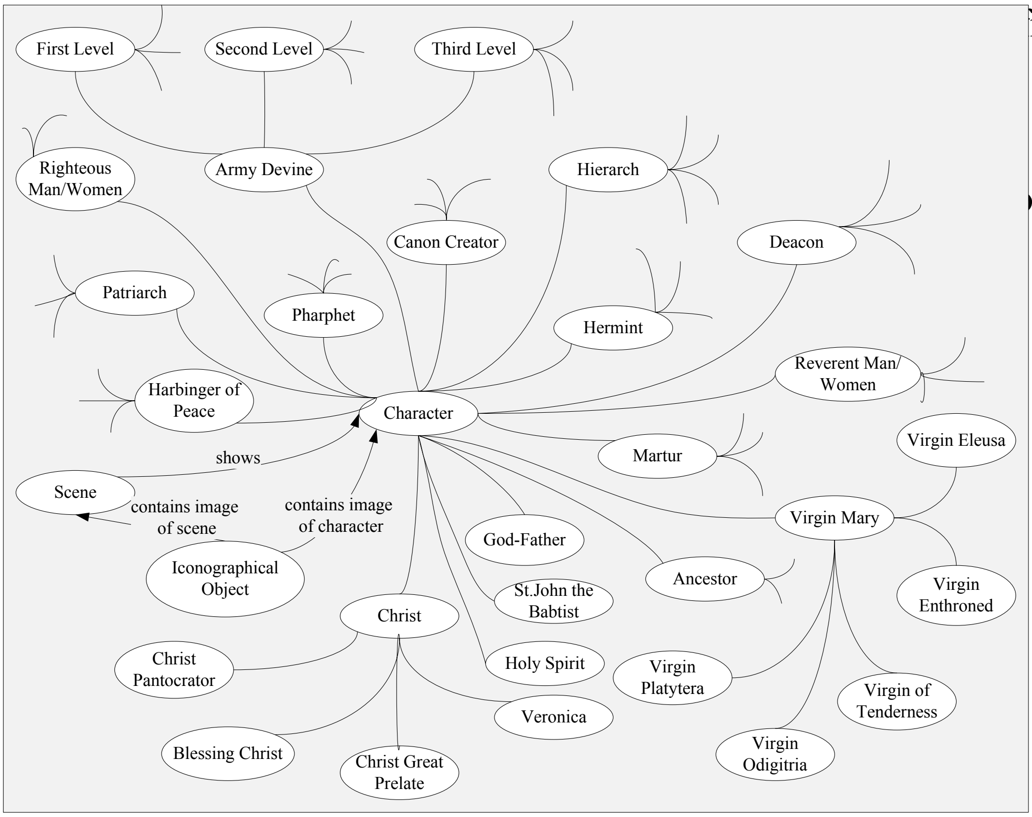


Figure 2: “Character” class and its subclasses in the “Bulgarian iconographical artefacts” ontology

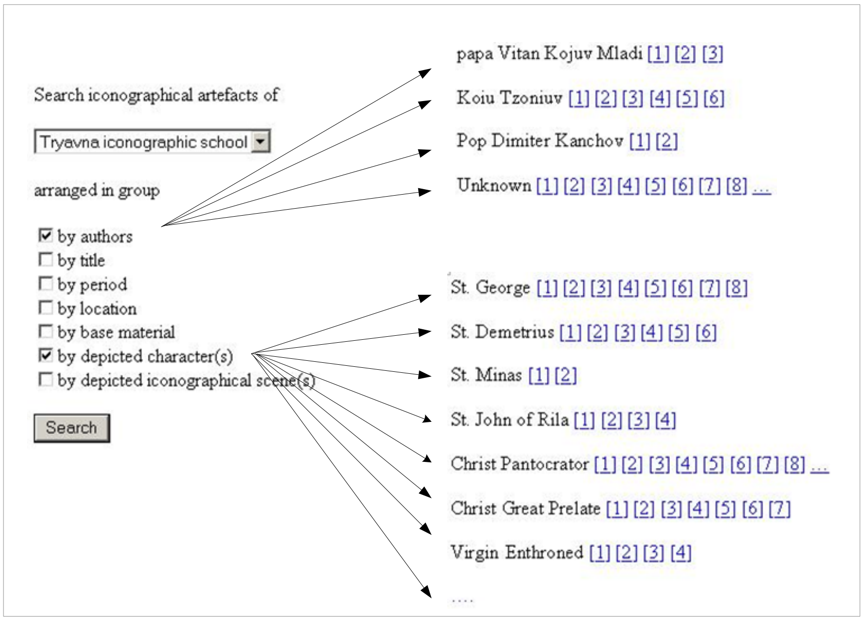
Figure 1 depicts the main classes and relations related to the concept ‘Iconographical Object’ in the ontology.

As it is shown on figure 1 in the “Bulgarian iconographical artefacts” ontology the concept ‘Iconographical Object’ is described with its title, author appellation, its clan and iconographic school, its current location and the period (time-span) of its creation, used base material and iconographic techniques, overall description. The ontology also captures the characters and scenes depicted on the iconographical object (icon, plastic iconographical object, mural painting, iconostasis, iconographic element in Psalm-book, *etc.*) in order to be defined its compoundness. Figure 2 depicts the main subclasses of the ‘Character” class.

## New content-sensitive and customizing services in Virtual encyclopedia of the Bulgarian iconography

Multimedia digital library “Virtual encyclopedia of the Bulgarian iconography” currently provides its users with several services for present and search iconographical artefacts. But, the development of the “Bulgarian iconographical artefacts” ontology allows the inclusion of new semantic-based and content-sensitive access services with customizing elements in it.

One of them is “semantic-based search with grouping” depicted on a figure 3. It provides searching for iconographical artefacts that are created by representatives from chosen iconographic school, for example “Tryavna iconographic school”. The results are lists of artefacts grouped according the several chosen criteria: authors, title, period, location, base material, depicted characters and scenes. This grouping opportunity will be very helpful for quick find of definite artefact in the iconographical object repository. During the search process the semantic-based service traces nodes of the ontological tree and presents instances of checked classes.



Search iconographical artefacts of

Tryavna iconographic school

arranged in group

☑ by authors
□ by title
□ by period
□ by location
□ by base material
☑ by depicted character(s)
□ by depicted iconographical scene(s)

papa Vitan Kojuv Mladi [1] [2] [3]
Koiu Tzoniuv [1] [2] [3] [4] [5] [6]
Pop Dimiter Kanchov [1] [2]
Unknown [1] [2] [3] [4] [5] [6] [7] [8] ...

St. George [1] [2] [3] [4] [5] [6] [7] [8]
St. Demetrius [1] [2] [3] [4] [5] [6]
St. Minas [1] [2]
St. John of Rila [1] [2] [3] [4]
Christ Pantocrator [1] [2] [3] [4] [5] [6] [7] [8] ...
Christ Great Prelate [1] [2] [3] [4] [5] [6] [7]
Virgin Enthroned [1] [2] [3] [4]
....

Figure 3: “Semantic-based search with grouping” in multimedia digital library “Virtual encyclopedia of the Bulgarian iconography”

Similar service could provide this grouping artefacts functionality during the multi-criteria search (at present available in the digital library), but the desired grouping criteria have to be selected by the user during its personal profile creation. This action dictates the proper iconographical object observation style [Paneva, ‘06].

Another content-sensitive service is the “content browsing”. It will display the ontology information graphically in order to support artists to easily navigate and browse through the concepts. Moreover, it provides them with information about the concepts and other related issues concerning the ontology. This service will be particularly useful to artists who are not familiar with concept searching and want to browse the information resources in a user-friendly way.

The displayed concepts will be obtained querying the ontology. If the artist requires the concrete content associated to any of the concepts displayed by the iconographical content browsing service, another query is done, this time, on the content database. In such way, the artist gets the information requested with the precise content to build his story.

## Conclusion

The “Bulgarian iconographical artefacts” ontology tries to capture the knowledge in the iconography domain in order to provide tool for semantically description and indexing of the raw audiovisual iconographical content digital objects in the multimedia digital library “Virtual encyclopedia of the Bulgarian iconography”. This ontology can be use for realization semantic-based access and search of concrete iconographical objects, as it shown in this paper. The future development of the digital library will continue to improve and extend the “Bulgarian iconographical artefacts” ontology and the DL services based on it.

## Bibliography

[Pavlov et al., ‘06b] Pavlov, R., Paneva D., Pavlova-Draganova L., Draganov L. (2006), Digital Libraries with Multimedia Content and Applications in Bulgarian Cultural Heritage (Analytical study), State Agency for Information Technologies and Communication (SAITC), by contract № 8/21.07.2005 between IMI-BAS and SAITC, 2006, Sofia, Bulgaria, Available at: http://mdl.cc.bas.bg/Digital_libraries_with_multimedia_content_and_applications_in_Bulgarian__cultural_heritage.pdf (accessed on April 10, 2007).

[Pavlova-Draganova et al., ‘07] Pavlova-Draganova L., Georgiev V., Draganov L. (2007), Virtual Encyclopeadia of Bulgarian Iconography, International Journal “Information Technologies and Knowledge”, vol.1, №3, pp. 267-271.

[Pavlov&Paneva, ‘06] Pavlov R., Paneva D. (2006), Toward Ubiquitous Learning Application of Digital Libraries with Multimedia Content, Cybernetics and Information Technologies, vol. 6, № 3, pp. 51-62.

[Paneva et al., ‘05] Paneva, D., Pavlova-Draganova L., Draganov L. (2005), Digital Libraries for Presentation and Preservation of East-Christian Heritage, Proceedings of the Second HUBUSKA Open Workshop “Generic Issues of Knowledge Technologies”, Budapest, Hungary, pp. 75-83.

[Paneva, ‘06] Paneva D. (2006), Use of Ontology-based Student Model in Semantic-oriented Access to the Knowledge in Digital Libraries, In proc. of HUBUSKA Fourth Open Workshop “Semantic Web and Knowledge Technologies Applications”, Varna, Bulgaria, pp. 31-41

[Pavlova-Draganova et al., ‘07] Pavlova-Draganova, L., D. Paneva, L. Draganov (2007), Knowledge Technologies for Description of the Semantics of the Bulgarian Iconographical Artefacts, In proc. of HUBUSKA Fifth Open Workshop “Knowledge Technologies and Applications”, Kosice, Slovakia, 31 May – 1 June, 2007

[Gruber, ‘93] Gruber T. (1993) Towards Principles of the Design of Ontologies Used for Knowledge Sharing. Int. Journal of Human Computer Studies 43, pp. 907-928, Available at: http://www.itee.uq.edu.au/~infs3101/_Readings/OntoEng.pdf (Accessed on April 10, 2007).

## Authors' Information

***Desislava Paneva** - Institute of Mathematics and Informatics, BAS, Bulgaria, 8, Acad. G. Bonchev str.; e-mail: dessi@cc.bas.bg*

***Lilia Pavlova-Draganova** – Laboratory of Telematics, BAS, Sofia, Bulgaria, 8, Acad. G. Bonchev str.; e-mail: lilia@cc.bas.bg*

***Lubomil Draganov** – Institute of Mathematics and Informatics, BAS, Sofia, Bulgaria, 8, Acad. G. Bonchev str.; e-mail: lubo@cc.bas.bg*

# USING WORDNET FOR BUILDING AN INTERLINGUAL DICTIONARY[1]

Juan Bekios, Igor Boguslavsky, Jesús Cardeñosa, Carolina Gallardo

***Abstract**: UNL is an enterprise to support multilinguality in Internet based on a common language called Universal Networking Language. One of the components of the language is a dictionary of Universal Words (UWs). Such dictionary constitutes the link between the vocabularies of the languages involved in the project. This article describes the process of creating the common UWs dictionary within the UNL context, using as an external resource Wordnet. The process is completely automatic. Implementation details and results of the process are shown.*

***Keywords**: Lexical Resources, Wordnet.*

***ACM Classification Keywords**: J.5. Arts and Humanities; H.2.8 Database Applications;*

## Introduction

The UNL Program was initiated by the Institute of Advanced Studies of the United Nations University under the UN auspices with an ambitious goal: to break down or at least to drastically lower the language barrier for the Internet users. It was launched in November 1996; the project embraced 14 groups from different countries representing a wide range of languages: Arabic, Chinese, German, French, Japanese, Hindi, Indonesian, Italian, Mongolian, Portuguese, Russian, Spanish and Thai.

The UNL Program pivots on the *Universal Networking Language*, a meaning representation language designed to represent informational content conveyed by natural languages. The complete specifications of the language are public and freely downloadable from Internet (see [Uchida et al., 2005]). One of the major applications of UNL is to serve as an interlingua between different natural languages. Besides that, UNL can also be used for other applications such as information retrieval, text summarization and the like. In fact, the specifications have known several versions, from version 1.0 in 1997 to current version of 2005, due to the fact that the language accommodates itself to new uses.

The UNL is composed of three main elements: **universal words** (UWs hereafter), **relations** and **attributes**. UWs form the vocabulary of the interlingua; relations express thematic roles and attributes represent the context and speaker dependent information. Formally, a UNL expression can be viewed as a semantic net, whose nodes are UWs, linked by arcs labeled with UNL relations. Universal Words are expanded by the attributes.

The complete set of UWs composes the **UNL dictionary**. The UNL dictionary is complemented with bilingual dictionaries, connecting UWs with words of different natural languages. Local dictionaries are formed by pairs of the form **<Word, UW>** where Word is any word of a given natural language and UW is the corresponding representation of one of its senses in UNL. The UNL dictionary constitutes a common lexical resource for all natural languages currently represented in the project, so that word senses of different natural languages become linked via their common UWs. Therefore, the UNL Dictionary can serve as an important lexical resource to construct multilingual dictionaries or other resources like thesauri, being UWs the pivot among the vocabulary of natural languages.

However, there is an apparent drawback in the UNL dictionary. The set of UWs is not formally defined, that is, the specifications do not provide either a complete knowledge base or precise instructions to create UWs. The absence of formalization of the lexical part of the language prevents the construction of a common dictionary of UWs for all the members of the project, the management of dictionaries and lexical resources being the hardest part of the project.

[1] This paper has been sponsored by the Ministry of Education and Science of Spain under project **CICYT HUM2005-07260**.

This paper presents a methodology and an application that tries to solve the main problems in the UNL dictionary management, namely, the standardization of the definition of UWs and the automatic construction of the common UNL dictionary on the basis of the existing lexical resources.

## Data Analysis

As already said, UWs constitute the vocabulary of the language. Broadly speaking, a UW is an English word modified by a series of semantic restrictions. The main purpose of semantic restrictions is to eliminate lexical ambiguity present in natural languages. Besides that, they establish major lexical relations with other words and specify an argument frame. In this way, UNL gets an expressive richness from the natural languages but without their ambiguity. For example, the verb "land" in English has several senses and different argument frames. In a sentence like "*The plane landed at the Geneva airport*", the corresponding UW for the sense of this verb would be **land(icl>do, plt>surface, agt>thing, plc>thing).** This UW is divided in two parts: the headword and the list of semantic restrictions enclosed in parenthesis and separated by commas, as shown in figure 1.

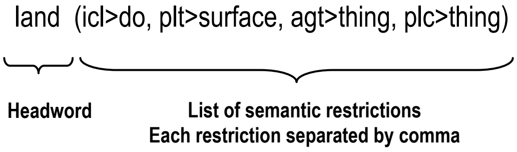

land (icl>do, plt>surface, agt>thing, plc>thing)
Headword
List of semantic restrictions
Each restriction separated by comma

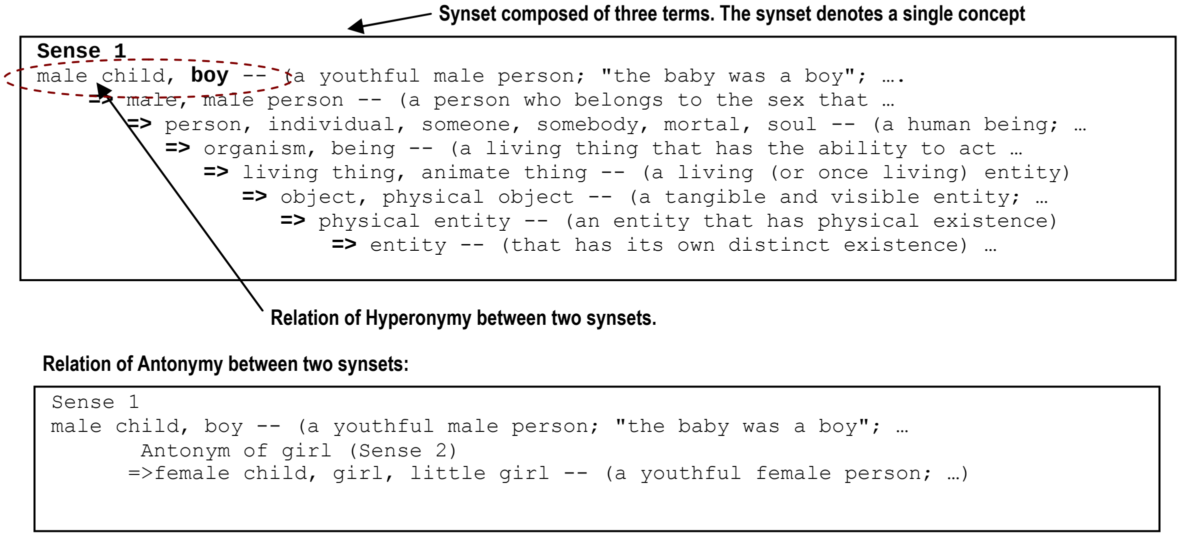
**Figure 1**. Parts of a UW

This UW corresponds to the definition "To alight upon or strike a surface". The proposed semantic restrictions stand for:

- **icl>do**: (where *icl* stands for *included*) establishes the type of action that "lands" belongs to, that is, actions initiated by an agent.
- **plt>surface**: (where *plt* stands for *place to*) expresses an inherent part of the verb meaning, namely that the final direction of the motion expressed by "land" is onto a surface.
- **agt>thing, plc>thing**: (where *plc* stands for *place*) establish the obligatory semantic arguments of the predicate "land".

As can be seen from this example, UNL semantic restrictions are based on lexical relations among terms, namely, the *hyponymy* relation (by means of "icl" relation), *synonymy* ("equ" relation) and *meronymy* ("pof" relation). Besides, the semantic arguments of predicates (that is, verbs, adjective and some nouns) must be specified. Since UWs are described by means of relations between terms, the result is a connected net of UWs, constituting the UW system. A more comprehensive view of the UW system is described in [Boguslavsky et al, 2005].

The organizing principles of the UW system are based on well-known lexical relations, like those present in Wordnet [Fellbaum, 1998]. Wordnet is a large lexical database of English, freely downloadable from Internet (http://wordnet.princeton.edu/). As opposed to most lexicographic works and similarly to the UW system, Wordnet is not ordered alphabetically but conceptually, by means of semantic relations. The main organizing relation in Wordnet is the **synset**, defined as a group of cognitive synonyms that expresses a single **concept.** Besides, synsets are interconnected by means of other lexico-semantic relations like hyperonymy (hierarchical relation between class and subclass), antonymy (an opposite term), metonymy (part-of) and other relations like *relative_to*, sentence frames for verbs, etc. Figure 2 shows two samples of Wordnet that illustrate the relations of hyperonymy and antonymy for the synset "male child, boy".



Synset composed of three terms. The synset denotes a single concept

```
Sense 1
male child, boy -- (a youthful male person; "the baby was a boy"; ….
    => male, male person -- (a person who belongs to the sex that …
       => person, individual, someone, somebody, mortal, soul -- (a human being; …
          => organism, being -- (a living thing that has the ability to act …
             => living thing, animate thing -- (a living (or once living) entity)
                => object, physical object -- (a tangible and visible entity; …
                   => physical entity -- (an entity that has physical existence)
                       => entity -- (that has its own distinct existence) …
```

Relation of Hyperonymy between two synsets.

Relation of Antonymy between two synsets:

```
Sense 1
male child, boy -- (a youthful male person; "the baby was a boy"; …
       Antonym of girl (Sense 2)
      =>female child, girl, little girl -- (a youthful female person; …)
```

**Fig 2**. Two samples of Wordnet

Wordnet includes nouns, adjectives, adverbs and verbs. Other categories like prepositions, determiners or conjunctions are spelled out from Wordnet, since they do not denote any semantic concept.

The use Wordnet as an ancillary resource to support the process of automatic dictionaries creation is not new in the UNL framework. The generation of UNL-English dictionaries for specific texts is depicted in [Bhattacharyya et al, 2004]. We have made use of the similarity of Wordnet and the UW system to use Wordnet as the main source to define and create a complete UW dictionary. The complete process and the final UW dictionary are described in the next sections.

## Design Issues

The main design issue when considering a UW Dictionary and Wordnet as the main source of data is that the structure of lexical relations in Wordnet can be used to construct the list of restrictions of UWs. To do that, we must first establish the main similarities between Wordnet and the UW system. Such similarities are exposed in table 1, where the first column describes lexical relations in Wordnet and the second column states their equivalent semantic restrictions in the UW system.

Table 1 shows how any **word** included in Wordnet can be used to represent the **headword** of a UW. Each different sense of an English word is delimited by means the set of synonyms, hypernyms, antonyms and other lexical relations associated to that word, in the same way that the sense of a headword in UNL is delimited by its list of semantic restrictions.

| ***SIMILARITY RELATIONS*** | |
|---|---|
| ***WordNet 2.1*** | ***UW System*** |
| ***An English Word.*** | ***Headword*** |
| ***Synset*** | ***Relation equ>*** |
| ***Hyperonym*** | ***Relation icl>*** |
| ***Antonym*** | ***Relation ant>*** |
| ***Relative to*** | ***Relation com>*** |

**Table 1**. Similarity Relations between Wordnet and the UW system

What is really important for us is that from these similarity relations, it is possible to devise a **method** that defines UWs in a **systematic way** using Wordnet. The method is described in figure 3.

```
1. Extract a Word from Wordnet
2. Obtain each of the senses of the Word
3. For each sense of the word, do the following:
    3.1. Assign the Word to the Headword of UW
    3.2. Depending on the syntactic category (noun, adjective, adverb, verb)
         and on the data obtained from WordNet; for each sense, apply a set
         of rules that will generate semantic restrictions.
    3.3. Taking the Headword and the obtained restrictions, construct the
         complete Universal Word.
    3.4. Store the UW in the dictionary.
4. If more UWs are to be constructed, return to step 1. Otherwise, finish.
```

**Fig. 3.** Method to define UWs from Wordnet

There are two aspects that require further explanations in this method. First, the number of UWs that are created per word and second the set of rules mentioned in step 3.2 of the method.

The method will generate one UW per word sense. For example, the word “bank” as a noun has 10 senses and thus generates 10 different UWs. In some cases, when the difference between the senses is too subtle, Wordnet relations are not sufficient to differentiate between them. In these cases, the method will generate identical UWs for different senses. These “duplicate” UWs must be treated in a special way.

On the other hand, the method is based on the similarity relations of table 1 along with **a set of rules** to systematically yield a dictionary of UWs. These rules are presented in the next section.

## Set of Rules

Only six rules are required to create the semantic restrictions of UWs. A rule takes as input a Wordnet word (that is, the set of senses for the word and the lexical relations each word is engaged in) and yields a semantic restriction suitable for the UW that is being created. The six rules are:

### 1. Rule for the Construction of Headword (HW)

***Definition:*** This rule turns a WordNet word into a Headword for a candidate Universal Word.

***Example:*** The word “*banking company*” in Wordnet returns the Headword “***banking_company***”.

### 2. Immediate Hypernym Rule (RHper)

***Definition:*** For a sense of a word, take its most immediate hypernym and establish an **icl>** relation type.

***Example:*** For the first sense of the word “*bank*” as a noun, take its immediate hypernym (“*financial institution)* and create a semantic restriction with icl>. The result is: “**icl> financial_institution**”

### 3. Immediate Hyponym rule (RHpo)

***Definition:*** For a sense of a word, take its most immediate hyponym and establish an **icl<** relation type. Use this relation only when there are duplicate UWs.

***Example:*** For the first sense of the word “*bank*” as a noun, it is possible to obtain navigating through WordNet an immediate hyponym (“for example *credit_union”)* and create a semantic restriction with icl<. The result is: “**icl<credit_union**”

### 4. Rule of First Synonym (RSyn)

***Definition:*** For a sense of a word, if the word is not the first element of the synset, take the first word of the synset and establish an **equ>** relation.

***Example:*** For the first sense of the word "*bank*", it synset is {*depository financial institution*, ***bank***, *banking concern*, *banking company*}. Since "bank" is not the first element, create the following semantic restriction: "**equ> depository_financial_institution**"

### 5. Rule of First Antonym (RAnt)

***Definition***: For a sense of a word, take its associated antonym (if any) and establish an "**ant>**" relation.

***Example:*** For the adjective "*good*" in its first sense, the antonym associated to its first sense is "*bad*", therefore the generated restriction will be: "**ant>bad**"

### 6. Rule of Relative_to (RRel)

***Definition:*** For a sense of a word (usually adjectives), take the associated noun by means of relation "pertains to" (if any) and establish an "**com>**" relation

***Example:*** For "the *legal*" adjective, WordNet establishes a relation belongs to the noun "*law*", therefore the following restriction is obtained: "**com>law**"

These rules are independent of each other and can be executed in any order. When constructing the complete UWs dictionary, the application of rules will depend on the syntactic category of the headword (that is, not all rules are relevant for a given syntactic category). For example, when working with verbs, the application of the Antonym rule is irrelevant, since the meaning of a verb is not characterized by its antonyms. Table 2 summarizes the rules that are triggered for each syntactic category.

| Syntactic category | Executed rules |
|---|---|
| *Noun* | *HW, RHper, RSyn, RAnt* |
| *Adjective* | *HW, RHper, RAnt, RSyn, RRel,* |
| *Adverb* | *HW, RSyn, RAnt, RRel* |

**Table 2**. Set of rules relevant for each syntactic category

That is, a given noun may produce at most 4 semantic restrictions. For example, the noun "boy" in its first sense produces the following semantic restrictions:

- icl>male>thing (by means of RHper)
- equ>male_child (by means of RS)
- ant>girl (by means of RA)

The final UW is the concatenation of the generated semantic restrictions following the same order of table 2:

**boy(icl>male>thing, equ>male_child, ant>girl)**

The order of semantic restrictions implicit in table 2 is a convention followed by all the team members of the project. A different ordering will not imply different semantics of the UW.

Verbs are treated in a different way. Whereas all the information required for creating good UWs for nouns, adverbs and adjectives is present in the Wordnet, the mapping between verbal UWs and verbs in Wordnet is not so straightforward. This is due to the following reasons:

- Verbal UWs are categorized into three basic types of events: "do", "occur" and "be". This categorization is absent in Wordnet.
- Verbal UWs should be provided with its semantic arguments. Verbs in Wordnet are assigned a Sentence Frame, which is a, often incomplete, description of syntactic arguments for verbs.

Since there is no one-to-one relation between verbal UWs and the verbs, it was necessary to infer the type of event and the semantic arguments from the scarce information present in Wordnet. For that, we made use of the so-called lexicographic files which define broad ontological categories. Some of these categories are "verbs of dressing and bodily care", "cognition verbs", "verbs of being and having". The combination of the ontological category together with the sentence frame of a verb gives us a hint about its type of event and semantic arguments. Table 3 shows an excerpt of the combinations that have been used to define verbal UWs.

| ***Wordnet*** | | ***UNL*** | | |
|---|---|---|---|---|
| ***Ontological Category*** | ***Sentence Frame*** | ***Event Type*** | ***Semantic Arguments*** | ***Example*** |
| *verbs of being, having, spatial relations* | *Somebody ----s to somebody* | *be* | *aoj>thing,obj>thing* | *conform(icl>be,aoj>thing,obj>thing)* |
| *verbs of weather* | *Somebody ----s* | *occur* | *obj>thing* | *steam(icl>occur,obj>thing)* |
| *verbs of creation* | *Somebody ----s something* | *do* | *agt>thing,obj>thing* | *cut(icl>do,agt>thing,obj>thing)* |

**Table 3**. Combinations to define verbal UWs.

## The Dictionary Application

The complete application is composed of the following modules, graphically shown in figure 4:

- **Conversion Module:** This component converts words from Wordnet into UWs. This module uses the **Rules** and the Wordnet data. The generated Universal Words are served to the Database Manager.
- **Database Manager:** This component manages all the communications to and from the Database. Thus, this module receives the set of generated UWs from the Conversion Module and serves them to the Database. On the other hand, this module manages the processes of searching, modifying, deleting and inserting UWs as requested by users through the **Web Browser**. This component was developed in Java, using the special library Hibernate. (www.hibernate.org). In the near future, the UW Dictionary is expected to store the translations of UWs not only into English but into the other languages of the project.
- **Web Browser:** It refers to any existent web browser like Explorer, Firefox, Opera, etc. which will be used by users in order to interact with the UW Dictionary.

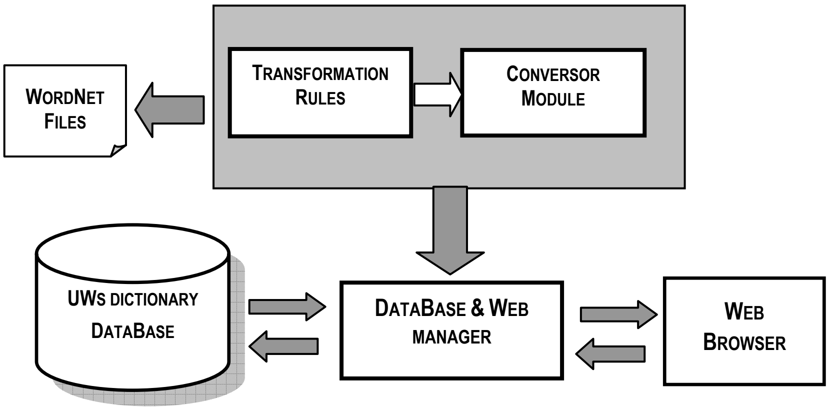


**Fig 4**. Components and relations of dependency of the Dictionary of Universal Words

The application can be accessed at the following address: http://chueca.dia.fi.upm.es:8080/UNLDicWeb/

## Results

All the UWs of the resulting UW Dictionary have been created automatically, without human intervention. Obtained results for a total amount of 207016 words that have been processed are summarized in table 4, where the total amount of generated UWs divided in syntactic categories is shown. The percentage of duplicate UWs for each syntactic category is also specified.

| | ***Nouns*** | ***Adjectives*** | ***Adverbs*** | ***Verbs*** |
|---|---|---|---|---|
| ***Unique UWs*** | *142343* | *26784* | *4958* | *23716* |
| ***Duplicate UWs*** | *2761* | *4518* | *762* | *1174* |
| ***Total*** | ***145104*** | ***31382*** | ***5728*** | ***24890*** |
| ***% duplicate UWs*** | ***1,9%*** | ***14,39%*** | ***13%*** | ***4,7%*** |

***Table 4**. Obtained results*

Since nouns are by far the most elaborated category in Wordnet, we considered as correct UWs the set of unique UWs, and as incorrect the set of duplicate UWs. As can be seen from table 4, the rate of duplicate UWs for nouns is less than 2%, a good result for the most polysemous syntactic category. Surprisingly, the results for verbs is rather good (less that 5% of error rate), although we assume that semantic arguments of verbs require human revision. On the other hand, both adjectives and adverbs yield an error rate quite high (around 14%). The possible reason for such an error rate may lie in the fact that the main lexical relations present in Wordnet are synonymy and hypernym, natural relations for nouns but not for predicates like adjectives or adverbs.

## Bibliography

[Bhattacharyya et al, 2004]. N. Verma and P. Bhattacharyya, *Automatic Lexicon Generation through WordNet,* Global WordNet Conference (GWC-2004), Czech Republic. Jan, 2004

[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

***Igor Boguslavsky** – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: igor@opera.dia.fi.upm.es http://www.vai.dia.fi.upm.es*

***Juan Bekios** – Group of Validation and Industrial Applications. Facultad de Informática. Universidad Politécnica de Madrid; Madrid 28660, Spain; e-mail: juan.bekios@opera.dia.fi.upm.es. http://www.vai.dia.fi.upm.es*

***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*

# WEB INTERFACES DESTINED FOR PEOPLE WITH DISABILITIES

Laura Ciocoiu, Ionuț Petre, Dragoş Smada, Dragoş Nicolau

***Abstract****: One of the main characteristics of the world that we live in is the access to information and one of the main ways to reach the information is the Internet. Most Internet sites put accessibility problem on a secondary plan. If we try to define this concept (accessibility) we could say that accessibility it's a way to offer access to information for the people with disabilities.*

*For example blind people can't navigate on the Internet like usual people. For that reason Internet sites have to put at their disposal ways to make their content known to this people. Accessibility does not refer only at blind people the web accessibility refers to all people who lost their ability to access the Internet sites.*

*The web accessibility includes every disability that stops people with disabilities to access the web sites content like hearing disability, neurological and cognitive. People that have low speed Internet connection or with low performance computers can use the web accessibility.*

***Keywords****: Accessibility, Adaptability, Adaptable interfaces, Interactive database*

***ACM Classification****: H 5.2. Information interfaces and presentation (e.g., HCI): Miscellaneous.*

## Introduction

"A society that eliminates some of his members it's a poor society"

In the entire world exists over 500 millions people with disabilities that are in-title to have the same rights like normal people. The first article of the Universal Declaration of the Human Rights says that all human beings are free and equal in dignity and rights.

The web accessibility is a different procedure to structure and organize a web page after strict accessibility criteria for people with disabilities. And like this people with disabilities can easily navigate on web and understand their content.

The Internet represents a important information source in most of our life aspects like commerce, health, education and job finding. From this point of view web accessibility helps disability people to integrate more easily in to the society.

Expediently – induced by user satisfaction and navigation facility and retain the structure of the site. Expediently is a quality: you realise her absence only when is not available anymore.

Accessibility and Expediently they are in strong relationship because both improve the efficiency and the satisfaction of the viewer. While the accessibility tends to make a product more accessible to the visitor, Expediently targets to please only a category of visitors that use that service or product.

The transformation of a web site that didn't respect the aspects related with accessibility into a site that respects those aspects could be an easy task or a hard one depending of the complexity of the site.

The interactive data base- represents a model of the real world and can only represent a limited number of characteristics necessary in different applications. No matter how perfect this application is, there are some applications that we can conceive that could not be satisfied by the database.

For building a database corresponding with a real system we must make a general appreciation of the system. This appreciation contains information regarding the structure of the system and essential system elements that are contained in a sketch.

For the relation model of the data base choosing the relations that are contained in the data base it's very important. The information contained in the database cannot be randomly chosen from the domain associated with their attributes. This kind of errors can be detected imposing some restrictions over the data. There are two kinds of restrictions:

- that depend on semantics of elements domain's
- produces by comparative values.

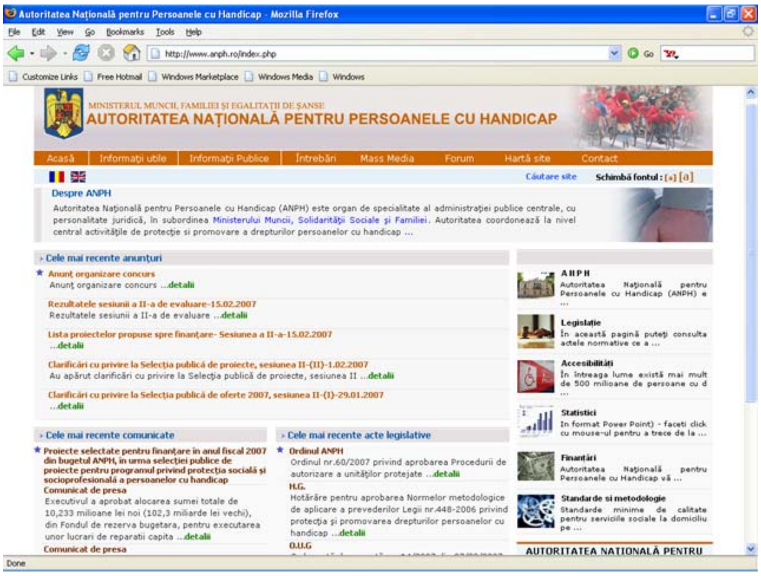
## The realisation of the web interface

### Starting page

An web site well made is loading very fast, and offers to the visitors a very well functionality, a complete content, the architecture of the information what is simple and clear and assures a intuitive navigation, quick access to the information that you were looking for. A quality design means a pleasant look that shows the site functionality. Will use for exemplification the site of the National Authority for People with Disabilities site made by this team, www.anph.ro

The first page must contain links to the main sections of the site; those are most interesting for the user and permit the user to detail in other pages the wanted sections.

That's way is preferred that the structure of the database based on hierarchical decomposition. For example divisions like Categories –under category –Details were the category represents the main entity and contains the under categories and the detail represents the lowest entity and is a part of under category.


Autoritatea Națională pentru Persoanele cu Handicap - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.anph.ro/index.php
Go
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
MINISTERUL MUNCII, FAMILIEI ȘI EGALITAȚII DE ȘANSE
AUTORITATEA NAȚIONALĂ PENTRU PERSOANELE CU HANDICAP
Acasă
Informații utile
Informații Publice
Întrebări
Mass Media
Forum
Hartă site
Contact
Căutare site
Schimbă fontul : [a] [a]
Done

Despre ANPH

Autoritatea Națională pentru Persoanele cu Handicap (ANPH) este organ de specialitate al administrației publice centrale, cu personalitate juridică, în subordinea Ministerului Muncii, Solidarității Sociale și Familiei. Autoritatea coordonează la nivel central activitățile de protecție si promovare a drepturilor persoanelor cu handicap ...

› Cele mai recente anunțuri

★ Anunț organizare concurs
Anunț organizare concurs ...detalii

Rezultatele sesiunii a II-a de evaluare-15.02.2007
Rezultatele sesiunii a II-a de evaluare ...detalii

Lista proiectelor propuse spre finanțare- Sesiunea a II-a-15.02.2007
...detalii

Clarificări cu privire la Selecția publică de proiecte, sesiunea II-(II)-1.02.2007
Au apărut clarificări cu privire la Selecția publică de proiecte, sesiunea II ...detalii

Clarificări cu privire la Selecția publică de oferte 2007, sesiunea II-(I)-29.01.2007
...detalii

› Cele mai recente comunicate

★ Proiecte selectate pentru finanțare în anul fiscal 2007 din bugetul ANPH, în urma selecției publice de proiecte pentru programul privind protecția socială și socioprofesională a persoanelor cu handicap

Comunicat de presa
Executivul a aprobat alocarea sumei totale de 10,233 milioane lei noi (102,3 miliarde lei vechi), din Fondul de rezerva bugetara, pentru executarea unor lucrari de reparatii capita ...detalii

Comunicat de presa

› Cele mai recente acte legislative

★ Ordinul ANPH
Ordinul nr.60/2007 privind aprobarea Procedurii de autorizare a unităților protejate ...detalii

H.G.
Hotărâre pentru aprobarea Normelor metodologice de aplicare a prevederilor Legii nr.448-2006 privind protecția și promovarea drepturilor persoanelor cu handicap ...detalii

O.U.G

ANPH
Autoritatea Națională pentru Persoanele cu Handicap (ANPH) e ...

Legislație
În această pagină puteți consulta actele normative ce a ...

Accesibilități
În intreaga lume există mai mult de 500 milioane de persoane cu d ...

Statistici
In format Power Point) - faceti click cu mouse-ul pentru a trece de la ...

Finanțări
Autoritatea Națională pentru Persoanele cu Handicap vă ...

Standarde si metodologie
Standarde minime de calitate pentru serviciile sociale la domiciliu pe ...

AUTORITATEA NATIONALĂ PENTRU

Figure 1. Starting Page

The first page must contain links to the main sections of the site; those are most interesting for the user and permit the user to detail in other pages the wanted sections as we can see in figure 1.

On the first page we have to find most recent information, for example last minute regulations announcements, press communicates and anything that is new for the site visitor.

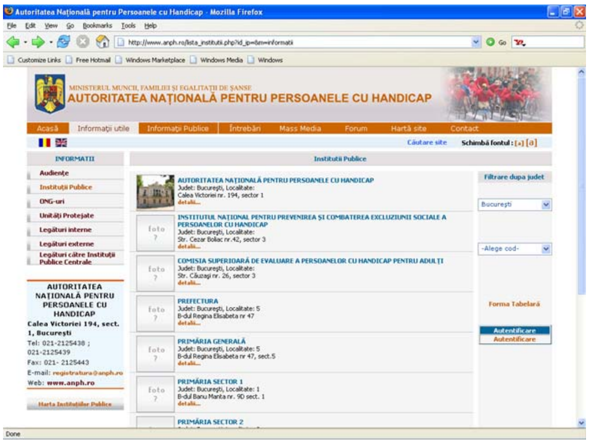
### Site menus

In this case in the horizontal menu we find the links that goes to the main sections that very important for the user, Useful Information, Public Information, Frequent Questions, Forum, Site Map. In the vertical menu are sections that are not very important and they are addressing to a restrained area of users and contains the to the Legislation, Statistics, Financings, Standards and Methodology, Accessibility and General information about National Authority for People with Disabilities.

## Sections

By accessing from the menu of useful information the user can find information about the units and institutions all over the country that the main occupation to treat people with disabilities, and they access information for every unit as we can see in figure 2.a.

Beside all this lists and units details, the visitor can find out the schedule for audiences contact addresses for the personal of this units and other links for same category sites as we can see in the figure 2.b.



AUTORITATEA NAŢIONALĂ PENTRU PERSOANELE CU HANDICAP
Judet: Bucureşti, Localitate:
Calea Victoriei nr. 194, sector 1
detalii...

INSTITUTUL NAŢIONAL PENTRU PREVENIREA ŞI COMBATEREA EXCLUZIUNII SOCIALE A PERSOANELOR CU HANDICAP
Judet: Bucureşti, Localitate:
Str. Cezar Boliac nr.42, sector 3
detalii...

COMISIA SUPERIOARĂ DE EVALUARE A PERSOANELOR CU HANDICAP PENTRU ADULŢI
Judet: Bucureşti, Localitate:
Str. Căuzaşi nr. 26, sector 3
detalii...

PREFECTURA
Judet: Bucureşti, Localitate: 5
B-dul Regina Elisabeta nr 47
detalii...

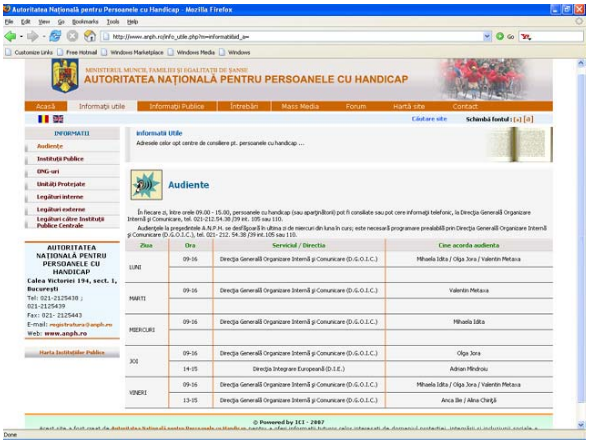
PRIMĂRIA GENERALĂ
Judet: Bucureşti, Localitate: 5
B-dul Regina Elisabeta nr 47, sect.5
detalii...

PRIMĂRIA SECTOR 1
Judet: Bucureşti, Localitate: 1
B-dul Banu Manta nr. 9D sect. 1
detalii...

PRIMĂRIA SECTOR 2

AUTORITATEA NAŢIONALĂ PENTRU PERSOANELE CU HANDICAP
Calea Victoriei 194, sect. 1, Bucureşti
Tel: 021-2125438 ;
021-2125439
Fax: 021- 2125443
E-mail: registratura@anph.ro
Web: www.anph.ro

Figure 2.a. Useful Information


Autoritatea Naţională pentru Persoanele cu Handicap - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.anph.ro/info_utile.php?m=informatii&id_a=
Go
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
MINISTERUL MUNCII, FAMILIEI ŞI EGALITĂŢII DE ŞANSE
AUTORITATEA NAŢIONALĂ PENTRU PERSOANELE CU HANDICAP
Acasă
Informaţii utile
Informaţii Publice
Întrebări
Mass Media
Forum
Hartă site
Contact
Căutare site
Schimbă fontul : [a] [a]
INFORMATII
Audienţe
Instituţii Publice
ONG-uri
Unităţi Protejate
Legături interne
Legături externe
Legături către Instituţii Publice Centrale
Harta Instituţiilor Publice
Done

informatii Utile
Adresele celor opt centre de consiliere pt. persoanele cu handicap ...

Audiente

În fiecare zi, între orele 09.00 - 15.00, persoanele cu handicap (sau aparţinătorii) pot fi consiliate sau pot cere informaţii telefonic, la Direcţia Generală Organizare Internă şi Comunicare, tel. 021-212.54.38 /39 int. 105 sau 110.

Audienţele la preşedintele A.N.P.H. se desfăşoară în ultima zi de miercuri din luna în curs; este necesară programare prealabilă prin Direcţia Generală Organizare Internă şi Comunicare (D.G.O.I.C.), tel. 021- 212. 54.38 /39 int.105 sau 110.

| Ziua | Ora | Serviciul / Directia | Cine acorda audienta |
|---|---|---|---|
| LUNI | 09-16 | Direcţia Generală Organizare Internă şi Comunicare (D.G.O.I.C.) | Mihaela Idita / Olga Jora / Valentin Metaxa |
| | | | |
| MARTI | 09-16 | Direcţia Generală Organizare Internă şi Comunicare (D.G.O.I.C.) | Valentin Metaxa |
| | | | |
| MIERCURI | 09-16 | Direcţia Generală Organizare Internă şi Comunicare (D.G.O.I.C.) | Mihaela Idita |
| | | | |
| JOI | 09-16 | Direcţia Generală Organizare Internă şi Comunicare (D.G.O.I.C.) | Olga Jora |
| | 14-15 | Direcţia Integrare Europeană (D.I.E.) | Adrian Mîndroiu |
| VINERI | 09-16 | Direcţia Generală Organizare Internă şi Comunicare (D.G.O.I.C.) | Mihaela Idita / Olga Jora / Valentin Metaxa |
| | 13-15 | Direcţia Generală Organizare Internă şi Comunicare (D.G.O.I.C.) | Anca Ilie / Alina Chiriţă |

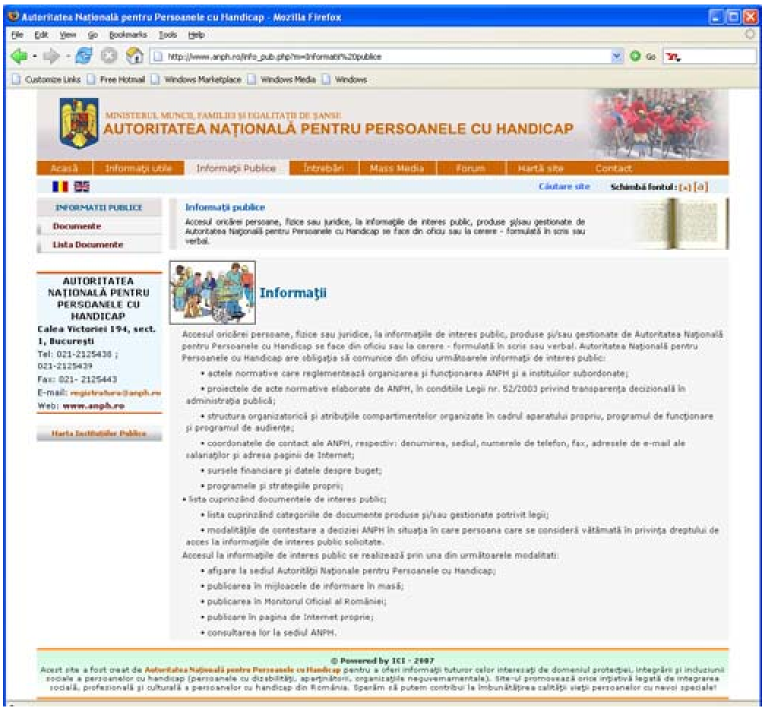
AUTORITATEA NAŢIONALĂ PENTRU PERSOANELE CU HANDICAP
Calea Victoriei 194, sect. 1, Bucureşti
Tel: 021-2125438 ;
021-2125439
Fax: 021- 2125443
E-mail: registratura@anph.ro
Web: www.anph.ro

© Powered by ICI - 2007

Figure 2.b. Useful Information

In every page of this site there exists a box that contains National Authority for People with Disabilities contacts and a link to a sensitive map for every unit in the country.

In the Public Information sections we find petition models, forms and other documents of this type and information about all kind of documents that National Authority for People with Disabilities elaborates and those are public interest information (figure 3).


Autoritatea Naţională pentru Persoanele cu Handicap - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.anph.ro/info_pub.php?m=Informatii%20publice
Go
Customize Links
Free Hotmail
Windows Marketplace
Windows Media
Windows
MINISTERUL MUNCII, FAMILIEI ŞI EGALITĂŢII DE ŞANSE
AUTORITATEA NAŢIONALĂ PENTRU PERSOANELE CU HANDICAP
Acasă
Informaţii utile
Informaţii Publice
Întrebări
Mass Media
Forum
Hartă site
Contact
Căutare site
Schimbă fontul : [a] [a]
INFORMATII PUBLICE
Documente
Lista Documente
Harta Instituţiilor Publice

Informaţii publice
Accesul oricărei persoane, fizice sau juridice, la informaţiile de interes public, produse şi/sau gestionate de Autoritatea Naţională pentru Persoanele cu Handicap se face din oficiu sau la cerere - formulată în scris sau verbal.

Informaţii

Accesul oricărei persoane, fizice sau juridice, la informaţiile de interes public, produse şi/sau gestionate de Autoritatea Naţională pentru Persoanele cu Handicap se face din oficiu sau la cerere - formulată în scris sau verbal. Autoritatea Naţională pentru Persoanele cu Handicap are obligaţia să comunice din oficiu următoarele informaţii de interes public:

- actele normative care reglementează organizarea şi funcţionarea ANPH şi a instituţiilor subordonate;
- proiectele de acte normative elaborate de ANPH, în condiţiile Legii nr. 52/2003 privind transparenţa decizională în administraţia publică;
- structura organizatorică şi atribuţiile compartimentelor organizate în cadrul aparatului propriu, programul de funcţionare şi programul de audienţe;
- coordonatele de contact ale ANPH, respectiv: denumirea, sediul, numerele de telefon, fax, adresele de e-mail ale salariaţilor şi adresa paginii de Internet;
- sursele financiare şi datele despre buget;
- programele şi strategiile proprii;
- lista cuprinzând documentele de interes public;
- lista cuprinzând categoriile de documente produse şi/sau gestionate potrivit legii;
- modalităţile de contestare a deciziei ANPH în situaţia în care persoana care se consideră vătămată în privinţa dreptului de acces la informaţiile de interes public solicitate.

Accesul la informaţiile de interes public se realizează prin una din următoarele modalităţi:

- afişare la sediul Autorităţii Naţionale pentru Persoanele cu Handicap;
- publicarea în mijloacele de informare în masă;
- publicarea în Monitorul Oficial al României;
- publicare în pagina de Internet proprie;
- consultarea lor la sediul ANPH.

AUTORITATEA NAŢIONALĂ PENTRU PERSOANELE CU HANDICAP
Calea Victoriei 194, sect. 1, Bucureşti
Tel: 021-2125438 ;
021-2125439
Fax: 021- 2125443
E-mail: registratura@anph.ro
Web: www.anph.ro

© Powered by ICI - 2007
Acest site a fost creat de Autoritatea Naţională pentru Persoanele cu Handicap pentru a oferi informaţii tuturor celor interesaţi de domeniul protecţiei, integrării şi incluziunii sociale a persoanelor cu handicap (persoanele cu dizabilităţi, aparţinători, organizaţiile neguvernamentale). Site-ul promovează orice iniţiativă legată de integrarea socială, profesională şi culturală a persoanelor cu handicap din România. Sperăm să putem contribui la îmbunătăţirea calităţii vieţii persoanelor cu nevoi speciale!

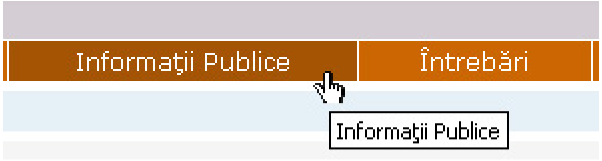
Figure 3.Public Information

Forum sections is very important for the visitors of the site because you can use it to find out useful information and you can discuss on every subject with other visitors

**Specific Facilities**

During the development of this project we respected all the accessibility and usability rules that a site for people with disabilities must obey.

For blind people the applications must supply equal access of all content and visual aspects of technology wherewith we wish to transmit the information. We used the description of the text (alternative text) of all static pictures for example (pictures, logos).


Informaţii Publice | Întrebări

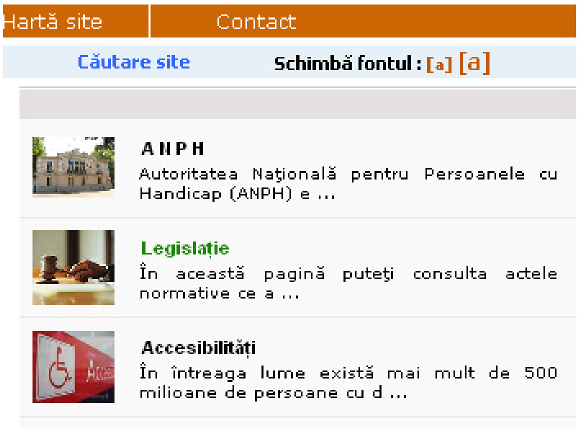
Informaţii Publice

Figure 4. ALT Option

Thereby the text can be read with screen readers and Braille machines (figure 4).

There are programs that synthesize the text and read the text with loud voice. The screen reader technology has limitations doesn't recognize graphic elements like buttons or other image elements without having a text attached attribute compliance "alt" and "title". We used a long description of the images that are very important we have left a considerable space between all the items we avoided t use ostentatious colours. Another facility of this site is that you can make fonts larger.

People that have sight deficiency can make fonts larger on every page of this site but in any moment they could turn back to normal size fonts as we can see in the picture below (figure 5.b).


Hartă site | Contact

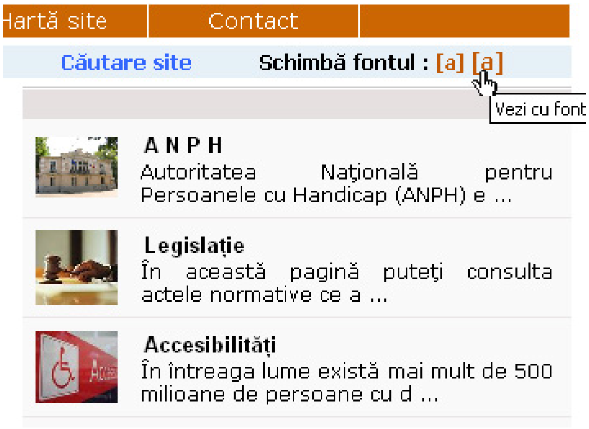
Căutare site    Schimbă fontul : [a] [a]

**ANPH**
Autoritatea Naţională pentru Persoanele cu Handicap (ANPH) e ...

**Legislaţie**
În această pagină puteţi consulta actele normative ce a ...

**Accesibilităţi**
În întreaga lume există mai mult de 500 milioane de persoane cu d ...

Figure 5.a. Normal Fonts


Hartă site | Contact

Căutare site    Schimbă fontul : [a] [a]

Vezi cu font

**A N P H**
Autoritatea Naţională pentru Persoanele cu Handicap (ANPH) e ...

**Legislaţie**
În această pagină puteţi consulta actele normative ce a ...

**Accesibilităţi**
În întreaga lume există mai mult de 500 milioane de persoane cu d ...

Figure 5.b. Enlarged Fonts

**Administration page**

Represents an entire section of pages and web forms, used for introducing and for modification of the site content. The access in this section is restricted (see figure 6), divided on more levels and the access is based on a username and password. Some of the administrators have limited rights over the site.

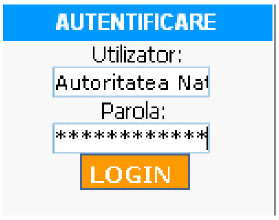

AUTENTIFICARE
Utilizator:
Autoritatea Nat
Parola:
************
LOGIN

Figure 6.Administrator Authentication

**Data Insertion**

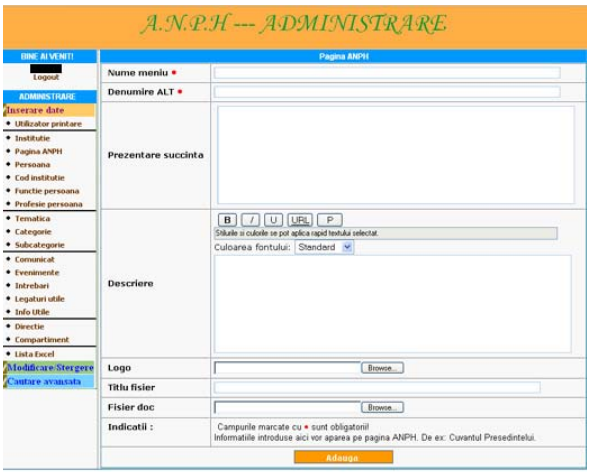
This section permits to right new information and data in the database. We can insert text type details for every category and we can upload files and images (see figure 7).

After we introduced the data it can be made actual from the section.

### Data modification

This part is more complex accountable the data insertion section. Because assumes the partial or total change of already existent files (figure 8).

We can add additional data or we could eliminate certain data or we can totally erase a category from the database. We can modify the pictures that are attached to categories.



[redacted]

Pagina ANPH

Nume meniu •

Denumire ALT •

Prezentare succinta

Descriere

Stilurile si culorile se pot aplica rapid textului selectat.

Culoarea fontului: Standard

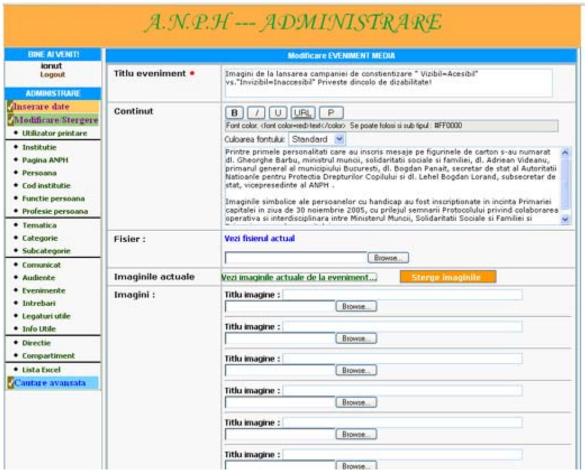
Logo

Titlu fisier

Fisier doc

Indicatii : Campurile marcate cu • sunt obligatorii!
Informatiile introduse aici vor aparea pe pagina ANPH. De ex: Cuvantul Presedintelui.

Figure 7. Data Insertion



Modificare EVENIMENT MEDIA

Titlu eveniment • Imagini de la lansarea campaniei de constientizare " Vizibil=Acesibil" vs."Invizibil=Inaccesibil" Priveste dincolo de dizabilitate!

Continut

Font color: <font color=red>text</color> Se poate folosi si sub tipul : #FF0000

Culoarea fontului: Standard

Printre primele personalitati care au inscris mesaje pe figurinele de carton s-au numarat dl. Gheorghe Barbu, ministrul muncii, solidaritatii sociale si familiei, dl. Adriean Videanu, primarul general al municipiului Bucuresti, dl. Bogdan Panait, secretar de stat al Autoritatii Natioanle pentru Protectia Drepturilor Copilului si dl. Lehel Bogdan Lorand, subsecretar de stat, vicepresedinte al ANPH .

Imaginile simbolice ale persoanelor cu handicap au fost inscriptionate in incinta Primariei capitalei in ziua de 30 noiembrie 2005, cu prilejul semnarii Protocolului privind colaborarea operativa si interdisciplinara intre Ministerul Muncii, Solidaritatii Sociale si Familiei si

Fisier : Vezi fisierul actual

Imaginile actuale: Vezi imaginile actuale de la eveniment...

Imagini :
Titlu imagine :
Titlu imagine :
Titlu imagine :
Titlu imagine :
Titlu imagine :
Titlu imagine :

Figure 8. Data Modification

## Conclusion

Web interface must be made in such a way that will offer equal access to all the people, whatsoever if they are persons with disabilities or normal people. We noticed that the information is different understood.

Online navigation permits to every user to interact with the material and preferred way supporting his strong points and trying to reduce weal points.

## Bibliography

L.Ciocoiu, I. Petre, D. Smada, D. Barbu, D. Nicolau – Multimedia Portal dedicated for people with disabilities for National Authority for People with Disabilities – 2006; http://www.anph.ro

L.Ciocoiu, L.Constantinescu, I. Petre, D. Smada, D. Barbu, D. Nicolau - eBiMuz - Integrated multimedia system for access to the multicultural thesaurus of the areas inhabited by Romanians, as integrated part of the European culture – CEEX 142/2005; http://ebimuz.ici.ro

L.Ciocoiu, I. Petre, D. Smada, D. Barbu, D. Nicolau – eMeditur – A tool for delivering information for online services in the medical assistance and tourism - R2130/2005 – http://emeditur.ici.ro

## Authors' information

***Laura Ciocoiu*** *– senior researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: ciocoiu@ici.ro; http://intelligent-agents.ici.ro*

***Ionuț Petre*** *– researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: ipetre@ici.ro; http://intelligent-agents.ici.ro*

***Dragoş Smada*** *– researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: dsmada@ici.ro; http://intelligent-agents.ici.ro*

***Dragoş Nicolau*** *– senior researcher; National Institute for Research and Development in Informatics; 8-10 Averescu Avenue, 011455 Bucharest 1 Romania; e-mail: dragos@ici.ro; http://intelligent-agents.ici.ro*

# Knowledge Discovery and Engineering

## AN ONTOLOGY- CONTENT-BASED FILTERING METHOD

Peretz Shoval, Veronica Maidel, Bracha Shapira

(Keynote Speech)

***Abstract:*** *Traditional content-based filtering methods usually utilize text extraction and classification techniques for building user profiles as well as for representations of contents, i.e. item profiles. These methods have some disadvantages e.g. mismatch between user profile terms and item profile terms, leading to low performance. Some of the disadvantages can be overcome by incorporating a common ontology which enables representing both the users' and the items' profiles with concepts taken from the same vocabulary.*

*We propose a new content-based method for filtering and ranking the relevancy of items for users, which utilizes a hierarchical ontology. The method measures the similarity of the user's profile to the items' profiles, considering the existing of mutual concepts in the two profiles, as well as the existence of "related" concepts, according to their position in the ontology. The proposed filtering algorithm computes the similarity between the users' profiles and the items' profiles, and rank-orders the relevant items according to their relevancy to each user. The method is being implemented in ePaper, a personalized electronic newspaper project, utilizing a hierarchical ontology designed specifically for classification of News items. It can, however, be utilized in other domains and extended to other ontologies.*

***Keywords***: *Ontology, Retrieval models, Information filtering, Content-based filtering, User profiles.*

***ACM Classification Keywords***: *H.3 Information Storage and Retrieval, H.3.1 Content Analysis and Indexing, H.3.3 Information Search and Retrieval, I.7 Document and Text Processing.*

## Introduction

In content-based filtering, the representations of the contents of items (e.g. documents, News), i.e. the items' profiles, are compared with the representation of the users, i.e. the users' profiles, in order to find the items whose contents are most relevant to each user. It is assumed that a user's profile and an item's profile share a common method of representation (e.g., by keywords) in order to enable matching and measuring the similarity between the profiles. The output of the matching process can be expressed as a ranking score, indicating the similarity between the user's profile and each item.

A user profile can be generated in various ways, including explicit definition by the user, or implicit analysis of the user's behavior (e.g. by logging and analyzing what the user read). An item's profile too can be generated in various ways, e.g. explicitly, by asking the originator (author) to specify proper index terms, or automatically, using a text classification algorithm which extracts terms representing the item's content in the best way. At any rate, no matter which method is used for creating either type of profile, content-based filtering has drawbacks due to well known problems of term ambiguity. For example, different terms may be used to represent the same content or the same user (synonymy); or, the same term may be used to represent different contents or different users (homonymy).

A possible way to overcome such problems of ambiguity might be through the use of ontology, i.e., a controlled vocabulary of terms or concepts, and semantic relationships among them. An ontology can bridge the gap between the terms in the users' profile and the terms representing the items. An ontology can be organized in various ways. For example, a taxonomy is a hierarchical structure with is-a relationships; in a thesaurus there are a few more types of relationships, e.g. BT/NT (broader-term; narrower terms) and general relatedness. Note that a thesaurus is a graph, not a hierarchy, because a term may have many NTs and more than one BT. In the newspapers domain, which is exemplified in this study, there exists an ontology specifically generated for classification of News named NewsCodes, created by IPTC [Le Meur and Steidl, 2004].

Assuming that there exists an ontology of a specific domain, which is used for representing both users (user profiles) and contents of items (item profiles), the research question we deal with is how exactly to match and measure the similarity between a user's profile and an items' profile. Obviously, if a user's profile includes exactly the same concept (terms) as an item's profile, there is some similarity between them; but the two profiles may include different concepts and still be similar to a certain degree – depending on if and how "close" are concepts in the two profiles with respect to the common ontology.

This research is conducted within the framework of *ePaper*, an electronic personalized newspaper research project, which is aimed to provide a personalized electronic newspaper to each reader. In the News domain, instant filtering of News items is important. Since a new item has no reading history, the filtering and personalization cannot rely on collaborative filtering (as opposed to other domains such as recommendation of books, movies, etc.), but rather need to rely on content-based filtering, so that once a new item arrives to the repository of News, the content-based filtering algorithm can perform the necessary matching with the users' profiles and determine the degree of relevancy of each item to potential users. If many News items accumulate in a certain period of time, the content-based filtering algorithm can rank-order the items according to their relevancy to each of the potential readers.

The remaining of this paper is structured as follows: The next section provides a background on content-based filtering and on ontological modeling, and reviews related research on conceptual and ontological modeling employed in content-based filtering. The third main section of the paper presents the proposed method for the ontology- content-based filtering, along with an example. The fourth section describes the evaluations that we plan to conduct with the proposed method, and the last section summarizes and proposes further research and extensions to the proposed method.

## Background on Content-Based Filtering and Ontological Modeling

### Content-Based Filtering

The information filtering approach is based on the information retrieval (IR) domain and employs many of the same techniques [Hanani et al., 2001]. One aspect by which information filtering differs from IR is with respect to the uses' interests: while in IR the users poses ad-hoc queries, in information filtering the users have profiles which represent their long-term interests, and the filtering system tries to provide to each user relevant items on a long-term basis. As said, the user profiles, as well as the item profiles, may consist of sets of terms. Based on some measure of similarity between the respective profiles, the filtering system selects and rank-orders the relevant items and provides them to the user.

The actual relevancy of an item provided by the system to a user can be determined by explicit or implicit user feedback. Explicit feedback requires the user to express the degree of relevancy of the provided item, while in implicit feedback the relevancy of the item is inferred by observing the user's behaviour, e.g. the reading time. Implicit feedback may be more convenient for the user but more difficult to implement and less accurate. User feedback enables to update the user's profile according to what he/she actually read, liked or disliked.

There exist two main approaches in information filtering: collaborative and content-based. In collaborative filtering, the system selects and rank-orders items for a user based on the similarity of the user to other users who read/liked similar items in the past. In content-based filtering, the system selects and rank-orders items based on the similarity of the user's profile and the items' profiles.

A major advantage of content-based filtering is that users can get insight into the motivation why items are considered relevant to them, because the content of each item is known from its representation. Content-based filters are less affected by problems of collaborative filtering systems [Claypool et al., 1999] such as "cold start" and sparsity: if a new item is added to the repository, it cannot be recommended to a user by a collaborative filter before enough users read/rated it. Moreover, if the number of users is small relative to the volume of items in the repository, there is a danger of the coverage of ratings becoming very sparse, thinning the collection of recommendable items [Balabanovic and Shoham, 1997]. For a user whose tastes are unusual compared to the rest of the population, the system will not be able to locate users who are particularly similar, leading to poor recommendations.

But content-based filtering has disadvantages too. One of them is that it focuses on keyword similarity. This approach, however, is incapable of capturing more complex relationships at a deeper semantic level, based on different types of attributes associated with structured objects of the text [Dai and Mobasher, 2001]. Consequently, many items are missed and many irrelevant items are retrieved [Blair and Maron, 1985].

Unlike humans, content-based techniques have difficulty in distinguishing between high quality and low quality information, since both good and bad information might be represented by the same terms. As the number of items increases, the number of items in the same content-based category increases too, further decreasing the effectiveness of content-based approaches [Claypool et al., 1999]. Another disadvantage of content-based methods is that they require analyzing the content of the document, which is computationally expensive and even impossible to perform on multimedia items which do not contain text.

To expand the first point of the disadvantages, it can be added that there is a tremendous diversity in the words people use to describe the same concept (synonymy) and this places strict and low limits on the expected performance of keyword-based systems. If the user uses different words from the organizer (indexer) of the information, relevant materials might be missed. On the other hand, the same word can have more than one meaning (homonyms), leading to irrelevant materials being retrieved [Dumais et al., 1988]. This disadvantage is added to the fact that the basic models of content-based filtering assume a representation of documents as sets or vectors of index-terms and typically employ only primitive search strategies based solely on the occurrence of term or combinations of terms [Knappe, 2005].

Thus, extensions to the traditional content-based filtering methods should be considered. Extensions may include additional knowledge in the form of a coherent taxonomy of concepts in the domain spanned by the items. This type of conceptual knowledge would provide means for item and user profile representation.

There is a need for devising a content-based approach which extends the classical models, where the use of simple natural language analysis in combination with the knowledge contained in an ontology forms the basis for representations of both user profiles and items. Consequently, items can be described using a concept language and be directly mapped into the ontology. The similarity between such representation of the user and the representations of the items will be based on the proximity principle stating that the distance of two descriptive items in the ontology is directly related to their similarity [Knappe, 2005].

## Ontological Modeling

Ontology is a specification of a conceptualization. It can be described by defining a set of representational concepts. These definitions are used to associate the names of entities in the universe (e.g., classes, relations, functions or other objects) with human-readable text, describing what the names mean, and formal axioms that constrain the interpretation and focus the well-formed use of these concepts [Khan, 2000]. When constructing an ontology, not only concepts and relationships are defined, but also the context in which the concept (relationship) applies. Therefore, an ontology defines a set of representational terms which are called concepts, and the interrelationships among the concepts.

Linguistic ontologies (e.g., WordNet) and thesauri express various relationships between concepts (e.g. synonyms, antonyms, is-a, contains-a), and have a hierarchical structure based on the relations between concepts. But they do not explicitly and formally describe what a concept means [Khan, 2000]. WordNet, for example, is an electronic lexical database that contains nouns, verbs, adjectives and adverbs which are organized into synonym sets (*synsets*), each representing one underlying lexical concept [Magnini and Strapparava, 2001]. It is offering two distinct services: a vocabulary which describes the various word senses, and an ontology which describes the semantic relationships among senses [Guarino et al., 1999].

An example of a domain ontology is the IPTC NewsCodes [Le Meur and Steidl, 2004]. This is a 3-level hierarchical ontology of concepts targeted to News description; it currently contains approximately 1,400 concepts. A first level concept of NewsCodes is called Subject; a second level – Subject Matter, and a third – Subject Detail.

Figure 1 demonstrates an example.

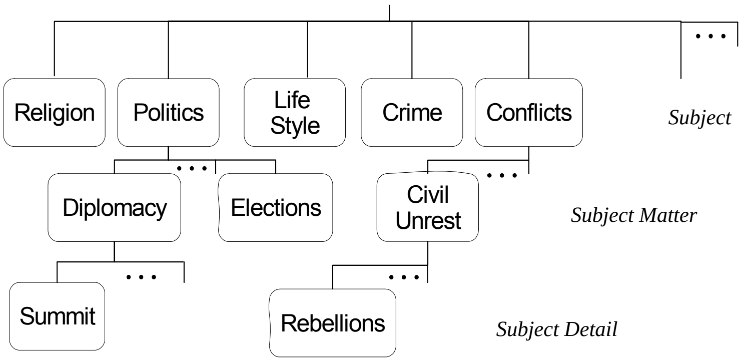


Figure 1: Example of IPTC NewsCodes ontology

## Related Work

Savia et al. [1998] was one of the first to present a hierarchical representation for describing documents and user profiles by attaching metadata to each document and using the same method to generate a compatible representation of users' interests. A hierarchical representation was chosen in order to develop a document classification system understandable to humans and yet not restricted to text documents. Savia et al. chose to take advantage of the hierarchical metadata concepts along with an asymmetric distance measure, which considers not only the concepts appearing both in a user's profile and in the document's profile, but also concepts appearing in an a document's profile which do not appear in the user's profile. The underlying assumption for the asymmetric measure was that the best matching documents are not necessarily those that cover all the interests at the same time. Distance computations were performed on different levels of the hierarchy and the metadata was represented in a fuzzy distribution among the leaf nodes of the concept tree.

Ontological and conceptual modeling was used in order to extract user profiles, such as the four-level ontology used in the Quickstep system [Middleton et al., 2001] which recommends papers to researchers by combining both content-based and collaborative filtering techniques. Papers were represented as term vectors with term frequency normalized by the total number of terms used for a term's weight. Whenever a research paper was browsed and had a classified topic, it accumulated an interest score of that topic for the particular user. In the ontology-based user profile, whenever a topic received some interest all its super classes gained a share: the immediate super-class gained 50% of the main topics value; the next super-class gained 25%, and so on. This way, general topics rather than just the most specific ones were also included in the profile and thus produced a broader profile. Recommendations were computed based on the correlation between the user's topics of interest and papers classified to those topics.

Another work which used ontology for content-based retrieval was the electronic publishing system CoMet [Puustjärvi and Yli-Koivisto, 2001]. CoMet extracted metadata information both about users and about contents of documents (document profiles) and stored the metadata in hierarchical ontology structures. Comparison between a user's profile and the documents' profiles was performed by finding the largest combined hierarchy (LCH), which is the largest hierarchy that the user profile and the document profile share in the ontology. By using the weights on the nodes in each level, a similarity measure was calculated between the documents that had an LCH with a user profile. In a weighted LCH-matching, the deepness of the LCH was emphasized in the matching calculations, since the depth of the hierarchy has a significant effect on the expression power of the incorporated ontology. The depth of the profile was also suggested to be used as a generalization tool. For example, if a user is interested in news items on F1 (a sport car), one can assume that she would like to view other motor sport related items when F1 news items are not available. The result of the matching generated a set of news items most suitable for the user according to the calculation result of LCH-matching.

Pereira and Tettamanzi [2006] illustrated a novel approach to learning users' interests on the basis of a fuzzy conceptual representation of documents, by using information contained in an ontology. Instead of a keyword representation, documents were represented as a vector of components expressing the “importance” of the concepts. In order to choose the concepts that would represent a document, they considered both the leaf concepts and the internal nodes of the ontology. The internal nodes were implicitly represented in the importance vector by “distributing” their importance to all their descendants down to the leaf concepts. All documents with a certain level of similarity were grouped together into fuzzy clusters, in order to express user interests with respect to clusters instead of individual documents. Since the clusters were fuzzy, each document received its membership degree for that cluster, meaning it could belong to more than one cluster. A user model was represented as a vector of membership degrees which described the model’s guess of the extent to which the user was interested in each document cluster. A user profile was set up by adding to the list of its interest groups the instances of clusters with features similar to those requested by the user.

In the above survey we have emphasized methods involving the incorporation of ontologies both for user profile generation and for representation of items. Some of the methods employed ontology in order to acquire user profiles more accurately, while others used ontology in order to perform disambiguation of a query/user profile. In most cases, the ontology was used in all of the steps taken towards the retrieval of items according to the query/user profile. All studies which incorporated ontology in their content-based filtering method provided better and more accurate results compared to traditional content-based methods. This encouraged us to adopt the ontology approach and inspired us to introduce a novel filtering method which incorporates an ontology.

## The New Method for Ontological-Content-based Filtering

### Research Goal

The aim of this research is to develop, implement and evaluate a new ontology-based filtering method, which filters and ranks relevant items by measuring the similarity of user profiles and item profiles, both consisting of the ontology concepts, by considering the "closeness" (or distance) of concepts in the profiles based on their location in the ontology. We utilize the method in the News domain, as part of *ePaper*, a research project which includes the development of a personalized electronic newspaper system. In this research we incorporation an ontology for the News domain and exploit its three-level hierarchy in the representation of user profiles and News items profiles, and in the process of matching between them.

## The Ontological- Content-based Filtering Method

The filtering method, initially proposed by Shoval [2006], is based on the assumption that each item (e.g. a News item) and each user profile (e.g. a reader of the *ePaper*) is represented with a set of concepts taken from the ontology. In the *ePaper* system we use the IPTC NewsCodes ontology. It may be assumed that the generation of an item's representation (profile) is done automatically, utilizing some classification technique which analyses both the metadata describing the item and the actual text of the item. (We do not elaborate here on how this is done because as it is not an essential part of the proposed method.) Similarly, it may be assumed that an initial user profile is generated explicitly by the user who selects concepts from the ontology and assigns them weights of importance. Subsequently, the concepts in the initial user profile and their weights are updated implicitly, based on monitoring the items actually read by the user and considering the ontology concepts by which those items are represented. (This part too is not elaborated here; suffice is to know that at any point in time a user's profile contains an up-to-date weighted set of ontology concepts.)

Following are the details of the filtering method.

**Representation of contents – an item's profile:**

An item's profile consists of a set of ontology concepts which represent its content. The concepts representing an item are the most specific ones in a certain branch of the hierarchy. For example, if an item deals with 'sport' and specifically with 'football', it is represented with 'football' only; the ontology can tell that latter is a child (subtype) of the former.

Obviously, an item may be represented with many ontology concepts; each concept may appear in any branch of the ontology hierarchy and at any level – all depending on the actual content of that item. For example, an item's profile may include the concepts 'politics' (a top-level concept), 'football' (child of 'sport') and 'rebellions' (grandchild of 'conflicts'). Note that the profile may include sibling concepts, i.e. children of the same super concept. For example, an item's profile may include both 'football' and 'basketball' (children of 'sport').

Note that we do not assume that the concepts representing an item are weighted, although the proposed filtering algorithm can be adjusted for such possibility.

**Representation of users – a user's profile:**

A user's content-based profile consists of a weighted list of ontology concepts representing his/her interests. Obviously, a user's profile may consist of many ontology concepts, each appearing in different branches and at different levels of the hierarchy. For example, a user's profile may include 'sport' only, or 'sport' and 'football', or 'football' and 'basketball', or all the three – besides many other concepts. This means that a certain concept in an item's profile may be "matched" (i.e. compared) with more than one equivalent concept in the user's profile. For example, if an item's profile includes 'football' and a user's profile includes both 'sport' and 'football', then there is a "perfect match" between the two profiles because of the common concept 'football', and also a "partial match" because of the parent concept 'sport'.

As stated before, the user's content-based profile may be generated initially by the user who selects concepts from the ontology and assign them weights of importance. (The total of the weights is normalized 100%). Then, the user's profile is updated all the time according to implicit feedback from the user: when a user reads an item and finds it interesting, the concepts in that item's profile which are not yet in the user's profile are added to it, and the weights of all concepts in this profile are recalculated as follows: a new concept it is added with 1 'click' (a 'click' indicates how many times that concept was found relevant to the user); the weight of an existing concepts is increased by 1 'click'. The weight of each concept in the user's profile is the number of its 'clicks' divided by the total number of 'clicks' in the user's profile. (Hence, the weights sum up to 100 %.)

**Measuring similarity between an item and a user:**

An item and a user are similar to a certain degree if their profiles include common (the same) concepts or related concepts, i.e. concepts having some kind of parent-child relationship. An item's profile and a user's profile may have many common or related concepts; obviously, the more common or related concepts, the stronger is the similarity between them. For example, if a user's profile includes 'football' and 'sport', this profile is similar (to a certain degree) to an item including these two concepts, but it is less similar to an item including just 'sport', and it is more similar to an item including 'sport' and 'football' and 'basketball'.

In the *ePaper* project, we adopted the 3-level NewsCodes ontology, so related concepts may be only one or two levels apart (parent-child or grandparent-grandchild), but generally concepts may be more levels apart. It is obvious that the closer two concepts are in the ontology, the closer are the two objects which they represent (i.e. the user and the item).

When dealing with related concepts appearing a user's profile and in an item's profile, two different cases can be distinguished: in one case, the concept in the user's profile is more general than the related concept in the item's profile (one or two levels apart), meaning that the user has a more general interest in the topic which the item deals with. In the other case, the concept in the user's profile is more specific than the related concept in the item's profile (one or two levels apart), meaning that the user has more specific interests in the topic dealt in the item. In any of the above cases of "partial match" between the user and item concepts, we should also consider the distance between the concepts: two related concepts which are only one level apart are closer (i.e. more similar) than two concepts which are two levels apart.

**Scores of similarity:**

Based on the above, in a 3-level hierarchical ontology we can distinguish between 9 different possible cases of similarity between concepts in a user's profile and an item's profile, as portrayed in Figure 2.

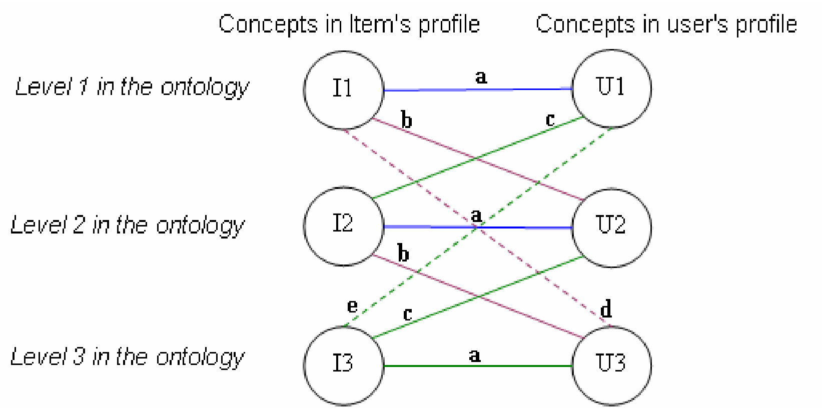


Figure 2: Hierarchical similarity measure

- **"Perfect match"**: the concept appears in both the user's profile and the item's profile. I1, I2, I3 (see Figure 1) denote the level of a concept in an item's profile, and U1, U2, U3 - the level of a concept in the user's profile. A 'perfect match' can occur in 3 cases:
  - I1=U1 (e.g. both item and user profiles include 'sport')
  - I2=U2 (e.g. both item and user profiles include 'football')
  - I3=U3 (e.g. both item and user profiles include 'Mondeal games')
- **"Close match"**: a concept appears only in one of profiles, while a parent or child of that concept appears in the other profile. A 'close match' can occur in 2 **pairs** of cases:
  - I1=U2 (e.g. item concept is 'sport', while user concept is 'football')
  - I2=U3 (e.g. item concept is 'football' while user concept is 'Mondeal games')

  In the above 2 cases, the item's concept is more general than the user's concept (1 level apart), i.e. the user interest is more precise/specific than the item.

  - I2=U1 (e.g. item concept is 'basketball' while user concept is 'sport')
  - I3=U2 (e.g. item concept is 'Euro league' while user concept is 'basketball')

  In the above 2 cases, the item concept is more specific than the user concept, i.e. the user's interest is more general.

  Note that in all the above 4 cases there may be more than one occurrence of 'close match' between the concepts. For example, in the case I1=U2, assume the item's concept is 'sport' while the user's profile includes both 'football' and 'basketball'. When measuring similarity we have to consider all possible 'close matches' between parent and children concepts.

- **"Weak match":** a concept appears in one profile, while a grandparent concept or a grandchild concept appears in the other profile (concepts are 2 levels apart). A 'weak match' can occur in 2 cases:
  - I1=U3 (e.g. item concept is 'sport' while user concept is 'Mondeal games') – in this case the item is much more general than the user's interest.
  - I3=U1 (e.g. item concept is 'Euro league' while user concept is 'sport') – in this case the item is much more specific than the user's interest.

  Recall that there may be more than one occurrence of 'weak match' between the concepts. For example, in the case I3-U1 the user concept is 'sport' while the item concepts include 'Euro league' and 'Mondeal games'.

For each of the 9 possible cases we determine a score of similarity. In the 3 cases of "perfect" match' labeled 'a' (see Figure 2) the score is 1 (maximal); in all other cases the score should be less than 1, depending if it is a 'close' or a 'weak' match and on the "direction" of the relationship, i.e., whether the user's concept is more general or more specific than the item's concept. For example, the score for the case I1=U2 (the item's concept is more general than the user's concept) may be 2/5, while the case I2=U1 (the item's concept is more specific than the user's concept) may score 2/3 – higher. The rationale for this may be that in the first case the item deals with a more general concept than the user's interest, yielding lower Precision than in the other case, where the item deals with a more specific concept than the user's interest, thus yielding higher Precision. But this assumption, as well as the exact scores of similarity for all possible cases is subject to experimentation.

The following is a possible scoring scheme for the 9 possible cases:

- I1=U1 → 1; I2=U2 → 1; I3=U3 →1 (3 cases of "perfect match"; marked **a** in Figure 2)
- I1=U2 → 2/5; I2=U3 → 2/5 (2 cases of "close match" - item concept is more general; marked **b**)
- I2=U1 → 2/3; I3=U2 → 2/3 ( 2 cases of "close match" - item concept is more specific; marked **c**)
- I1=U3 →1/3 (case of "weak match" - item concept is much more general; marked **d**)
- I3=U1 → 1/2 (case of "weak match" - item concept is much more specific; marked **e**)

**Measure of similarity between item and user:**

The similarity of an item's profile to a user's profile is based on the number of "perfect match", "close match" and "weak match" of concepts in the two profiles, and on the weights of the concepts in the user's profile. The overall Item Similarity score (IS) is computed as follows:

$$IS = \frac{\sum_{i \in Z} N_i \cdot S_i}{\sum_{j \in U} N_j}$$

where:

*Z* - number of concepts in item's profile

*U* - number of concepts in user's profile

*i* - index of the concepts in item's profile

*j* - index of the concepts in user's profile

$S_i$ - score of similarity, depending if it is a "perfect", "close" or a "weak" match of concept *i* to a respective concepts in the user's profile. (Note that in case of no match at all, $S_i = 0$.)

$N_i$ - number of clicks on the concept (used to determine the concepts' weights)

**The matching algorithm:**

The algorithm can be applied for measuring the similarity of a single item to a single user, or for rank ordering by relevancy a batch of items for a single user, or for rank ordering by relevancy a batch of users for a single item, or for rank ordering by relevancy a batch of items for a batch of users – all depending on the specific need/application.

The algorithm described below is for measuring the similarity of a single item's profile to a single user's profile. The algorithm is expressed in pseudo-code; it does not refer to any specific programming language, database system and other implementation aspects. However, it may be assumed that due to size on one hand and efficiency on the other hand, during execution the ontology resides in memory.

Since a user's profile may include many concepts (depending, among else, on how many items he already read), some with very low weights ('clicks'), it might be worthwhile to include in the computation of similarity only the most important concepts, e.g., the top 10 concepts or the concepts having weight above a certain threshold. The exact number of concepts has to be determined in experiments.

The algorithm consists of two loops: one over the concepts in the Item-list (i.e., list of concepts in the item's profile), searching for matches in the User-list (i.e., list of concepts in the user's profile); the other loop is over the User-list, searching for matches in the Item-list. Within each loop, if there is no "perfect match" the search is for a match with the parent or grand-parent of the item. (There is no need to search for children and grandchildren, a time-consuming task, because the first loop finds matches from the other list of concepts.)

Legend:

- Score: total score of similarity b/w item and concept
- I-concept: a concept in Item-list
- U-concept: a concept in User-list
- w: weight of concept in User-list that is being matched.

```
Begin
Score=0
Repeat for each I-concept in Item-list:
  Do case:
    - If I-concept is in User-list then Score= ++1*w      /*"perfect match"/
    - If parent of I-concept is in User-list then Score= ++ 2/3*w      /*"partial match" type c: I2=U1 or I3=U2/
    - If grandparent of I-concept is in User-list then Score= ++ 1/2*w      /*"weak match" type e: I3=U1/
  End case.
Until end of Item-list.
Repeat for each U-concept in User-list:
  Do case:      /*no need to check again for "perfect match" between concepts of same item and user profiles/
    - If parent of U-concept) is in Item-list then Score= ++ 2/5*w     /*"partial match" type b: I1=U2 or I2=U3/
    - If grandparent of U-concept is in Item-list then Score= ++ 1/3*w      /*"weak match" type d: I1=U3/
  End case.
Until end of User-list.
End.
```

Notes:

1) The scores for each type of match used in the algorithm are given as examples, as described above.

2) Not all user concepts must participate in the computation; as said, only the n-top concepts might be considered.

## Example

The following example demonstrates the application of the filtering method using a few simulated items' profiles and a user's profile. The calculations are based on the matching scores demonstrated above.

*Items' Profiles:*

| Item # | Ontology concepts representing the item* |
|---|---|
| Item 1 | Crime → Laws |
| | Unrest → Civil unrest → Social conflict |
| Item 2 | Sport → American Football |
| | Health → Injury |
| Item 3 | Science → Natural science → Astronomy |
| Item 4 | Life style and leisure |
| | Disaster and accident → Emergency incident |

*A User's Profile:*

| Ontology concepts in the user's profile | Number of clicks (weight) |
|---|---|
| Sport | 20 |
| Health | 12 |
| Crime → Laws → Criminal | 3 |
| Unrest | 10 |
| Lifestyle and leisure → Fishing | 8 |

** An arrow represents parent-child relationship. The item's profile includes only the lower-level concepts.*

The application of the algorithm yields the following rank ordered list of items:

| **Item #** | **Ranking score** |
|---|---|
| **Item 2** | **0.40** |
| **Item 1** | **0.11** |
| **Item 4** | **0.06** |
| **Item 3** | **0.00** |

It can be observed that Item 2 gets the highest score because its profile includes 'American football', a child of 'Sport' in the user's profile; and 'Injury', a child of 'Health' in the user's profile – and both concepts in the user's profile have relatively high weights. Here is the exact computation of the ranking score, assuming we consider the scoring scheme in which I2=U1 → 2/3:

$$IS = \frac{\frac{2}{3}\cdot 20+\frac{2}{3}\cdot 12}{20+12+3+10+8} = 0.4$$

Item 1 gets the second highest score because of the 'close match' between its 'Laws' concept and 'Crime' in the user's profile, and also because of the 'weak match' between its 'Social conflict' concept and 'Unrest' in the user's profile. Item 1 gets a lower ranking than Item 2 because of two reasons: 1) lower scores of similarity; 2) lower weight of the matched concepts. Item 4 gets even a lower ranking because it has only one concept having any match with the user's profile: its concept 'Lifestyle and leisure' is a 'close match' with 'Fishing' in the user's profile. Item 3 gets a ranking score 0 because it has no match at all with the user's profile.

## Evaluations of the Filtering Method

We plan to evaluate the filtering method in a controlled setting utilizing a prototype of the *ePaper* system. The main objective of the evaluations is to examine the effectiveness of the method, including the contribution of the various matching types (i.e. "perfect", "close" and "weak" matches) to performance, and to determine the optimal values for the various matching scores.

## Measures of Effectiveness

Traditional measures of effectiveness of information retrieval systems usually include Precision and Recall. But these measures may not be appropriate for evaluating the quality of rank ordered items because the user might read only some of the top ranked items, while Precision and Recall are based on the total number of relevant or retrieved items, respectively. We are considering several rank accuracy measures which are more appropriate to evaluate rank-ordered results, and where the users' preferences in recommendations are non-binary. Following Herlocker et al. [2004], we are considering the following measures:

- *Rank Correlations*, such as Spearman's $\rho$ and Kendall's Tau, which measure the extent to which two different rankings agree independent of the actual values of the variables.
- *Half-life Utility* metric, which attempts to evaluate the utility of a ranked list to the user. The utility is defined as the difference between the user's rating for an item and the "default rating" for an item.
- *NDPM Measure*, which is used to compare two different weakly-ordered ratings.

## What will be Evaluated

The evaluations will include the following objectives:

1. **Determination of the matching scores**: The filtering method assumes different matching scores to the various possible types of matching between concepts in the user's profile and the item's profile: the highest score (1) is given to a "perfect" match, while a "close" match and a "weak" match get lower scores, considering also the direction of the hierarchical relation between the concepts (i.e., whether the user's concept is more general or more specific than the item's concept). This part of the experimental evaluations is aimed to determine the optimal scores for the different types of match.
2. **Evaluation of the contribution of the various types of match between user concepts and item concepts:** It is obvious that the more common concepts appear in both the user's profile and the item's profile, and the closer the user's concepts is to the item's concepts - the more relevant is the item to the user. The question is: what is the residual contribution of the different types of match (i.e. "closeness") to the quality of the results. For example, what is the quality of results if only "perfect" matches are considered? What is the additional contribution of "close" matches? What is the additional contribution of "weak" matches? Results of these evaluations may enable us to determine if it is worthwhile to consider all types of relatedness, or perhaps only some of them are sufficient to obtain quality results.
3. **Considering more than one match between related concepts in the user's and item's profiles:** A user's profile may contain concepts form various level of one branch of the hierarchy (e.g., the profile may include the concepts 'sport' and 'football'). The question is whether all concepts along the branch should be considered when compared to the item's profile, or perhaps only the concept having the highest score*weight. (Note that the score itself is determined according to the "closeness" factor, while the weight is determined according to the number of read items which included the concept).
4. **Determining the number of concepts in a user's profile to consider:** A user's profile may include many concepts, each having a certain weight (as explained above). Considering all concepts in the profile might be time consuming (in terms of processing time). It is likely that concepts having low weights will not contribute much to the quality of the filtering results. We will examine the contribution of low-weight concepts in order to determine a threshold for an optimal number of concepts or for concept weights. Initially, the algorithm will consider all concepts; then we will omit certain concepts (beyond a certain number or below a certain weight) and see to what degree it affects performance.

## The Evaluation Plan

We plan to conduct user studies with real users (subjects), each having a content-based profile representing his interests. Some of the subjects will have similar (overlapping) profiles and some will have dissimilar profiles, to enable finding out how the filtering method affect similar and dissimilar subjects.

The subjects will read News items delivered to them by the *ePaper* system and rate each item as "interesting" or "not interesting". Alternatively we are considering to use a scale bar (say from 1 to 5) to expresses the level of interest.

Some of the items read by a subject will be used for updating his/her profile (training set), while the remaining items will be used for the various tests described above (test sets). A test set of items, rank ordered by the algorithm (in any of its variations) will be compared to the subject's ratings of the items, and the measures of effectiveness (as described above) will be applied to determine the quality of the result. As described, the algorithm will vary from test to test:

1. As a result of the first set of tests (determination of the matching scores) we will adopt the best set of matching scores of similarity; these will be used in all the subsequent evaluations.
2. As a result of the second set of tests (evaluation of the contribution of the various types of match between user concepts and item concepts) we will determine the contribution of each level of proximity relatively to the performance obtained at the prior level, and determine if it is worthwhile to consider all or only parts of match types (e.g. only 'perfect' and 'close' matches).
3. As a result of the third set of tests (considering more than one match between related concepts in the user's and item's profiles) we will determine whether all concepts along a branch should be considered or only the concept having the highest score*weight. Based on that the filtering algorithm will be adjusted.
4. As a result of the fourth set of tests (determining the number of concepts in a user's profile to consider) we will calibrate the algorithm to consider only a certain number of concepts in the user's profile, limited by a threshold number or weight.

## Summary and Further Research

We presented a new content-based filtering method that uses ontology for representing user and item profiles and for ranking items according to their relevancy in the electronic newspapers domain. The method is being implemented in the *ePaper* system for personalized electronic newspaper. The filtering method considers the hierarchical distance, or closeness, between concepts in the user's profile and concepts in the items' profile.

The method can be enhanced in various aspects. One possible enhancement is to assign more importance to concepts appearing together in items read in the past by the user. An item including co-occurring concepts might get a higher score than an item including the same concepts that did not co-occur in past read items. The added value of the incorporation this enhancement will be examined before being implemented in the method.

Another possible enhancement of the method is to consider penalty scores for concepts appearing in an item but not in the user's profile. This idea, adopted from Savia et al. [1998], means that a concept in an item's profile which does not appear in the user's profile might be given a negative (penalizing) score. The contribution of such penalty to the quality of the filter can be determined in empirical experiments.

The proposed filtering method utilizes a 3-level hierarchical ontology of News. It can, however, be generalized to other domains with their specific ontologies; and it must not be restricted to three levels. Moreover, the method can be enhanced to deal not just with a hierarchical but also with a network-based (DAG) ontology, where a concept may have many parent concepts, not only child concepts. Another possible extension to the method is to consider more types of relations between concepts, besides parent-child and grandparent-grandchild, e.g. twins of concepts. For example, a use's profile may include 'football' while an item may include 'basketball'. These extensions will be dealt with in further research.

## Acknowledgment

This work is supported by Deutsche Telekom Laboratories at Ben-Gurion University.

## Bibliography

[Balabanovic et al., 1997] Balabanovic, M. & Shoham, Y. Fab: Content-based, collaborative recommendation. Communications of the ACM, 40(3), 66-72.

[Blair and Maron, 1985] Blair, D.C. & Maron, M. E. An evaluation of retrieval effectiveness for a full-text document retrieval system. Communications of the ACM, 28, 289-299.

[Claypool et al., 1999] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. & Sartin M. Combining content-based and collaborative filters in an online newspaper. Proc. of ACM SIGIR Workshop on Recommender Systems.

[Dai and Mobasher, 2002] Dai, H., & Mobasher, B. Using ontologies to discover domain-level web usage profiles. Proc. of the Second Semantic Web Mining Workshop at PKDD 2001, Helsinki, Finland.

[Dumais et al., 1988] Dumais, S.T., Furnas, G.W., Landauer, T.K. & Deerwester, S. Using latent semantic analysis to improve information retrieval. Proc. of CHI'88 Conference on Human Factors in Computing, New York: ACM, 281-285.

[Guarino et al., 1999] Guarino, N., Masolo, C. & Vetere, G. OntoSeek: Content-based access to the Web. IEEE Intelligent Systems 14(3), 70-80.

[Hanani et al., 2001] Hanani, U., Shapira, B. & Shoval, P. Information filtering: overview of issues, research and systems. User Modeling and User-Adapted Interaction (UMUAI), 11(3), 203-259.

[Herlocker et al., 2004] Herlocker, J.L., Konstan, J.A., Terveen, L.G. & Riedl, J.T. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5-53.

[Khan, 2000] Khan, L.. Ontology-based Information Selection. Ph.D. Thesis, University of South California.

[Knappe, 2005] Knappe, R. Measures of semantic similarity and relatedness for use in ontology-based information retrieval. Ph.D. Thesis, Roskilde University, Department of Communication, Journalism and Computer Science.

[Le Meur and Steidl, 2004] Le Meur, L. & Steidl, M. NewsML 1.2 – Guidelines V1.00. Int'l Press Telecommunications Council. Retrieved: Dec. 07, 2006: http://www.newsml.org/IPTC/NewsML/1.2/documentation/NewsML_1.2-doc-Guidelines_1.00.pdf

[Magnini and Strapparava, 2001] Magnini, B. & Strapparava, C. Improving user modelling with content-based techniques. Proc. of the 8th Int'l Conference on User Modeling 2001. M. In: Bauer, P., Gmytrasiewicz, J. & Vassileva, J. (Eds.): Lecture Notes in Computer Science, 2109. Springer-Verlag, London, 74-83.

[Middleton et al., 2001] Middleton, S.E., De Roure, D.C. & Shadbolt, N.R. Capturing knowledge of user preferences: ontologies in recommender systems. Proc. of 1st Int'l Conf. on Knowledge Capture, 100-107, Victoria, BC, Canada.

[Pereira and Tettamanzi, 2001] Pereira, C. C. & Tettamanzi, A. G. An ontology-based method for user model acquisition. Soft Computing in Ontologies and Semantic Web, Berlin, Springer.

[Puustjärvi and Yli-Koivisto, 2001] Puustjärvi, J. & Yli-Koivisto, J. Using metadata in electronic publishing. Project internal publication, available at http://www.soberit.hut.fi/comet/.

[Savia et al., 1998] Savia, E., Koskinen, T. & Jokela, S. Metadata based matching of documents and user profiles. Proc. of Finnish Artificial Intelligence Conference, STeP'98.

[Shoval, 2006] Shoval, P. Ontology and content-based filtering for the ePaper project. Working Paper, BGU.

## Authors' Information

***Peretz Shoval** – Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: shoval@bgu.ac.il*

***Veronica Maidel** – Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: maidel@bgu.ac.il*

***Bracha Shapira** – Department of Information Systems Engineering, Ben-Gurion University, Beer-Sheva 84105, Israel; e-mail: bshapira@bgu.ac.il*

# LOGIC BASED PATTERN RECOGNITION - ONTOLOGY CONTENT (2)[1]

**Levon Aslanyan, Vladimir Ryazanov**

***Abstract**: Logic based Pattern Recognition extends the well known similarity models, where the distance measure is the base instrument for recognition. Initial part (1) of current publication in iTECH-06 reduces the logic based recognition models to the reduced disjunctive normal forms of partially defined Boolean functions. This step appears as a way to alternative pattern recognition instruments through combining metric and logic hypotheses and features, leading to studies of logic forms, hypotheses, hierarchies of hypotheses and effective algorithmic solutions. Current part (2) provides probabilistic conclusions on effective recognition by logic means in a model environment of binary attributes.*

## 1. Introduction

Pattern Recognition consists in reasonable formalization (ontology) of informal relations between object's visible/measurable properties and of object classification by an automatic or a learnable procedure [1]. Similarity measure [1] is the basic instrument for many recognition formalisms but additional means are available such as logical terms discussed in part (1) of current research [2]. Huge number of recognition models follow the direct goal of increasing recognition speed and accuracy. Several models use control sets above ordinary learning sets, others use optimization and other direct forces. Besides, more alternative notions are available to describe algorithmic properties. In existing studies the role of these notions is underestimated and less attention is paid to these components. In part (1) the attention is paid to implementing the learning set through its pairs of elements rather than the elements separately. The following framework is considered: given a set of logical variables (properties) $x_1, x_2, ..., x_n$ to code the studied objects, and let we have two types/classes for classification of objects: $K_1$ and $K_2$. Let $\beta \in K_1$, and $\gamma \in K_2$, and $\alpha$ is an unknown object in sense of classification. We say, that $\gamma$ is separated by the information of $\beta$ for $\alpha$ if $\beta \oplus \gamma \le \beta \oplus \alpha$, where $\oplus$ is $mod\,2$ summation. Formally, after this assumption, the reduced disjunctive normal forms of two complementary partially defined Boolean functions appear to describe the structure of information enlargement of the learning sets. The idea used is in knowledge comparison. $\alpha$ is an object of interest. Relation $\beta \oplus \gamma \le \beta \oplus \alpha$ informs that the descriptive knowledge difference of $\beta$ and $\alpha$ is larger than the same difference of $\beta$ and $\gamma$. This approach we call logic separation. While notion of similarity gives the measure of descriptive knowledge differences, the logic separation describes areas which are preferable for classes and learning set elements. In general the question is in better use of learning set. The learning set based knowledge, which is used by recognition procedure, at least is supposed to reconstruct the learning set itself. It is indeed negative when these information is not able to reconstruct the learning set. It is easy to check that the similarity knowledge can't reconstruct an arbitrary learning set, and only special sets allow reconstructing of objects by their distances [3]. Restructuring power is high when comparison is used for the set of all attribute subsets. Theoretically such structures are studied in discrete tomography problems [4], but practically even the use of pairs draws to known hard computational area of disjunctive normal forms.

Consider pairs of elements of the learning set, where each pair contains elements of different classes (the case of 2 learning classes is supposed). It was shown [2] that the logical separators divide the object space into three areas, where only one of these areas needs to be treated afterward by AEA (algorithms of estimation analogies – voting algorithms) [1]. This set is large enough for almost all weakly defined Boolean functions, but for the functions with compactness property it is small. Let, for $0 \le k_0 < k_1 \le n$, $F_{n,k_0,k_1}$ be the set of all Boolean

[1] The research is supported partly by INTAS: 04-77-7173 project, http://www.intas.be

functions defined as follows: each of them has zero (false) value on the vertices of $k_0$-sphere centered at $\widetilde{0}$, and has one (true) value on $(n-k_1)$-sphere centered at $\widetilde{1}$. On the remainder vertices of $n$-cube the assignment/evaluation is arbitrary. These functions (for appropriate choice of $k_0$ and $k_1$) satisfy the compactness assumptions [8], and their quantity is not less than $2^{\varepsilon(n)2^n}$ for an appropriate $\varepsilon(n) \to 0$ with $n \to 0$. For these functions we have also, that for recovering the full classification by means of logical separators procedure, it is enough to consider a learning set which consists of any $n2^{n-\varepsilon(n)\sqrt{n}}$ or more arbitrary points. This is an example of postulations which will be considered below. It is relating the metric and logic structures and suppositions, although separately studies of these structures are also important. The follow up articles will describe the mixed hierarchy of recognition metric-logic interpretable hypotheses, which helps to allocate classification algorithms to the application problems.

## 2. Structuring by Logic Separation

Let $f$ be a Boolean function (it might be partially or completely defined). Let $N_f$ denotes the reduced disjunctive normal form of $f$ and sets $N_0^f, \ldots, N_3^f$ [2] define areas, in which $N_f$ and $N_{\bar{f}}$ take values {0,1}, {1,0}, {0,0} and {1,1} correspondingly. Identical to $N_0^f, \ldots, N_3^f$, similar areas are defined by logic separation - $M_0^f, \ldots, M_3^f$.

Let $f_0 \in P_2(n)$ (a completely defined Boolean function of $n$ variables) and $f_0(\widetilde{\alpha}) = 1$. Denote by $t(f_0, \widetilde{\alpha})$ the number of k-subcubes included in $N_{f_0}$ and covering the vertex $\widetilde{\alpha}$. Let $m_k$ is the average number of $t(f_0, \widetilde{\alpha})$ calculated for all $f_0 \in P_2(n)$ and $\widetilde{\alpha} \in N_{f_0}$. It is easy to check that $m_k = \dfrac{2^n C_n^k 2^{2^n - 2^k}}{2^n 2^{2^n - 1}} = \dfrac{C_n^k}{2^{2^k - 1}}$.

Dispersion $d_k$ of the same value $t(f_0, \widetilde{\alpha})$ is expressed as $d_k = C_n^k \sum\limits_{j=0}^{k} C_k^j C_{n-k}^{k-j} 2^{-2^{k+1}+2^j+1} - \left( \dfrac{C_n^k}{2^{2^k - 1}} \right)^2$.

Applying the Chebishev inequality to above measures $t(f_0, \widetilde{\alpha})$, $m_k$, $d_k$ leads to the conclusion:

**Proposition 1(8).** $t(f_0, \widetilde{\alpha}) \sim \dfrac{C_n^k}{2^{2^k - 1}}$ for almost all pairs $f_0 \in P_2(n)$ and $\widetilde{\alpha} \in N_{f_0}$, when $n \to \infty$ and $\dfrac{C_n^k}{2^{2^k}} \to \infty$.

Taking into account that for almost all Boolean functions the number of 1-vertices is equivalent to $2^{n-1}, n \to \infty$, we obtain that for almost all functions $f_0 \in P_2(n)$, almost all 1-vertices are covered by the number of k-intervals from $N_{f_0}$, which is equivalent to $\dfrac{C_n^k}{2^{2^k - 1}}$, when $n \to \infty$ and $\dfrac{C_n^k}{2^{2^k}} \to \infty$. Particularly, this fact might be used to adjust the postulation in Proposition 7, [2]. Indeed, the $\dfrac{C_n^k}{2^{2^k - 1}}$ intervals, coming from a common fixed vertex, cover not less than $\dfrac{C_n^k}{2^{2^k - 1}}$ vertices of an n-cube.

Now consider arbitrary placement of any $l$ points into the vertices of an n-cube $M$. Estimate for almost all functions $f_0 \in P_2(n)$ (see Proposition 1(8)) the main value of the number of vertices $\tilde{\alpha} \in N_{f0}$, which are not covered by any of the k-intervals included in $N_{f0}$ which is pricked by our $l$ vertices:

$$\mu(n,k,l) \prec (1+\varepsilon_1(n))2^{n-1}\frac{C^{l}_{2^n-(1+\varepsilon_2(n))C_n^k/2^{2^k-1}}}{C^{l}_{2_n}}, n \to \infty,\ \varepsilon_1(n) \to 0,\ \varepsilon_2(n) \to 0, \text{ and } \frac{C_n^k}{2^{2^k}} \to \infty.$$

**Proposition 2(9)**. If $\frac{C_n^k}{2^{2^k}} \to \infty$ and $l \geq \varphi(n)\frac{2^n 2^{2^k}}{C_n^k}$, where $\varphi(n) \to \infty$ as $n \to \infty$, then random $l$ vertices for almost all functions $f_0 \in P_2(n)$ prick such sets of k-subcubes included in $N_{f0}$, which cover almost all $N_{f0}$.

In case of $k = [\log\log n]$ we conclude that the minimal number $l$ satisfying the above proposition, is not greater than $2^n n^2 / C_n^{[\log\log n]}$.

Notice, that in conditions of Proposition 7 [2] and Proposition 2(9) only the usability of condition $F_0$ (logic separation) is mentioned, so that these are the conditions, when usage of $F_0$, as a rule, doesn't imply to significant errors. Also, it is important, that we applied the condition $F_0$ to the whole class $P_2(n)$, although it was supposed for problems, satisfying compactness suppositions. So, it is interesting to know how completely the class $P_2(n)$ satisfies to these suppositions.

Let us bring now a particular justification of compactness conception [8]. Let $f_0 \in P_2(n)$. We call the vertex $\tilde{\alpha} \in M$ boundary vertex for function $f_0$, if the sphere $S(\tilde{\alpha},1)$ of radius 1 centered at $\tilde{\alpha}$, contains a vertex for which $f_0$ has the opposite value to $f_0(\tilde{\alpha})$. Denote by $\Gamma(f_0)$ the set of all boundary vertices of $f_0$. We will say that the function hipping (completion) procedure obeys the compactness conditions, if $|\Gamma(f_0)| = o(2^n), n \to \infty.$

It is easy to calculate that the average number of boundary vertices of functions $f_0 \in P_2(n)$ is almost $2^n$. This shows that $P_2(n)$ contradicts the compactness conditions. The same time we proved that the use of the $F_0$ rule in a very wide area $P_2(n)$ doesn't move to a sensitive error. Below we consider an example problem, which obeys the compactness assumptions, and will follow the action of the rule $F_0$ on that class. Before that we justify some estimates for the set $M_3^f$.

Consider the class $\Phi_2(n,k(n),l(n))$ of all of partial Boolean functions, for which $|M_0| = l(n)$ and $|M_1| = k(n)$. We'll deal with the case $l(n) = o(2^n)$ and $k(n) = o(2^n)$. Estimate now the quantitative characteristics of sets $M_0^f$, $M_1^f$ and $M_3^f$.

First estimate the average number of vertices of the cube, which are achievable from set $M_0$:

$$C_{03} \geq 2^n \frac{C^{k(n)}_{2^n-l(n)-2^j}}{C^{k(n)}_{2^n-l(n)}}\left(1-\frac{C^{l(n)}_{2^n-\sum_0^j C_n^j}}{C^{l(n)}_{2^n}}\right), j = 0,1,\cdots$$

**Proposition 3(10)**. If $k(n)$ and $l(n)$ are $o(2^n)$, $n \to \infty$ and there exists a $j_0$, that $k(n)2^{j_0-n} \to 0$ and $2^{-n}\sum_0^{j_0} C_n^i l(n) \to \infty$, then for almost all functions of class $\Phi_2(n,k(n),l(n))$, $\left|M_1^f\right| \approx o(2^n), n \to \infty$.

To except the trivial cases in the pattern recognition problems we have to suppose, that $k(n) \cong l(n), n \to \infty$. Then it is clear that choosing appropriate values for $j_0$ we get $\left|M_1^f\right| = o(2^n)$ and $\left|M_0^f\right| = o(2^n)$ for almost all functions of class $\Phi_2(n,k(n),l(n)), n \to \infty$.

Let us give an other estimation of $c_{03}$: $c_{03} \geq \sum_{j=0}^{n} C_n^j \frac{C_{2^n-2^j}^{l(n)}}{C_{2^n-1}^{l(n)}}$.

If $\lambda(n)$ is the minimal value for which $\sum_{i=n/2-\lambda(n)}^{n/2+\lambda(n)} C_n^i \sim 2^n, n \to \infty$, then $\lambda(n) \approx \sqrt{n}$.

From here we conclude:

**Proposition 4(11)**. If $l(n) \geq 0$ and $2^{-n}k^2(n)2^{C(n)\sqrt{n}} \to 0$ as $n \to \infty$ for $\forall c(n)$ - restricted, then almost ever $M_1^f \sim o(2^n)$.

So, for the small values of $k(n)$ and $l(n)$ from the each vertex of set $M_0 \cup M_1$, almost all vertices of the n-unite-cube almost ever are achievable. Comparing this, for example with [5] we find that for these classes $F_0$ works ineffectively.

## 3. Logic Separation on Compact Classes

Consider problems, satisfying the compactness assumptions. First of all it is evident, that for $M_0 \cup M_1 \supseteq \Gamma(f_0)$ the continuation of function $f$ made on base of $F_0$, exactly correspond to the final result $f_0$. Taking into account that by the given description of the compactness assumptions $\Gamma(f_0) = o(2^n), n \to \infty$, we receive that in problems, satisfying the compactness assumptions we can point out learning sets of size $o(2^n), n \to \infty$, which allow to complete and exact continuation of function $f_0$ on base of condition $F_0$ only.

Let $\tilde{\alpha} \in M$ and $0 \leq k_1 \leq k_2 \leq n$. Consider functions $f_0 \in P_2(n)$, for which $M_0(f_0) \supseteq S(\tilde{\overline{\alpha}}, n-k_2)$, $M_1(f_0) \supseteq S(\tilde{\alpha}, n-k_1)$ and which receive arbitrary values on vertices of sets $S(\tilde{\alpha}, k_2-1) \supseteq S(\tilde{\alpha}, k_1)$.

Denote the class of these functions by $K(n)$. It is evident, that for $\left|S(\tilde{\alpha}, k_2-1) \setminus S(\tilde{\alpha}, k_1)\right| = o(2^n)$ all the constructed functions satisfy the given formalisms for the compactness assumptions, and that the quantity of these functions is not less than $2^{\varepsilon_1(n)2^n}$, where $\varepsilon_1(n)$ is an arbitrary function of $n, \varepsilon_1(n) \to 0$ with the $n \to \infty$.

Take a point $\tilde{\beta} \in M, \rho(\tilde{\alpha}, \tilde{\beta}) = k, k < k_1$. It is evident that no more than $C_n^{[n/2]} \cong 2^n\sqrt{\frac{2}{\pi n}}, n \to \infty$ subsets of any fixed size are coming out from any point of $n$-cube. From the other hand it is evident, that it is enough to take $k_1 - k = o(\sqrt{n})$ as $n \to \infty$ to get the $\left|S(\tilde{\alpha}, k_1) \setminus S(\tilde{\alpha}, k)\right| = o(2^n)$. Suppose, that $\left|M_0(f_0) \cup M_1(f_0)\right| = l$, and that $l$ points appear as the result of their appropriate placement on the vertices of the $n$-cube $M$, when all of these placements are equally probable. Estimate the probability of reaching of this point $\widetilde{\beta}$ from zeros of function $f_0$.

$$\tau_l \le 2^n \sqrt{\frac{2}{\pi n}} \frac{C_{2^n - 2^{\varepsilon 2(n)\sqrt{n}}}}{C_{2^n}^l}, \varepsilon_{2(n)} \to 0 \text{ with } n \to \infty .$$

From here we conclude:

**Proposition 5(12)**. Let $f_0 \in K(n,k_1,k_2)$ and $f_1$ -- is the continuation of function $f$ on base of condition $F_0$. If $l \ge n2^{n-\varepsilon_2(n)\sqrt{n}} = o(2^n)$, $n \to \infty$ and the set $M_0(f) \cup M_1(f)$ is formed as a random collection of points of size $l$ from the set $M$, then almost ever the function $f_1$ is the continuation for $f$, which converges to the $f_0$ by the accuracy, tending to $1$ with the $n \to \infty$.

## Conclusion

Logic Separation is an alternative approach to pattern recognition hypotheses and formalisms, while the base concept uses the similarity approach. Structures appearing in this relation are based on terms of Reduced Disjunctive Normal Forms of Boolean Functions. Propositions 1-5(8-12) provide additional knowledge on quantitative properties of areas appearing in extending classification by means of compactness and logic separation principles.

## Bibliography

1. Zhuravlev Yu. I. On an algorithmic approach to the problems of recognition and classification. Problemi Kibernetiki, 33, (1978) 5--68.
2. Aslanyan L. and Castellanos J., Logic based pattern recognition – ontology content (1), iTECH-06, 20-25 June 2006, Varna, Bulgaria, Proceedings, pp. 61-66.
3. Gavrilov G. and Sapojenko A., Collection of exersises of discrete mathematics, NAUKA, Moscow, 1973, 368 p.
4. Herman G.T. and Kuba A., editors. Discrete Tomography: Foundations, Algorithms and Applications. Birkhauser, 1999.
5. Zhuravlev Yu. I. Selected Scientific Works, Publishing House Magister, Moscow, (1998) 417p.
6. Aslanyan L. H. On a pattern recognition method based on the separation by the disjunctive normal forms. Kibernetika, 5, (1975), 103--110.
7. Vapnik V. and Chervonenkis A. Theory of Pattern Recognition. "Nauka", 1974.
8. Aslanyan L. H. The Discrete Isoperimetric Problem and Related Extremal Problems for Discrete Spaces. Problemy Kibernetiki, 36, (1979), 85--128.
9. Nechiporuk E. I. On topological principles of self-correcting. Problemy Kibernetiki, 21, (1969), 5--102.
10. Graham N., Harary F., Livingston M. and Stout Q. Subcube Fault-Tolerance in Hypercubes. Information and Computation 102 (1993), pp. 280{314.
11. Glagolev V. V. Some Estimations of D.N.F. for functions of Algebra of Logic. Problemy Kibernetiki, 19, (1967), 75--94.

## Authors' Information

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

***Vladimir Ryazanov -*** *Computing Centre of the Russian Academy of Sciences, 40 Vavilova St., Moscow, GSP-1, 119991, Russian Federation, rvvccas@mail.ru*

# ON STRUCTURAL RESOURCE OF MONOTONE RECOGNITION[1]

Hasmik Sahakyan, Levon Aslanyan

***Abstract**: Algorithmic resources are considered for elaboration and identification of monotone functions and some alternate structures are brought, which are more explicit in sense of structure and quantities and which can serve as elements of practical identification algorithms. General monotone recognition is considered on multi-dimensional grid structure. Particular reconstructing problem is reduced to the monotone recognition through the multi-dimensional grid partitioning into the set of binary cubes.*

## 1. Introduction

Monotone Boolean functions have an important role in research area since they arise in various application models, such as design of schemes, pattern recognition, etc.

Monotone Boolean functions are studied in different viewpoint and are known as objects of high complexity. First results, obtained by Mickeev [M, 1959] and Korobkov [K, 1965], characterize Sperner families in unit cube. After enormous investigations and overcoming difficulties, Korshunov [K, 1981] obtained the asymptotical estimate of the number of Monotone Boolean functions. It is characteristic that analytical formulas are not known at this point.

Another cluster of research work solves problems of algorithmic identification of monotone Boolean functions. Hansel [H, ] constructed the best algorithm in sense of Shannon criterion, then Tonoyan [T, 1979] constructed a similar algorithm with minimal use of memory. Later on there obtained some generalizations for multi-valued cube. Alekseev [A, 1976] generalized Hansel's result, Katerinochkina [K, 1978] gave precise description of structure of Sperner families.

It is typical that for multi-valued cube there is no explicit formula not only for the number of monotone functions, but also for the cardinality of middle layer. It makes difficult choice of algorithms for a concrete problem and estimation of their complexity.

Below in this paper some algorithmic resources are considered for elaboration and identification of grid defined monotone functions and some alternate structures are brought, which are more explicit in sense of structure and quantities and which can serve as elements of practical identification algorithms.

## 2. Learning Monotone Functions on Multi-valued Cube

Let $\Xi_{m+1}^n$ denotes the grid of vertices of $n$ dimensional, $m+1$ valued cube, i.e. the set of all integer-valued vectors $S=(s_1,s_2,\cdots,s_n)$ with $0 \le s_i \le m$, $i=1,\cdots,n$. For any two vertices $S^{'}=(s_1^{'},s_2^{'},\cdots,s_n^{'})$ and $S^{''}=(s_1^{''},s_2^{''},\cdots,s_n^{''})$ of $\Xi_{m+1}^n$ we say that $S^{'}$ is greater than $S^{''}$, $S^{'} \ge S^{''}$ if $s_i^{'} \ge s_i^{''}$, $i=1,\cdots,n$. We call pair of vectors $S^{'}$, $S^{''}$ comparable if $S^{'} \ge S^{''}$ or $S^{'} \le S^{''}$, otherwise these vectors are incomparable. Set of pair wise incomparable vectors composes a Sperner family.

Usually vertices of $\Xi_{m+1}^n$ are placed schematically among the $m \cdot n+1$ layers of $\Xi_{m+1}^n$ according to their weights, – sums of all coordinates. Vector $\widetilde{0}=(0,\cdots,0)$ is located on the 0-th layer; then the $i$-th layer consists of all vectors, with the weight $i$. An element of $i$-th layer might be greater than some vector from the $i-1$-th layer, exactly by one component and exactly by one unit of value (such vector pairs are called neighbors and are connected by an edge). The vector $\widetilde{m}=(m,\cdots,m)$ is located on the $m \cdot n$-th layer.

[1] The research is supported partly by INTAS: 04-77-7173 project, http://www.intas.be

Consider a binary function $f$ on $\Xi_{m+1}^{n}$, $f : \Xi_{m+1}^{n} \rightarrow \{0,1\}$. We say that $f$ is monotone if for any two vertices $S^{'}, S^{''}$ notion $S^{'} \geq S^{''}$ implies $f(S^{'}) \geq f(S^{''})$. The vector $S^{1} \in \Xi_{m+1}^{n}$ is a lower unit of monotone function $f$ if $f(S^{1}) = 1$ and for arbitrary $S \in \Xi_{m+1}^{n}$, such that $S < S^{1}$ it is true that $f(S) = 0$. The vector $S^{0} \in \Xi_{m+1}^{n}$ is an upper zero of monotone function $f$ if $f(S^{0}) = 0$ and for arbitrary $S \in \Xi_{m+1}^{n}$, such that $S > S^{0}$ it is true that $f(S) = 1$.

Above defined monotone functions are known also as increasing monotone in contrast with a decreasing monotone function. A function $f$ is decreasing monotone if for any two vertices $S^{'}, S^{''}$, $S^{'} \geq S^{''}$ implies $f(S^{'}) \leq f(S^{''})$. For $f$ decreasing monotone, the vector $S^{1} \in \Xi_{m+1}^{n}$ is an upper unit if $f(S^{1}) = 1$ and for any $S \in \Xi_{m+1}^{n}$, such that $S > S^{1}$ we get $f(S) = 0$. The vector $S^{0} \in \Xi_{m+1}^{n}$ is a lower zero of function $f$ if $f(S^{0}) = 0$ and for any $S \in \Xi_{m+1}^{n}$, such that $S < S^{0}$ we get $f(S) = 1$.

When $m = 1$ the definitions above lead to ordinary monotone Boolean functions defined on binary cube $E^{n}$.

Let a monotone function $f$ be defined with the help of an oracle which, receiving any vector $S \in \Xi_{m+1}^{n}$, gives the value $f(S)$. The problem is in identification of arbitrary monotone function $f$ by as far as possible small number of accesses to the oracle. Similar problems are interested in finding all or the maximal/minimal upper zeros or alternatively the minimal/maximal lower 1's of the given Boolean function. Consider an example. Let, it is given a set of $n$ linear inequalities. A consistent subset of inequalities is coded by a vertex of $E^{n}$, where we define $f$ as 0. The problem of finding the maximal consistent subset of inequalities is a known hard problem and the use of oracle reduces the problem to solving several subsystems of inequalities, which is just an alternative way of solving the main problem. The monotone binary function recognition on $\Xi_{m+1}^{n}$ is the weighted inequalities version of the above given example model.

In [A, 1976] an algorithm of complexity $\leq |M| + \lfloor log_{2} m \rfloor \cdot |N|$ is constructed to learn the binary monotone functions above the multi-valued discrete grid, which generalizes the Hansel's method ([H, ]) for the case of monotone Boolean functions, here $M$ and $N$ denote the sets of vertices of middle layers of multi-valued grid/cube, i.e. layers which contain vectors with sums of coordinates equal to $\lfloor (m \cdot n)/2 \rfloor$ and $\lfloor (m \cdot n)/2 \rfloor + 1$ respectively. It is also proven that the complexity of the algorithm is approximately $\sqrt{n}$ time less than the whole number of vertices of the grid.

## 3. $\Xi_{m+1}^{n}$ partitioning through binary cubes

In this section an alternate approach to traditional means is considered for identification of monotone functions defined on $\Xi_{m+1}^{n}$. First $\Xi_{m+1}^{n}$ is partitioned into binary cube like structures and then Hansel's method is applied for identification of monotone Boolean functions. This approach may serve as a separate element of practical identification algorithms.

In $\Xi_{m+1}^{n}$ we distinguish several classes of vectors.

**Upper and lower homogeneous vectors.** A vector of $\Xi_{m+1}^{n}$ is called an upper $h$-vector (upper homogeneous) if the values of all its coordinates are at least $m/2$ for even $m$, and are at least $(m+1)/2$ for odd $m$. Similarly, a vector of $\Xi_{m+1}^{n}$ is called a lower $h$-vector (lower homogeneous) if the values of all its coordinates are at most $m/2$ for even $m$, and are at most $(m-1)/2$ for odd $m$.

We denote by $\widehat{H}$ the set of all upper $h$-vectors and by $\breve{H}$ the set of all lower $h$-vectors. The cardinalities of sets $\widehat{H}$ and $\breve{H}$ are equal to $((m+1)/2)^n$ for odd $m$ and to $(m/2+1)^n$ - for $m$ even.

**Middle vectors** $\tilde{m}_{mid+}$ and $\tilde{m}_{mid-}$. $\tilde{m}_{mid+} = ((m+1)/2,\cdots,(m+1)/2)$ and $\tilde{m}_{mid-} = ((m-1)/2,\cdots,(m-1)/2)$ for odd $m$ and $\tilde{m}_{mid+} = \tilde{m}_{mid-} = (m/2,\cdots,m/2)$ for even $m$.

$\tilde{m}_{mid+}$ is located on the $n\cdot(m+1)/2$-th layer of $\Xi^n_{m+1}$ (the lowest layer that contains vector from $\widehat{H}$) and $\tilde{m}_{mid-}$ is located on the $n\cdot(m-1)/2$-th layer of $\Xi^n_{m+1}$ (the highest layer that contains vector from $\breve{H}$) for odd $m$; for even $m$ the vector $\tilde{m}_{mid+} = \tilde{m}_{mid-}$ is located on the layer $n\cdot m/2$ and this is the only common vector of $\widehat{H}$ and $\breve{H}$.

**Vertical equivalent vectors.** Let $S^{'},S^{''} \in \Xi^n_{m+1}$. $S^{'}$ and $S^{''}$ are called $v$-equivalent (vertically equivalent) if one of them is obtained from the other by inverting some coordinates (that is replacing some coordinates by their complements up to the $m$).

For a given vector $S$ denote by $V(S)$ the class of all $v$-equivalent vectors to $S$ and call it the $v$-equivalency class of $S$. This structure $V(S)$ is congruent to a cube $E^k$, where $k$ is the number of coordinates of $S$ not equal to $m/2$ (this is valid for even $m$). For odd $m$ $k=n$. It is also evident, that $V(S^{'})=V(S)$ for an arbitrary $S^{'} \in V(S)$.

In $V(S)$ we distinguish two vectors $\widehat{S}=(\widehat{s}_1,\cdots,\widehat{s}_n)$ and $\breve{S}=(\breve{s}_1,\cdots,\breve{s}_n)$ - upper and lower vectors, which coordinates are defined as follows:

$$\widehat{s}_i = \begin{cases} s_i, & s_i \ge m-s_i \\ m-s_i, & s_i < m-s_i \end{cases} \text{ and } \breve{s}_i = \begin{cases} m-s_i, & s_i \ge m-s_i \\ s_i, & s_i < m-s_i \end{cases}, \quad i \in \overline{1,n}.$$

These are the only vectors of $V(S)$ that belong to sets $\widehat{H}$ and $\breve{H}$ respectively.

Consequently, for any $S$ the class of its $v$-equivalency can be constructed by the upper vector $\widehat{S}$ and/or by the lower vector $\breve{S}$ of that class by coordinate inversions. $v$-equivalency classes of different upper homogeneous vectors are none intersecting.

This proves partitioning of the whole structure $\Xi^n_{m+1}$ through binary cube like vertical extensions of elements of $\widehat{H}$ or $\breve{H}$. The following formula shows the picture of factorization of structure of $\Xi^n_{m+1}$ through these cubical elements: $(m+1)^n = \sum\limits_{k=0}^{n}\left(C_n^k \cdot 2^k \cdot (m/2)^k\right) = \sum\limits_{k=0}^{n}\left(C_n^k \cdot m^k\right)$ for even $m$ and $(m+1)^n = ((m+1)/2)^n \cdot 2^n$ for odd $m$.

Thus, we get $\left|\widehat{H}\right|$ disjoint subsets, congruent to binary cubes, which cover $\Xi^n_{m+1}$. Notice that if we construct the corresponding binary cubes, then a pair of vertices, comparable in a binary cube, is comparable also in $\Xi^n_{m+1}$. Therefore monotonicity in $\Xi^n_{m+1}$ implies monotonicity in all received binary cubes and starting by a monotone function in $\Xi^n_{m+1}$ and reconstructing the implied functions on cubes the initial function will be reconstructed in a unique way.

We recall now the problem of identification of monotone binary functions defined on $\Xi^n_{m+1}$.

By Hansel's result [H, 1966] an arbitrary monotone Boolean function with $k$ variables can be identified by $C_k^{\lfloor k/2 \rfloor} + C_k^{\lfloor k/2 \rfloor + 1}$ accesses to the oracle.

Hence an arbitrary monotone function defined on $\Xi_{m+1}^{n}$ can be identified by $\sum_{k=0}^{n}\left(C_n^k\cdot(m/2)^k\cdot\left(C_k^{\lfloor k/2\rfloor}+C_k^{\lfloor k/2\rfloor+1}\right)\right)$ accesses for even $m$ and by $((m+1)/2)^n\cdot\left(C_n^{\lfloor n/2\rfloor}+C_n^{\lfloor n/2\rfloor+1}\right)$ - for odd $m$.

## 4. Characteristic Vectors of Subsets Partitions of $E^n$ and Identification of Monotone functions in $\widehat{H}$

For a given $m$, $0\le m\le 2^n$ let $\psi_m$ denote the set of all **characteristic vectors of partitions** of $m$-subsets of $E^n$. A nonnegative integer-valued vector $S=(s_1,s_2,\cdots,s_n)$ is called characteristic vector of partitions of a vertex subset $M$, $M\subseteq E^n$ if its coordinates are the sizes of partition-subsets of $M$ by coordinates $x_1,x_2,\cdots,x_n$, which are the Boolean variables composing $E^n$. $s_i$ is the size of one of partition-subsets of $M$ in the $i$-th direction and $m-s_i$ is the complementary part of partition. For simplicity we will later assume that $s_i$ is the size of the partition with $x_i=1$.

If $m\neq 0$ then $\psi_m$ is not empty. It is also evident that $\psi_m\subseteq\Xi_{m+1}^{n}$. As other exceptions distinguish between the 2 boundary cases: if $m=1$ then $\psi_m=\Xi_{m+1}^{n}$ and so $|\psi_m|=\left|\Xi_{m+1}^{n}\right|=2^n$; if $m=2^n$ then $|\psi_m|=1$ and the vector with all coordinates $2^{n-1}$ indeed belongs to $\Xi_{m+1}^{n}$.

In [S, 2006] the entire description of $\psi_m$ is given in terms of $\Xi_{m+1}^{n}$ geometry. It is particularly proven that the main problem of describing characteristic vectors can be moved from the $\Xi_{m+1}^{n}$ to the area of $\widehat{H}$ ($\breve{H}$), where the vector set $\psi_m$ has monotonous structure, – it corresponds to the units of some monotone decreasing binary function defined on $\widehat{H}$ (monotone increasing binary function defined on $\breve{H}$).

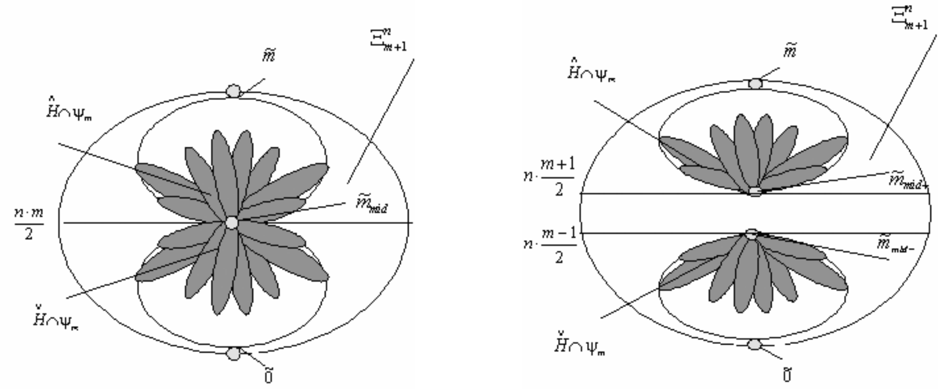


Figure 1

Figure 1 illustrates the sets $\widehat{H}\cap\psi_m$ and $\breve{H}\cap\psi_m$ for even and odd $m$ values, correspondingly.

Thus for entire description of $\psi_m$ it is sufficient to consider all monotone functions defined on $\widehat{H}$ or $\breve{H}$. We shall restrict ourselves to the consideration of decreasing monotone functions defined on $\widehat{H}$. Let $\widehat{\psi}_m$ be the subset of $\widehat{H}\cap\psi_m$ consisting of all upper units of corresponding monotone function.

In [AS, 2001] additional resource is introduced: $L_{min}$ and $L_{max}$, - minimal and maximal numbers of layers of $\widehat{H}$, - are calculated, such that all vectors of $\widehat{\psi}_m$ are located between them. It importantly follows that the entire description of $\psi_m$ is reduced to the identification of monotone functions with upper units between the layers $L_{min}$ and $L_{max}$.

Summarizing all the above consideration we come to the conclusion:

1) Algorithmic resource of learning monotone binary functions defined on $\Xi_{m+1}^n$ includes structures such as:

- generalized Hansel's method and constructions, for monotone binary functions defined on multi-valued cube,
- $\Xi_{m+1}^n$ partitioning through binary cube like vertical extensions of the elements of $\widehat{H}$ together with applying Hansel's result for monotone Boolean functions defined on that cubes,

2) For the entire description of $\psi_m$ we reduce the problem to $\widehat{H}$ becoming able to possess with additional resources:

- learning monotone binary functions defined on $\widehat{H}$ by means of generalized Hansel's method,
- partitioning $\widehat{H}$ into binary cube like vertical extensions of its upper homogeneous elements and applying Hansel's method for them,
- identifying monotone functions defined on $\widehat{H}$ with use of additional information on location of their upper units through $L_{min}$ and $L_{max}$.

The choice of concrete resource set depends on requirements of certain applications.

## Conclusion

Algorithmic resources are considered for elaboration and identification of monotone functions. Current research proposes two new components - partitioning the multi-valued cube through binary cube like vertical extensions of its upper homogeneous elements; and learning upper homogeneous area through the analogous partitioning. The choice of concrete resource depends on requirements of certain application.

## Bibliography

[M, 1959] V. Mickeev. On sets, containing maximal number of pair wise incomparable Boolean vectors, Prob.Cyb, 2, 1959.

[K, 1965] V. Korobkov. On monotone functions of algebra of logic, Prob.Cyb, 13, 1965.

[K, 1981] A. Korshunov. On the number of monotone Boolean functions, Prob.Cyb, 38, 1981.

[H, 1966] G. Hansel. Sur le nombre des fonctiones booleennes monotones de n variables, C. R. Acad. Sci., Paris, 262, N% 20, 1966.

[T, 1979] G. Tonoyan. Cain partitioning of n-cube vertces and deciphering of monotone Boolean functions, Journal of Computational Mathematics and Math. Physics, 1979, vol. 19, N.6.

[A, 1976] V. Alexeev. On deciphering of some classes of monotone many valued functions, Journal of Computational Mathematics and Math. Physics, 1976, vol. 16, N.1.

[K, 1978] Katerinochkina N., On sets, containing maximal number of pair wise incomparable n-dimensional k-valued vectors, Mathematical notes, v. 24, no, 3, Sept. 1978.

[AS, 2001] L. Aslanyan, H. Sahakyan. On the boundary cases of partitioning of subsets of the n-dimensional unit cube, "Computer Science & Information Technologies" Conference, Yerevan, September 17-20, 2001

[S, 2006] H. Sahakyan. Numerical characterization on n-cube subset partitioning, submitted to Discrete Applied Mathematics.

## Authors' Information

***Hasmik Sahakyan*** *– Institute for Informatics and Automation Problems, NAS Armenia, P.Sevak St. 1, Yerevan-14, Armenia; e-mail: hasmik@ipia.sci.am*

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

# CROSSOVER OPERATOR IN DNA SIMULATION OF GENETIC ALGORITHMS

Angel Goñi Moreno

***Abstract:*** *In this paper a crossover operator to use in the simulation of genetic algorithms with DNA is presented. The aim of the paper is to follow the path of creating a new computational model based on DNA molecules and genetic operations. It shows also the using of this operator applied to the travelling salesman problem (TSP). After individual encoding and fitness evaluation, a protocol of the next step in a genetic algorithm, crossover, is needed. The simulation of GA using DNA will resolve the problem of exponentially size algorithms like the ones that were proposed in the beginning of DNA computing.*

***Keywords:*** *DNA Computing, Evolutionary Computing, Genetic Algorithms.*

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

## Introduction

In a short period of time DNA based computations have shown lots of advantages compared with electronic computers. DNA computers can 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]. However, 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.

On the other hand there are genetic algorithms (or short GA) which are categorized as global search heuristics and use techniques inspired by evolutionary biology [Holland, 1975]. It seems to be perfect to combine DNA computing and GAs.

Previous work on molecular computation for genetic algorithms [J.Castellanos, 1998] show the possibility of solving optimization problems without generating or exploring the complete search space and give a solution to the first step to be done in a GA, the coding of the population and the evaluation of individuals (fitness). A recent work [M.Calviño, 2006] produced a new approach to the problem of fitness evaluation saying 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].

In this paper the crossover (also called recombination) of DNA-strands has been resolved satisfactorily by making a crossover operator suitable for DNA computing and its primitive operations. This crossover operator is used in the simulation of the travelling salesman problem (TSP) with both genetic algorithm and DNA computing continuing the work previously done about the coding of information. The Lipton [Lipton, 1995] encoding is used to obtain each individual coded by a sequence of zeros and ones, and when using DNA strands this information is translated into the four different bases that are presented in DNA – adenine (A), thymine (T), cytosine (C), and guanine (G).

## Molecular Computing

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.

Here a short description of the tool box of techniques for manipulating DNA is provided so that the reader can have a clear intuition about the nature of the techniques involved.

- Strands separation:
    - Denaturation of DNA strands. Denaturation of DNA is usually achieved by heat treatment or high pH, which causes the double-stranded helix to dissociate into single strands.
    - According to their length using gel electrophoresis. This technique is used to push or pull the molecules through a gel matrix by applying an electric current. The molecules will move through the matrix at different rates depending on their size.
    - According to a determinated subchain using complementary probes anchored to magnetic beads.
- Strands fusing:
    - Renaturation. If the soup is cooled down again, the separated strands fuse again.
    - Hybridization. Originally it was used for describing the complementary base pairing of single strands of different origin (e.g., DNA with RNA).
- Cutting DNA. Using restriction enzymes which destroy internal phosphodiester bonds in the DNA.
- Linking (pasting) DNA. Molecules can be linked together by certain enzymes called ligases.
- PCR mutagenesis. To incorporate the primer as the new (mutant) sequence.

## Genetic Algorithms

Genetic Algorithms 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 [Holland, 1975] was the first one to study an algorithm based on an analogy with the genetic structure and behaviour of chromosomes. Genetic algorithms has been widely studied and experimented. 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 individuals, (3) cross and mutate selected individuals, (4) evaluate and introduce the new created individuals in the initial population. In that way, the successive generation will become more suited to their environment.

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.

## Crossover

As it has already been said the first thing to do when a problem is presented is the codification of individuals. In our case the problem is TSP. How can we code the paths? A possible solution was provided in a previous work [J.Castellanos, 1998] giving a representation of individuals like a DNA-strand for each path. This encoding is based on a sequence of genes each one represents an arch between two cities. Here the fitness would be an extra field placed at the beginning of the DNA-strand and its length is proportional to the value it represents (in fact it depends on the problem. For example in the travelling salesman problem, TSP, the length should be inversely proportional to the value of the path). Between the DNA code belonging to the genes a cutting site for a restriction enzyme will be inserted. The final encoding for a path is:

PCR Primer Np REp **Fitness** RE1 gene REn-1 …… RE0 gene REp Np PCR Primer

A recent work approached individual encoding by eliminating the field fitness. In that case, the fitness is embedded in the genes. The advantage of this work is that when all the individuals of the population are generated, there is no need to evaluate them because they have already been evaluated by themselves. After solving the problem of the selection by adding a specific field in each gene which tells the distance between both cities, it is necessary to see if the same format of the strands is valid in the next step of de GA, crossover.

First of all let's try to solve this step using the technique "cut and splice" like it is done in vivo. A single cut point is selected and after cut we splice both ends. An example is shown in figure 1 with two different chromosomes.

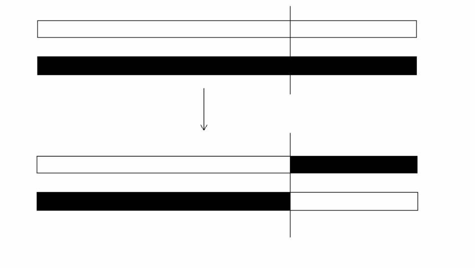
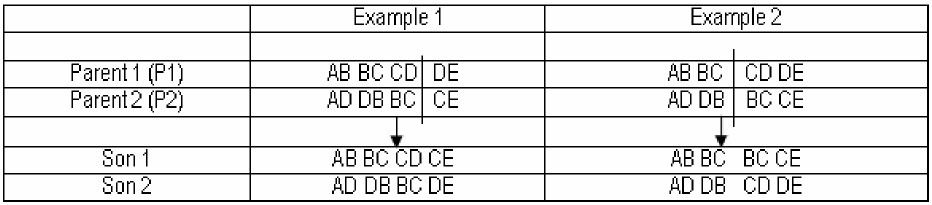


Fig 1

Solving TSP crossover with cut and splice:



| | Example 1 | Example 2 |
|---|---|---|
| | | |
| Parent 1 (P1) | AB BC CD DE | AB BC CD DE |
| Parent 2 (P2) | AD DB BC CE | AD DB BC CE |
| | ↓ | ↓ |
| Son 1 | AB BC CD CE | AB BC BC CE |
| Son 2 | AD DB BC DE | AD DB CD DE |

Table 1

The results show that this method must be discarded because all the sons it produces are invalid. Obviously, it has no sense to create a son that contains a specific city twice.

I proceed then to apply a different protocol called "order crossover" (OX). Adjacency information in the total ordering is important and this crossover preserves relative ordering. Two parents are selected between the population then a random mask is selected (with 0's and 1's which are chosen randomly with the same probability both bits). During the first step, the sons are filled with the genes of the parents which the mask allows. To complete the sons, we put the genes missing in S1 in the order they appear in P2 and the same with S2. An example is shown in Table 2.

| | |
|---|---|
| P1 | A C D B E |
| P2 | A D B C E |
| Mask | 1 0 1 0 1 |
| S1 (C, B missing) | A - D - E |
| S2 (D, C missing) | A - B - E |
| S1 (B before C in P2) | A B D C E |
| S2 (C before D in P1) | A C B D E |

Table 2

## Suitable mask for order crossover

Let's try to apply OX to TSP. It is remembered that each gene represents an arch between two different cities. Like a first attempt it is used the mask: 1001 (one bit for gene).

As we see in Table 3, the result of the computation of OX using that mask is invalid because the genes CD and DB do not even exist in P2 so S1 cannot be completed. By using this mask we will never get valid individuals. Now, OX is computed with the mask 10 01 01 11, using two bits for each gene. Each bit represents a city. As usually, the mask is chosen randomly (every bit).

| | |
|---|---|
| P1 | AC CD DB BE |
| P2 | AD DB BC CE |
| Mask | 1 0 0 1 |
| S1 (missing CD, DB) | AC - - BE |

Table 3

| | |
|---|---|
| P1 | AC CD DB BE |
| P2 | AD DB BC CE |
| Mask | 10 01 01 11 |
| S1 (missing C(twice), D) | A- -D -B BE |

Table 4

Once again the mask is not correct. In the example (Table 4) we can see how in the third gene of P1 (DB) there is missing only one city, city D. That has no sense at all, because the second city of the second gene must be the same as the first city of the third gene. That give us the idea of how the definitive mask should be. Let's try now with the mask 10 01 10 01. In this mask we choose randomly the pair of bits that represents the same city, for example we choose if the second bit of the first gene (1**0**) and the first bit of the second gene (**0**1) are 0 or 1 both of them but not different.

In this example (Table 5) the sons are correct. So that is the suitable mask. In order to find less invalid individuals we force the mask to one last rule: the first bit and the last must be 1 because in TSP the first city we visit and the last one must be always the same.

| | |
|---|---|
| P1 | AC CD DB BE |
| P2 | AD DB BC CE |
| Mask | 10 01 10 01 |
| S1 (C, B missing) | A - -D D- -E |
| S2 (D, C missing) | A - -B B- -E |

| | |
|---|---|
| S1 (B before C in P2) | AB BD DC CE |
| S2 (C before D in P1) | AC CB BD DE |

Table 5

## Translating order crossover to DNA computing

How can be translated into DNA computing the previous crossover operator? Firstly, imagine that we have in a test tube the individuals that we had before but in the encoding which is explained above, representing each individual like a sequence of genes in a DNA strand. That is shown in Fig 2.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P1 (AC CD DB BE) | | | | | | | | | | | | |
| ... | A | C | RE | C | D | RE | D | B | RE | B | E | … |
| P2 (AD DB BC CE) | | | | | | | | | | | | |
| … | A | D | RE | D | B | RE | B | C | RE | C | E | … |

Fig 2

When the problem of representing the mask is tackled a different view of the mask is given. This mask is suitable for our problem (TSP) using DNA computing and is obtained by following these steps:

1. Imagine that in our problem we have 5 cities: A B C D E.
2. To be discarded are the initial city and the final city. Then we have: B C D.
3. Randomly we choose one or several cities. For example: C D
4. Introduce in the soup (into the test tube) the following strands:

$\overline{C}$
(*)
$\overline{C}$
$\overline{D}$
(*)
$\overline{D}$

This represents the complementary bases of the cities C and D so that when introduced in the soup they can match the original strands there were in the soup before. See Fig 3.

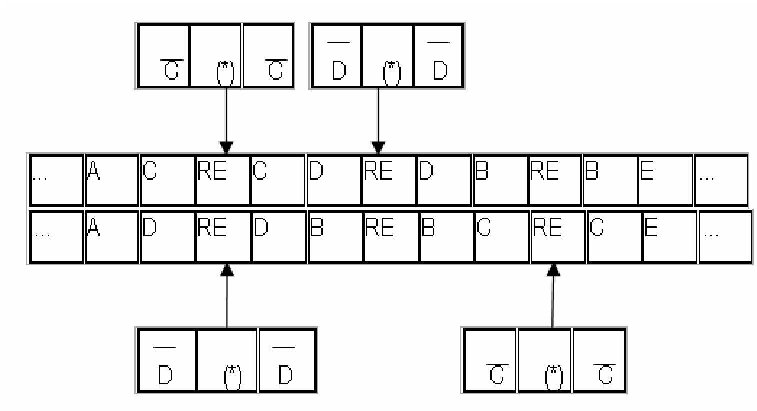


Fig 3

| P1 | AC CD DB BE |
|---|---|
| P2 | AD DB BC CE |
| Mask | C and D |
| S1 (C, D missing) | A - -- -B BE |
| S2 (D, C missing) | A - -B B- -E |

| S1 (D before C in P2) | AD DC CB BE |
|---|---|
| S2 (C before D in P1) | AC CB BD DE |

Table 6

Computing crossover of Fig 3 (see Table 6).

As a result of all this steps the crossover operator has changed a lot. Now it is still order crossover but with a very particular way of choosing the genes that must be changed. Instead of the initial mask we saw in Table 2, the mask it is used now consists on a selection of which cities (not genes) of P1 must be changed in the order they are found in P2. In the example shown in Table 5 the mask (10 01 10 01) showed the position of the cities that should be changed by the crossing operator. Now the mask for the same example (C, D) doesn't tell the position but de name of the cities to change.

However, if we try to apply this crossover operator in a genetic algorithm which uses the individual encoding that M.Calviño presented [M.Calviño, 2006] there is a big problem found. Spouse we have the gene AFC, in which F means the fitness between cities A and C. If we try to carry out the crossover operation of Fig 3, city C must be changed by D and then the gene would be AFD. Obviously F is not the fitness between A and D so this crossover operator only works with the strand-format proposed by J.Castellanos [J.Castellanos, 1998] though the other one works much better in the previous step of the GA, evaluation and selection.

## Conclusion

The problem of crossover in a genetic algorithm using DNA has been resolved satisfactorily. Although the crossover technique might be different depending on the problem to be solved, it has been proved that it is possible to find a suitable crossover for NP-Complete problems such as TSP. This represents a new approach to the simulation of genetic algorithms with DNA.

Since the beginning of DNA computing, the lack of algorithms to be applied to this scientific area has been very large. Until recently, molecular computation has used “brute force” to solve NP-Complete problems. That is why the simulation of concepts of genetic evolution with DNA will help DNA computing to resolve hard computations. The crossover operator I have presented here give an idea of how important and useful genetic algorithms are for DNA computing.

## 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.

[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.

## Authors’ Information

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

# USING A QUERY EXPANSION TECHNIQUE TO IMPROVE DOCUMENT RETRIEVAL

## Abdelmgeid Amin Aly

***Abstract**: Query expansion (QE) is a potentially useful technique to help searchers formulate improved query statements, and ultimately retrieve better search results. The objective of our query expansion technique is to find a suitable additional term. Two query expansion methods are applied in sequence to reformulate the query. Experiments on test collections show that the retrieval effectiveness is considerably higher when the query expansion technique is applied.*

## 1. Introduction

Since the 1940s the problem of Information Retrieval (IR) has attracted increasing attention, especially because of the dramatically growing availability of documents. IR is the process of determining relevant documents from a collection of documents, based on a query presented by the user.

There are many IR systems based on Boolean, vector, and probabilistic models. All of them use their model to describe documents, queries, and algorithms to compute relevance between user's query and documents. Each model contains some constraints, which cause disproportion between expected (relevant) documents and documents returned by IR system. One of the possibilities (how to solve the disproportion) is systems for automatic query expansion, and topic development observing systems. In this respect, query expansion aims to reduce this query/document mismatch by expanding the query using highly "correlated" to the query terms, words or phrases with a similar meaning or some other statistical relation. To detect such correlations between terms, different based-statistical-measures approaches, requiring the analysis of the entire document collection, have been introduced, e.g., term Co-occurrence measures or lexical co-occurrence measures [1, 2]. Query expansion (or term expansion) is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance. The method itself is applicable to any situation irrespective of the retrieval technique(s) used. The initial query (as provided by the user) may be an inadequate or incomplete representation of the user's information need, either in itself or in relation to the representation of ideas in documents.

There are three types of QE: manual, automatic, and interactive. Manual QE takes place when the user refines the query by adding or deleting search terms without the assistance of the IR system. New search terms may be identified by reviewing previous retrieval results, communication with librarians or colleagues; other related documents, or a general vocabulary tool are not specific to the IR system (e.g., a dictionary or standard thesaurus) [3]. Decisions about the association of terms are up to the users themselves and are dependent on the expertise of the users with the search system and features [4].

Query expansion involves adding new words and phrases to the existing search terms to generate an expanded query. However, previous query expansion methods have been limited in extracting expansion terms from a subset of documents, but have not exploited the accumulated information on user interactions. We believe that the latter is extremely useful for adapting a search method to the users. In particular, we will be able to find out what queries have been used to retrieve what documents, and from that, to extract strong relationships between query terms and document terms and to use them in query expansion.

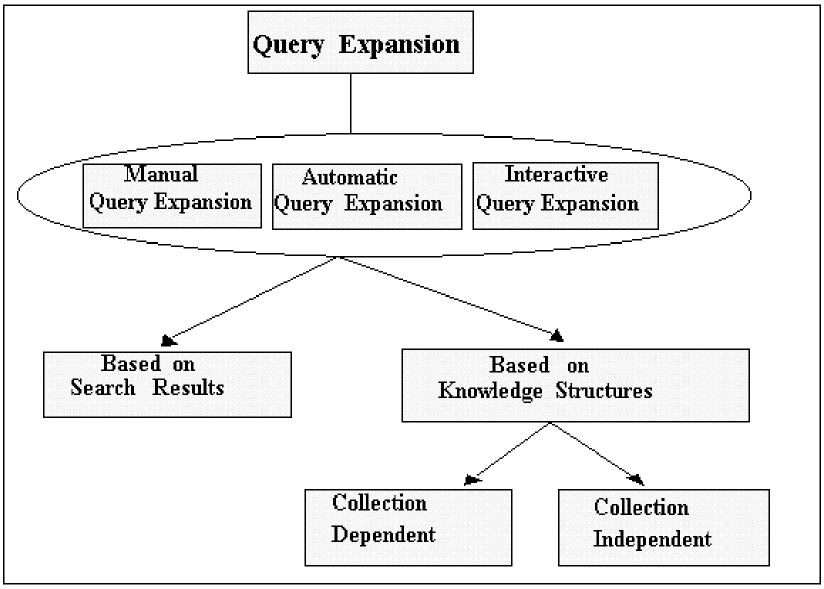


Figure 1: Query Expansion: Methods and Sources

Query expansion, as depicted in Figure 1, can be performed manually, automatically or interactively (also known as semi-automatic, user mediated, and user assisted).

## 2. Related Works

The existing state-of-the-art query expansion approaches can be classified mainly into two classes: global analysis and local analysis.

Global analysis is one of the first techniques to produce consistent and effective improvements through query expansion. One of the earliest global analysis techniques is term clustering [5], which groups document terms into clusters based on their co-occurrences. Queries are expanded by the terms in the same cluster. Other well-known global techniques include Latent Semantic Indexing [6], similarity thesauri [1], and Phrase Finder [7]. Global analysis requires corpus-wide statistics such as statistics of co-occurrences of pairs of terms, which results in a similarity matrix among terms. To expand a query, terms which are the most similar to the query terms are identified and added. The global analysis techniques are relatively robust; but corpus-wide statistical analysis consumes a considerable amount of computing resources. Moreover, since it only focuses on the document side and does not take into account the query side, global analysis cannot address the term mismatch problem well.

Different from global analysis, local analysis uses only some initially retrieved documents for further query expansion. The idea of local analysis can be traced back at least to a 1977 paper [8]. A well-known local analysis technique is relevance feedback [9,10], which modifies a query based on users' relevance judgments of the retrieved documents. Typically, expansion terms are extracted from the relevant documents. Relevance feedback can achieve very good performance if the users provide sufficient and correct relevance judgments. Unfortunately, in a real search context, users usually are reluctant to provide such relevance feedback information. Therefore, relevance feedback is seldom used by the commercial search engines.

To overcome the difficulty due to the lack of sufficient relevance judgments, pseudo-relevance feedback (also known as blind feedback) is commonly used. Local feedback mimics relevance feedback by assuming the top-ranked documents to be relevant [11]. Expansion terms are extracted from the top-ranked documents to formulate a new query for a second cycle retrieval.

In recent years, many improvements have been obtained on the basis of local feedback, including re-ranking the retrieved documents using automatically constructed fuzzy Boolean filters [12], clustering the top-ranked documents and removing the singleton clusters [13], clustering the retrieved documents and using the terms that best match the original query for expansion. In addition, recent TREC results show that local feedback approaches are effective and, in some cases, outperform global analysis techniques [14]. Nevertheless, this method has an obvious drawback: if a large fraction of the top-ranked documents is actually irrelevant, then the words added to the query (drawn from these documents) are likely to be unrelated to the topic and as a result, the quality of the retrieval using the expanded query is likely to be worse. Thus the effects of pseudo-feedback strongly depend on the quality of the initial retrieval.

Recently, Xu and Croft [15] proposed a local context analysis method, which combines both local analysis and global analysis. First, noun groups are used as concepts, which are selected according to their co-occurrences with the query terms. Then concepts are chosen from the top-ranked documents, similarly to local feedback.

## 3. Traditional Document Retrieval

The task of traditional document retrieval is to retrieve documents which are relevant to a given query from a fixed set of documents, i.e. a document database. In a common way to deal with documents as well as queries, they are represented using a set of index terms (simply called terms) by ignoring their positions in documents and queries. Terms are determined based on words of documents in the database, usually during pre-processing phases where some normalization procedures are incorporated (e.g. stemming and stop-word elimination).

### 3.1 Vector Space Model

The vector-processing model of retrieval is used to transform both the available information requests as well as the stored documents into vectors of the form:

$$D_i = (w_{i1}, w_{i2}, ...., w_{it}) \tag{1}$$

where $D_i$ represents a document (or query) text and $w_{ik}$ is the weight of term $T_k$ in document $D_i$. A weight of zero is used for terms that are absent from a particular document, and positive weights characterize terms actually assigned. The assumption is that $t$ terms in all are available for the representation of the information.

In choosing a term weighting system, low weights should be assigned to high-frequency terms that occur in many documents of a collection, and high weights to terms that are important in particular documents but unimportant in the remainder of the collection. The weight of terms that occur rarely in a collection is relatively unimportant because such terms contribute little to the needed similarity computation between different texts.

A well-known term weighting system following that prescription assigns weight $w_{ik}$ to term $T_k$ in query $Q_i$ in proportion to the frequency of occurrence of the term in $Q_i$, and in inverse proportion to the number of documents to which the term is assigned. [16, 17] Such a weighting system is known as a tf x idf (term frequency times inverse document frequency) weighting system. In practice the query lengths, and hence the number of of non-zero term weights assigned to a query, vary widely. To allow a meaningful final retrieval similarity, it is convenient to use a length normalization factor as part of the term weighting formula. A high- quality term weighting formula for $w_{ik}$, the weight of term $T_k$ in query $Q_i$ is

$$w_{ik} = \frac{(\log(f_{ik}) + 1.0) * \log(N/n_k)}{\sqrt{\sum_{j=1}^{t} [(\log(f_{ij}) + 1.0) * \log(N/n_j)]^2}} \tag{2}$$

where $f_{ik}$ is the occurrence frequency of $T_k$ in $Q_i$, N is the collection size, and $n_k$ the number of documents with term $T_k$ assigned. The factor $\log(N/n_k)$ is an inverse collection frequency ("idf") factor which decreases as terms are used widely in a collection, and the denominator in expression (2) is used for weight normalization.

The weight assigned to terms in *documents* are much the same. In practice, for both effectiveness and efficiency reasons the *idf* factor in the documents is dropped [18, 19]. The term $T_k$ included in a given vector can in principle represent any entities assigned to a document for content identification. Such terms are derived by a text transformation of the following kind: [20]

1. recognize individual text words
2. use stop lists to eliminate unwanted function words
3. perform suffix removal to generate word stems
4. optionally use term grouping methods based on statistical word co-occurrence or word adjacency computations to form term phrases (alternatively syntactic analysis computations can be used)
5. assign term weights to all remaining word stems and /or phrase stems to form the term vector for all information items.

Once term vectors are available for all information items, all subsequent processing is based on term vector manipulations.

The fact that the indexing of both documents and queries is completely automatic means that the results obtained are reasonably collected independently and should be valid across a wide range of collections.

### 3.1.1 Text Similarity Computation

When the text of document $D_i$ is represented by a vectors of the form ( $d_{i1}$, $d_{i2}$, …, $d_{it}$) and query $Q_j$ by the vector ($q_{j1}$ , $q_{j2}$, …,$q_{jt}$), a similarity (S) computation between the two items can conveniently be obtained as the inner product between corresponding weighted term vector as follows:

$$S(D_i, Q_j) = \sum_{k=1}^{t} (d_{ik} * q_{jk}) \tag{3}$$

Thus, the similarity between two texts (whether query or document) depends on the weights of coinciding terms in the two vectors.

In the following section we discuss the query expansion technique that will be used for comparison.

## 4. Query expansion

Query expansion algorithms at first evaluate given query on collection of documents, and then select from relevant documents appropriate terms. The original query is expanded with such selected terms. The expanded query is used to retrieve new set of relevant documents. In this paper we apply two query expansion methods in sequence to reformulate the query so that it will suit to the user's needs more appropriately. One method we applied is similarity thesaurus based expansion [1], and the other is local feedback method. The similarity thesaurus we use, based on [1], calculates the relevance between terms and queries and is constructed by interchanging the role of documents and terms in retrieval model. The relevance of a term in the similarity thesaurus to the concept of the query is the sum of the weighted relevance of the term to each term in the query. The queries are expanded by adding top *n* relevant terms, which are most similar to the concept of the query, rather than selecting terms that are similar to the query terms.

The local feedback method is similar to traditional relevance feedback method [21], which modifies queries by using the result of the initial retrieval, except that the latter uses the judgment set for calculating re-weighting while the former assumes that the terms in the top ranked *n* documents are relevant to the user's request. Queries are expanded by adding the weight of terms in relevant documents and reducing the weight of terms in last *m* documents of the initial retrieval.

We modify the traditional Rocchio expansion equation to include the query expanded by the thesaurus method and to include negative evidence from the lowest ranked documents rather than non-relevant documents. The new query $Q_{new}$, including thesaurus expansion, can be defined as the following:

$$Q_{new} = \alpha_1 Q_{org} + \alpha_2 + Q_{te} + \beta \sum_{top} D_i - \gamma \sum_{last} D_j \tag{4}$$

Here, $Q_{org}$ is a initial query, $Q_{te}$ is a query expanded by the similarity thesaurus based method, $\sum_{top} D_i$ represents terms in top ranked documents retrieved in the initial run, and $\sum_{last} D_j$ is terms in low ranked documents. The parameters $\alpha_1, \alpha_2, \beta$ and $\gamma$ represent the importance of each item. Currently, these parameters are given by human experience. For the initial retrieval, we used the queries expanded by thesaurus method. In this study, we set the parameters as following: $\alpha_1 = 1$, $\alpha_2 = 0.5$, $\beta = 0.6$, and $\gamma = 0.3$.

## 5. Experiments and their Results

In our experiments, we used the three standard test collections (CISI, NPL, and CACM). We evaluate the performance of the retrieval by average precision measure. Precision is the ratio of the number of relevant documents retrieved to the total number retrieved. The average precision of a query is the average of precisions calculated when a relevant document is found in the rank list. All the query's average precisions are averaged to evaluate an experiment.

Table (1) shows the retrieval quality difference between the original queries and the expanded queries. It seems that the improvement increases with the size of collection.

Table 1: Improvement using expanded queries

| Collection | CISI | CACM | NPL |
|---|---|---|---|
| Documents | 1035 | 3205 | 11430 |
| Avg. precision of original query | 0.5547 | 0.2819 | 0.1918 |
| Number of additional terms | 80 | 100 | 800 |
| Avg. precision of expanded query | 0.6445 | 0.3438 | 0.2448 |
| Improvement | 16.19 % | 21.96 % | 27.63 % |

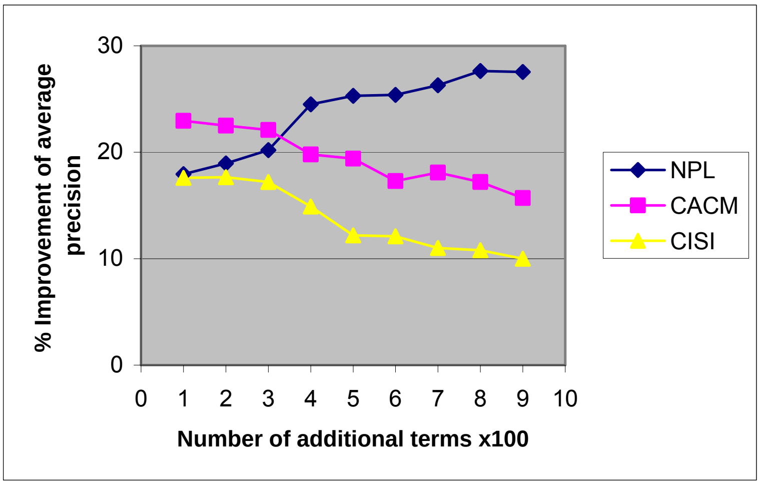


fig.2: Improvement using expanded queries with various numbers of additional terms

The figure indicates that our query expansion technique yields a considerable improvement in the retrieval effectiveness. It seems that the improvement increases with the size of the collection. In addition, the improvement increases with the number of additional search terms that expand the original query as long as the collection is large enough. In Fig. 2, we show how the number of additional terms affects the retrieval effectiveness. It can be seen easily that the improvement by expanded queries increases when the number of additional terms increases. When the number of additional terms is between 100 and 200, the improvement of the retrieval effectiveness remains constant in the small collections CISI and CACM. Once the number of additional terms gets to be larger than 200, the improvement decreases in the small collections, but continues to increase in the relatively large collection NPL. This could be explained by the fact that more search terms are needed to distinguish relevant documents from non-relevant documents in large collections.

## 6. Conclusion

We presented a two query expansion methods in sequence to reformulate the query. Our experiments made on three standard test collections with different sizes and different document types have shown considerable improvements vs. the original queries in the standard vector space model. Experiments on test collections showed that the improvement increases with the size of the collection. In addition, the improvement increases with the number of additional search terms that expand the original query as long as the collection is large enough. Also it has been pointed out how the number of additional terms affects the retrieval effectiveness.

## Bibliography

[1] Y., Qiu and H. P. Frei. Concept based query expansion. In Proceedings of the ACM-SIGIR Intl. Conference on Research and Development in Information Retrieval, pages 160-169, 1993.

[2] E. M., Voorhees. Query expansion using lexical-semantic relations. In Proceedings of the ACM-SIGIR Intl. Conference on Research and Development in Information Retrieval, Dublin, pages (61-70), 1994.

[3] J., Greenberg, Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology. Journal of the American Society for Information Science and Technology, 52(6), 487-498, (2001).

[4] E. N., Efthimiadis, Query expansion. Annual Review of Information Science and Technology, 31, 121-187, 1996.

[5] S. K. Jones, Automatic keyword classification for information retrieval, Butterworth's (London), 1971

[6] S. Deerwester, S. T. Dumai, G.W. Furnas, T. K. Landauer, R. Harshman, Indexing by latent semantic analysis, Journal of the American Society For Information Science volume 41(4)pages 391-407,1994

[7] Y. Jing, W. B. Croft, An association thesaurus for information retrieval, RIAO 94 Conference Proceedings pages 146-160,1994

[8] R. Attar, A. S. Fraenkel, Local feedback in full-text retrieval systems, Journal of the ACM volume 24(3) pages 397-417,1977

[9] J. Rocchio, Relevance feedback in information retrieval. The Smart Retrieval system—Experiments in Automatic Document Processing, Prentice Hall, 1971

[10] G. Salton, C. Buckley, Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science volume 41(4)pages 288-297,1990

[11] C. Buckley, G. Salton, J. Allan, A. Singhal, Automatic query expansion using SMART, Proceedings of the 3rd Text REtrieval Conference, 1994

[12] A. Singhal, M. Mitra, C. Buckley, Improving Automatic Query Expansion, SIGIR'98, 1998

[13] A. Lu, M. Ayoub, J. Dong, Ad hoc experiments using EUREKA, Proceedings of the 6th Text Retrieval Conference,1997

[14] J. Xu, W.B. Croft, Query expansion using local and global document analysis, Proceedings of the 17th ACM SIGIR,1994

[15] J. Xu, W. B. Croft, Improving the effectiveness of information retrieval with local context analysis, ACM Transactions on Information Systems,2000.

[16] G., Salton and C., Buckley, Term-weighting approaches in automatic text retrieval. Information Processing and Management, 24(5): 513-523, 1988.

[17] G., Salton, Automatic Text Processing – the Transformation, Analysis and Retrieval of Information by Computer. Addison –Wesley Publishing Co., Reading, MA, 1989.

[18] C., Buckley, G., Salton , and James Allan. Automatic retrieval with locality information using SMART. In D. K. Harman, editor, Proceedings of the First Text Retrieval conference (TREC-1), pages 59-72. NIST Special Publication 500-207, March 1993.

[19] C., Buckley, J., Allan, and G., Salton. Automatic routing and ad-hoc retrieval using SMART: TREC 2. In D. K. Harman, editor, Proceedings of the Second Text Retrieval conference (TREC-2), pages 45-56. NIST Special Publication 500-215, March 1994.

[20] G., Salton, Automatic Text Processing – the Transformation, Analysis and retrieval of Information by Computer. Addison –Wesley Publishing Co., Reading, MA, 1989.

[21] J., Rocchio, The SMART Retrieval System Experiments in Automatic Document Processing, Chapter: Relevance Feedback in Information Retrieval, 313{323, Prentice Hall, 1971.

## Author's Information

***A. A. Aly*** *– Computer Science Department, El - Minia University; e-mail: abdelmgeid@yahoo.com*

# DIGRAPHS DEFINITION FOR AN ARRAY MAINTENANCE PROBLEM

**Angel Herranz, Adriana Toni**

***Abstract**: In this paper we present a data structure which improves the average complexity of the operations of updating and a certain type of retrieving information on an array. The data structure is devised from a particular family of digraphs verifying conditions so that they represent solutions for this problem.*

***Keywords**: array maintenance, average complexity, data structures, models of computation*

## Introduction

Let A be an array of fixed length N with elements belonging to a commutative semigroup an let us consider two operations, Update and Retrieve, with the following intended effect:

- • Update(i, x) increments the i-th element of A in x (A(i):= A(i)+ x;).
- • Retrieve(i, j) outputs A(i)+ A(i + 1) + ... + A(j).

The less space consuming and, likely, the most natural data structure for implementing both operations is the array itself (from now on, expression i..j denotes the set $\{k \in \mathbb{N} \bullet i \leq k \wedge k \leq j\}$):

Example 1

```
Update(i, x) :
begin
   $A(i)$  :=  $A(i) + x$;
end;
```

```
Retrieve(i, j) :
begin
   return $\sum_{k \in i..j} A(i)$
end;
```

Running in a random access memory machine, the complexity of Update(i, x) is constant whilst, in the worst case, the complexity of Retrieve(i, j) is linear on N. To improve the complexity of Retrieve the data structure can be reified as an array S of length N + 1 with the property $S(i) = \sum_{k \in i..N} A(i)$. Then programs must be adapted:

Example 2

```
Update(i, x) :
for k in i..N loop
   $S(k)$  :=  $S(k) + x$;
end loop;
```

```
Retrieve(i, j) :
begin
   return $S(i) - S(j+1)$;
end;
```

For this implementation the complexity of Retrieve is constant whereas, in the worst case, the complexity of Update is linear on N. The design in Example 2 assumes the existence of - (the inverse of +) in the model. This consideration aside, under any execution sequence of operations Update and Retrieve, both implementations are indistinguishable from a functional point of view. This means that Example 1 and Example 2 are different solutions to the same problem definition.

In this paper we are interested in designs with a good average complexity of Update and Retrieve operations when the program variables store elements of a commutative semigroup. Obviously, programs must yield the correct result irrespective of the particular semigroup. Uniform probability of Update and Retrieve execution in programs is assumed in order to improve what we have called average complexity, in other words, we are trying to minimise the sumof the costs of all executions.

In the next Section the RQP (Range Query Problem) and its solutions in terms of digraphs are formalised. Then a particular family of digraphs that represent solutions to the RQP will be presented in an informal way.

## Range query Problem

In this section, the range query problem and its solutions are formalised. In this formalisation, arrays store elements that belongs to a commutative semigroup S. Let us start with the definition of arrays used in this paper.

**Definition 3**

An array A of length N is a total function from 1..N into S.

**Criterion 4**

Let A be an array of length N interpreted as a function from 1..N into S: A . 1..N . S. |A| denotes N, dom A denotes 1..N and ran A denotes S.

**Definition 5**

The Range Query Problem of size N (N-RQP) is the analysis and design of data structures for the implementation of the operations Update and Retrieve where both operations are interpreted as higher order functions:

$$Update\ :\ (1..N \rightarrow \mathbf{S}) \times 1..N \times \mathbf{S} \rightarrow (1..N \rightarrow \mathbf{S})$$

$$Update(A, i, x)(j) = \begin{cases} A(j) + x & \boldsymbol{if}\ i = j \\ A(j) & \boldsymbol{otherwise} \end{cases}$$

$$Retrieve\ :\ (1..N \rightarrow \mathbf{S}) \times 1..N \times 1..N \rightarrow \mathbf{S}$$

$$Retrieve(A, i, j) = \sum_{k \in i..j} A(k)$$

**Definition 6**

A N-RQP design is a triple (Z, U, R) where Z is an array of length M with N less or equal M, U is a family of subsets of 1..M indexed on 1..N and R is a family of subsets of 1..M indexed on 1..N × 1..N. Given a N-RQP design (Z, U, R), the implementation of the operations Update and Retrieve is:

```
procedure Update
   ($i$ : $1..N$, $x$ : $\mathbf{S}$) is
begin
   for $k$ in $U_i$ loop
      $Z(k)$ := $Z(k) + x$;
   end loop;
end Update;
```

```
function Retrieve
   ($i$ : $1..N$, $j$ : $1..N$)
   return $\mathbf{S}$ is
begin
   return $\sum_{k \in R_{ij}} Z(i)$
end Retrieve;
```

**Lemma 7**

The complexity of the implementation of Update(i, x) and Retrieve(i, j) in Definition 6 is linear on the cardinal of Ui and Rij, respectively, when running on a random access memory machine.

PROOF. Trivial

**Definition 8**

A N-RQP design (Z, U, R) is a N-RQP solution if and only if for every $i, j, k \in N$ and $x, y \in S$ the following triplets in the programming logic (annotated programs) are totally correct:

```
-- $i \leq k \ \wedge \ k \leq j$                  -- $(k < i \ \vee \ k > j)$
-- $\wedge \ Retrieve(i, j) = x$                  -- $\wedge \ Retrieve(i, j) = x$
$Update(k, y)$;                                   $Update(k, y)$;
-- $Retrieve(i, j) = x + y$                        -- $Retrieve(i, j) = x$
```

**Lemma 9**

A N-RQP design (Z, U, R) is a N-RQP solution if and only if

$$\forall i, j, k \in 1..N \ \bullet \ |R_{ij} \ \cap \ U_k| = \begin{cases} 1 \ \boldsymbol{if} \ i \leq k \ \wedge \ k \leq j \\ 0 \ \boldsymbol{otherwise} \end{cases}$$

PROOF. This is a well known result and a proof can be found in [1].

## Average Complexity in RQP Solutions

As we mentioned in previous sections, we will try to minimise the sum of the costs of all different executions of Update and Retrieve. A uniform probability distribution for each Update possible execution (N operations) and Retrieve possible execution ($\left(\frac{N+1}{2}\right)$ operations) is assumed.

**Definition 10**

The average complexity of a N-RQP design (Z, U, R) is

$$\frac{\sum_{i=1,j=i}^{N,N} |R_{ij}| + \sum_{i=1}^{N} |U_i|}{N + \binom{N+1}{2}}$$

Minimising function $\phi$ below, is enough to minimise the average complexity function above.

$$\phi(N) = \sum_{i=1,j=i}^{N,N} |R_{ij}| + \sum_{i=1}^{N} |U_i|$$

## RQP Solutions as Graphs

N-RQP designs can be described in terms of graphs where the content of Rij and Ui are represented as graph vertices and edges (Definition 16). Let us start with some basic definitions.

**Definition 11**

A digraph, or directed graph, G is a pair (V, E), where V is a finite set (vertex set) and E is a binary relation on V (edge set).

**Criterion 12**

Notation $u \rightarrow v$, instead of (u, v), is used to denote the edges

**Definition 13**

Let G =(V, E) be a digraph, the out-degree of a vertex u is $|\{v \in V \mid u \rightarrow v \in E\}|$ and the in-degree of a vertex $v$ is $|\{u \in V \mid u \rightarrow v \in E\}|$.

**Definition 14**

If there is a path from v1 to v2 in a digraph G =(V, E) we say that v2 is reachable from v1. Functions Successors and Ancestors are defined as:

$$Successors(G,u) = \{v \in V \mid v \text{ is reachable from } u\}$$

$$Ancestors(G,v) = \{u \in V \mid v \text{ is reachable from } u\}$$

$$Successors*(G,u) = \{u\} \cup Successors(G,u)$$

$$Ancestors*(G,v) = \{v\} \cup Ancestors(G,v)$$

**Definition 15**

An acyclic digraph G =(V, E) is a N-RQP graph if the following conditions hold:

*(1)* $V = 1..M$ *with* $N \leq M$.
*(2) For every vertex* $v \leq N$, *its out-degree is* $0$.
*(3) For every vertex* $v > N$, $Successors(G,v) \ \cap\ 1..N \neq \emptyset$.

**Definition 16**

Given a N-RQP graph G =(V, E), the N-RQP design (Z, U, R) is a N-RQP design in terms of G if it verifies the following properties:

*(1)* $|Z| = |V|$
*(2)* $U_i = Ancestors^*(G,i)$
*(3)* $R_{ij}$ *is the set of vertices with the smallest cardinal that verifies:*

$$\bigcup_{u \in R_{ij}} Successors^*(G,u) \ \cap\ 1..N = i..j$$

$$\bigcap_{u \in R_{ij}} Successors^*(G,u) \ \cap\ 1..N = \emptyset$$

The existence of $R_{ij}$ is guaranteed for every i, j such that $1 \leq i \leq j \leq N$ because in the absence of a set $R_{ij}$ with a cardinal less than j. i +1 we would end up with $R_{ij}$ = i..j. With respect to the uniqueness of $R_{ij}$ several sets could exist with a smallest cardinal verifying the conditions in Definition 16 so an arbitrary criterion should be given (lexicographic order, for instance).

The following theorem states that N-RQP graphs represent N-RQP solutions:

**Theorem 17**

Let G =(V, E) be a N-RQP graph, a N-RQP design in terms of G is a N-RQP solution.

PROOF. Let us consider the execution of an arbitrary program:

```
r := Retrieve(i,j);
Update(k,x);
r' := Retrieve(i,j);
```

with $i,k \in 1..N$ and $j \in i..2N$. We have to prove that if $i \leq k \leq j$ then $r' = r + x$; otherwise $r' = r$. The proof is based on the following obvious fact: $r' = r + |U_k \cap R_{ij}| x$ (observe that $Retrieve(i, j)$ is $Z(u_1)$+ ... + $Z(u_n)$ where $R_{ij} = \{u_1,...,u_n\}$).

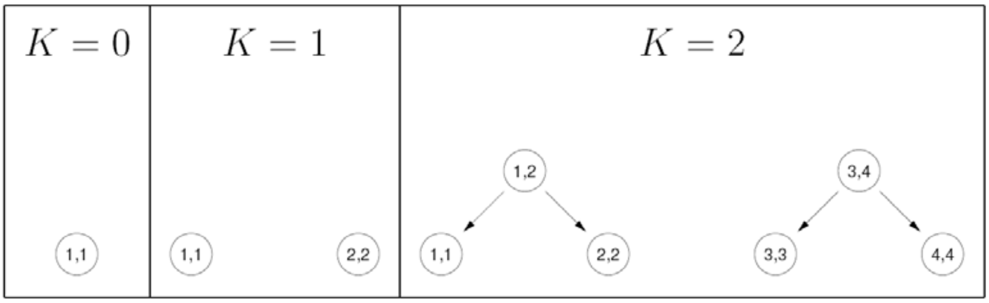


Fig. 1. $2^K$-RQP graphs for $K \in 0..2$

- Case $k < i \vee j < k$: in this case $U_k \cap R_{ij} = \emptyset$ therefore $r' = r$.
- Case $i \le k \le j$: in this case $|U_k \cap R_{ij}| = 1$ therefore $r' = r + x$.

## Constructing RQP Solutions

The inspiration of our approach comes from a particular family of N-RQP graphs where N is a power of 2. In the solution designs in terms of graphs of this kind, the cost of Retrieve operations is less or equal to 2.

Graphs of this family are called $2^k$- RQP graphs, with $K \in \mathbb{N}$. The construction method is described by induction on K and Figure 1 presents the trivial examples for $K$=0,1,2.

The reader will observe that, strictly speaking, graphs presented in this section are not N-RQP graphs because their vertices are not positive integers but pairs (i, j) of positive integers where $(i \le j \wedge j \le N)$. This is not an important problem, as pairs can be trivially encoded as positive integers 1 and an isomorphic N-RQP graph would be obtained. Authors pursue elegance in the presentation so vertices as pairs (i, j) are maintained.

The main characteristic of the construction of N-RQP solutions is that our graphs have the following property:

$$1 \le \left|R_{ij}\right| \wedge \left|R_{ij}\right| \le 2$$

Intuitively, the $2^{K+1}$-RQP graph can be built by cloning twice the $2^K$-RQP graph and then adding new vertices and edges that maintain the above mentioned property. To achieve this aim, after cloning the $2^K$-RQP graph, new vertices and edges will be added taking into account that the property on Rij holds if j $\le 2^K$ or i> $2^K$ .

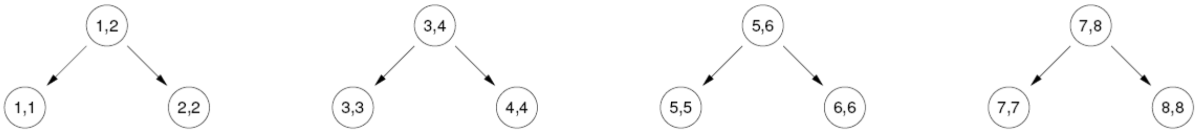


Fig. 2. A 8-RQP graph resulting after cloning the $2^2$-RQP graph

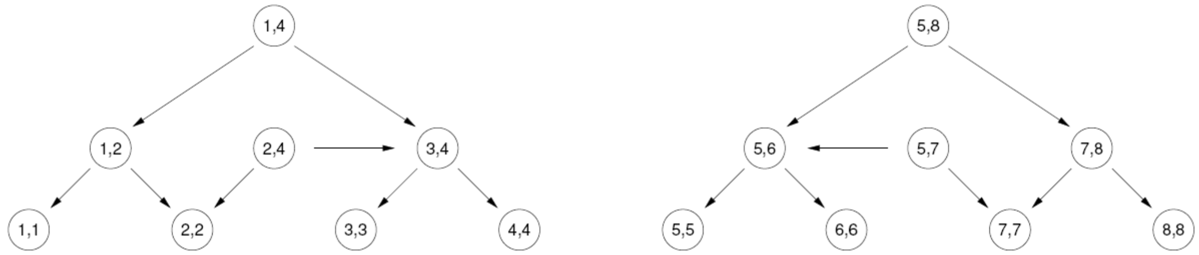


Fig. 3. The $2^3$-RQP graph

Let us show an example, the 8-RQP graph in Figure 2 is the result of cloning the $2^2$-RQP graph in Figure 1. In the 8-RQP design (Z, U, R) in terms of that graph, the values of all Ui and those Rij such that |Rij| > 2 are:

$U_1 = \{(1, 1), (1, 2)\}$ $U_2 = \{(2, 2), (1, 2)\}$ $U_3 = \{(3, 3), (3, 4)\}$ $U_4 = \{(4, 4), (3, 4)\}$

$U_5 = \{(5, 5), (5, 6)\}$ $U_6 = \{(6, 6), (5, 6)\}$ $U_7 = \{(7, 7), (7, 8)\}$ $U_8 = \{(8, 8), (7, 8)\}$

$R_{15} = \{(1, 2), (3, 4), (5, 5)\}$ $R_{16} = \{(1, 2), (3, 4), (5, 6)\}$

$R_{17} = \{(1, 2), (3, 4), (5, 6), (7, 7)\}$ $R_{18} = \{(1, 2), (3, 4), (5, 6), (7, 8)\}$

$R_{25} = \{(2, 2), (3, 4), (5, 5)\}$ $R_{26} = \{(2, 2), (3, 4), (5, 6)\}$

$R_{27} = \{(2, 2), (3, 4), (5, 6), (7, 7)\}$ $R_{28} = \{(2, 2), (3, 4), (5, 6), (7, 8)\}$

$R_{37} = \{(3, 4), (5, 6), (7, 7)\}$ $R_{38} = \{(3, 4), (5, 6), (7, 8)\}$

$R_{47} = \{(4, 4), (5, 6), (7, 7)\}$ $R_{48} = \{(4, 4), (5, 6), (7, 8)\}$

The idea is that new vertices and edges have to be added in order to decrease the cardinal of $R_{i4}$ and $R_{5j}$ to 1. $R_{ij}$ is then obtained as the union of $R_{i4}$ and $R_{5j}$ with a resulting cardinal of 2. In the example $|R_{14}|=|R_{24}|=2$ so a pair of vertices are added representing $R_{14}=R_{12}\cup R_{34}$ and $R_{24}=R_{22}\cup R_{34}$. Reasoning symmetrically with $R_{57}$ and $R_{58}$ we get the $2^3$-RQP graph in Figure 3. The application of the idea is shown in the left half of the 16-RQP graph (after clowning the $2^3$-RQP) in Figure 4.

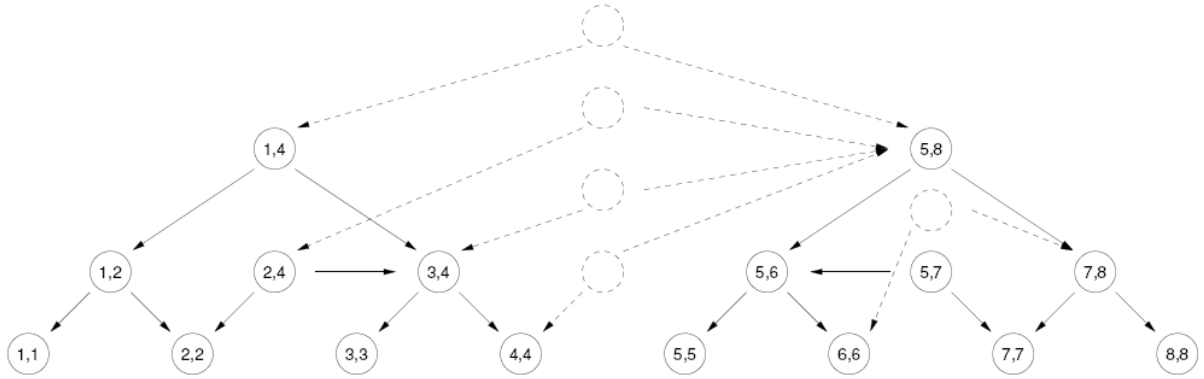


Fig. 4. Left half of the $2^4$-RQP graph

Next we present now the formalization of $2^k$-RQP graphs.

**Definition 18**

Let K be a natural number. A $2^K$ -RQP graph GK is defined inductively:

$$\text{(1) If } K = 0 \text{ then } G^K = (\{(1, 1)\}, \emptyset)$$
$$\text{(2) If } K > 0 \text{ then } G^K = Duplicate(G^{K-1})$$

where function Duplicate is defined as

```
function Duplicate ($G^K = (V^K,E^K)$ : Digraph) return Digraph is
  N : constant ℕ := $2^K$
  M : constant ℕ := $|V^K|$
  V : {(i, j) ∈ 1..N × 1..N • i ≤ j} := ∅;
  E : P(V×V) := 0;
  i, j : 1..(2N);
begin
  -- The ‘‘cloning’’ loops
  for $(i, j)$ in $V^K$ loop
    $V := V \cup \{(i, j), (i + N, j + N)\}$;
  end loop;
  for $(i, j) \rightarrow (i', j')$ in $E^K$ loop
    $E := E \cup \{(i, j) \rightarrow (i', j'), (i + N, j + N) \rightarrow (i' + N, j' + N)\}$;
  end loop;
  -- (V,E) is a graph with two subgraphs which are just like G
  -- but with different node numbering
  for i in $1..(N - 1)$ loop -- The ''left half'' loop
    $j := i + 1$;
    while $(i,N) \notin V \wedge j \leq N$ loop
      if $(i, j) \in V \wedge (j, N) \in V$ then
        $V := V \cup \{(i, N)\}$;
        $E := E \cup \{(i, N) \rightarrow (i, j), (i,N) \rightarrow (j,N)\}$;
      else
        $j := j + 1$;
      end if;
    end loop;
  end loop;

  for j in $(N + 2)..(2N)$ loop -- The ''left half'' loop
    $i := j - 1$;
    while $(N + 1, j) \notin V \wedge i \leq 2N$ loop
      if $(N, i) \in V \wedge (i, j) \in V$ then
        $V := V \cup \{(N, j)\}$;
        $E := E \cup \{(N, j) \rightarrow (N, i), (N, j) \rightarrow (i, j)\}$;
      else
        $i := i + 1$;
      end if;
    end loop;
  end loop;
  return (V,E);
end Duplicate;
```

## Bibliography

[1] D.J. Volper, M.L. Fredman, *Query Time Versus Redundancy Trade-offs for Range Queries*, Journal of Computer and System Sciences 23, (1981) pp.355--365.

[2] W.A. Burkhard, M.L. Fredman, D.J.Kleitman, *Inherent complexity trade-offs for range query problems*, Theoretical Computer science, North Holland Publishing Company 16, (1981) pp.279--290.

[3] M.L. Fredman, *The Complexity of Maintaining an Array and Computing its Partial Sums*, J.ACM, Vol.29, No.1 (1982) pp.250--260.

[4] A. Toni, *Lower Bounds on Zero-one Matrices*, Linear Algebra and its Applications, 376 (2004) 275--282.

[5] A. Toni, Complejidad y Estructuras de Datos para e*l problema de los rangos variables*, Doctoral Thesis, Facultad de Informática, Universidad Politécnica de Madrid, 2003.

[6] A. Toni, *Matricial Model for the Range Query Problem and Lower Bounds on Complexity*, submitted.

## Authors' Information

***Ángel Herranz Nieva** – Assistant Professor; Departamento de Lenguajes y Sistemas Informáticos; Facultad de Informática; Universidad Politécnica de Madrid; e-mail: aherranz@fi.upm.es*

***Adriana Toni** – Facultad de Informática, Universidad Politécnica de Madrid, Spain; e-mail: atoni@fi.upm.es*

# AN EFFECTIVE METHOD FOR CONSTRUCTING DATA STRUCTURES SOLVING AN ARRAY MAINTENANCE PROBLEM

Adriana Toni, Angel Herranz, Juan Castellanos

***Abstract***: *In this paper a constructive method of data structures solving an array maintenance problem is offered. These data structures are defined in terms of a family of digraphs which have previously been defined, representing solutions for this problem. We present as well a prototype of the method in Haskell*

***Keywords***: *array maintenance, average complexity, data structures, models of computation*

## Introduction

The Range Query Problem of size N (N-RQP) deals with the analysis and design of data structures for the implementation of the operations Update and Retrieve: let A be an array of length N of elements of a commutative semigroup, Update(i, x) increments A(i) (i-th element of A) in x and Retrieve(i, j) outputs the partial sum A(i)+..+A(j).

In [4] we find the following definition of N-RQP design.

**Definition N-RQP design**

A N-RQP design is a triple (Z, U, R) where Z is an array of length M with N less or equal M, U is a family of subsets of 1..M indexed on 1..N and R is a family of subsets of 1..M indexed on 1..N × 1..N. Given a N-RQP design (Z, U, R), the implementation of the operations Update and Retrieve is:

```
procedure Update
    (i: 1..N, x:S) is
begin
    for k in $U_i$ loop
     Z(k) := Z(k) + x;
    end loop;
end Update;
```

```
function Retrieve
    (i: 1..N, j: 1..N)
    return S is
begin
    return $\sum_{k \in R_{ij}} Z(i)$
end Retrieve;
```

It is a well known result that an N-RQP design (Z, U, R) is a N-RQP solution if and only if

$$\forall i, j, k \in 1..N \bullet \left| R_{ij} \cap U_k \right| = \begin{cases} 1 & \text{if } i \le k \wedge k \le j \\ 0 & \text{otherwise} \end{cases}$$

and a proof can be found in [1].

In [4] we find the three definitions below as well.

**Definition N-RQP graph**

An acyclic digraph G =(V, E) is a N-RQP graph if the following conditions hold:

*(1)* $V=1..M$ *with* $N \le M$.

*(2) For every vertex* $v \le N$*, its out-degree is 0.*

*(3) For every vertex* $v>N$*, Successors(G,v)*$\cap 1..N \neq 0$.

**Definition N-RQP design in terms of G**

Given a N-RQP graph G =(V, E), the N-RQP design (Z, U, R) is a N-RQP design in terms of G if it verifies the following properties:

*(1)* $|Z| = |V|$

*(2)* $U_i$*=Ancestors*$^\bullet$*(G,i)*

*(3)* $R_{ij}$ *is the set of vertices with the smallest cardinal that verifies:*

$$\bigcup_{u \in R_{ij}} Successors^\bullet(G,u) \cap 1..N = i..j$$

$$\bigcup_{u \in R_{ij}} Successors^\bullet(G,u) \cap 1..N = 0$$

being

$$Successors^\bullet(G,u) = \{u\} \cup Successors(G,u)$$

$$Ancestors^\bullet(G,v) = \{v\} \cup Ancestors(G,v)$$

and in the same paper it has been proved that given a N-RQP graph, a N-RQP design in terms of G is a N-RQP solution.

**Definition $2^K$ -RQP graph**

Let K be a natural number. A $2^K$ -RQP graph $G^K$ is defined inductively:

(1) $If \quad K = 0 \quad then \quad G^K = (\{(1,1)\}, 0)$

(2) $If \quad K > 0 \quad then \quad G^k = Duplicate(G^{K-1})$

where function Duplicate is defined as

```
function Duplicate (G^K = (V^K,E^K) : Digraph) return Digraph is
  N : constant ℕ := 2^K
  M : constant ℕ := |V^K|
  V : {(i, j) ∈ 1..N × 1..N • i ≤ j} := 0;
  E :  P(V×V) := 0;
  i, j : 1..(2N);
begin
  -- The ''cloning'' loops
  for (i, j) in V^K loop
    V := V ∪ {(i, j), (i + N, j + N)};
  end loop;
  for (i, j) → (i', j') in E^K loop
    E := E ∪ {(i, j) → (i', j'), (i + N, j + N) → (i' + N, j' + N)};
  end loop;
  -- (V,E) is a graph with two subgraphs which are just like G
  -- but with different node numbering
  for i in 1..(N−1) loop -- The ''left half'' loop
    j := i + 1;
    while (i,N) ∉ V ∧ j ≤ N loop
      if (i, j) ∈ V ∧ (j, N) ∈ V then
        V := V ∪ {(i, N)};
        E := E ∪ {(i, N) → (i, j), (i,N) → (j,N)};
      else
        j := j + 1;
      end if;
    end loop;
  end loop;
  for j in (N + 2)..(2N) loop -- The ''left half'' loop
    i := j − 1;
    while (N + 1, j) ∉ V ∧ i ≤ 2N loop
      if (N, i) ∈ V ∧ (i, j) ∈ V then
        V := V ∪ {(N, j)};
        E := E ∪ {(N, j) → (N, i), (N, j) → (i, j)};
      else
        i := i + 1;
      end if;
    end loop;
  end loop;

  return (V,E);
end Duplicate;
```

Obviously, the $2^K$ -RQP design (Z,U,R) can be computed after the construction of the $2^K$ -RQP graph as described by the following brute force algorithm:

**Algorithm 1** *The following algorithm computes $R_{ij}$ for a N-RQP design in terms of a N-RQP graph G=(V,E):*

```
R : P(1..|V|) := 1..N;
R' : P(1..|V|);
begin
  for R' in P(1..|V|) loop
    if |R'|≤|R|
      $\wedge \bigcup_{u\in R_{ij}} Successors^{\bullet}(G,u)\cap 1..N = i..j$
      $\wedge \bigcup_{u\in R_{ij}} Successors^{\bullet}(G,u)\cap 1..N = 0$ then
      R := R'
    end if;
  end loop;
  return R;
end;
```

The algorithm is correct for any N-RQP graph but in the case of $2^K$ -RQP graphs a refinement can be applied by filtering those R' with a cardinal greater than 2 reducing the complexity drasticly. Nevertheless, the user is just interested in the design and not in the graph so a direct constructive method that computes |Z|, U and R would be welcome. In this section a method for calculating $2^K$ -RQP designs is given..

As in the previous section, Z can be treated as a two dimensional array (where the variable Z(i, j) does not necessarily exist for all (i, j)) that is isomorphic to a one dimensional array Z' and where the isomorphism is given by an injective partial map such that $(i,j) \to i$ *when* $i=j$. .

The method presented in the following definition is the result of a deep analysis of the properties of $2^K$ -RQP graphs.

**Definition 1**

Let be the 2N-RQP with $N = 2^K$ and $K \in \mathbf{N}$ A $2^{K+1}$-RQP design (Z, U, R) is constructed in the following way:

$R_{i\ j}$ $(i \in 1..2N, j \in i..2N)$ ***is defined by the following cases:***

- ***A1.*** *If* $i = j$,

$$R_{i\ j} = \{(i, j)\} \qquad (4)$$

- ***A2.*** *For every* $l \in 1..K$,
  · *If* $i \in 1..2^{l-1}$,

$$R_{i\ 2^l} = \{(i, 2^l)\} \qquad (5)$$

*and for every* $c \in 1..(2^{K-l}-1)$ *and* $d = c2^{l+1}$,

$$R_{i+d\ \ 2^l+d} = \{(i+d, 2^l+d)\} \tag{6}$$

· *For every* $r \in 1..(l-2)$,
*If* $i \in (2^l - 2^{l-r} + 2)..(2^l - 2^{l-r-1})$,

$$R_{i\ \ 2^l} = \{(i, 2^l)\} \tag{7}$$

*and for every* $c \in 1..(2^{K-l}-1)$ *and* $d = c2^{l+1}$,

$$R_{i+d\ \ 2^l+d} = \{(i+d, 2^l+d)\} \tag{8}$$

- ***A3.*** *For every* $l \in 1..K$,
  · *If* $j \in (2^{l-1}3+1)..2^{l+1}$,

$$R_{2^l+1\ \ j} = \{(2^l+1, j)\} \tag{9}$$

*and for every* $c \in 1..(2^{K-l}-1)$ *and* $d = c2^{l+1}$,

$$R_{2^l+1+d\ \ j+d} = \{(2^l+1+d, j+d)\} \tag{10}$$

· *For every* $r \in 1..(l-2)$,
*If* $j \in (2^l + 1 + \frac{2^l}{2^{r+1}})..(2^l - 1 + \frac{2^l}{2^r})$,

$$R_{2^l+1\ \ j} = \{(2^l+1, j)\} \tag{11}$$

*and for every* $c \in 1..(2^{K-l}-1)$ *and* $d = c2^{l+1}$,

$$R_{2^l+1+d\ \ j+d} = \{(2^l+1+d, j+d)\} \tag{12}$$

- ***B1.*** *If* $i \in 1..N$ *and* $j \in N+1..2N$,

$$R_{i\ \ j} = \{(i, N), (N+1, j)\} \tag{13}$$

- ***B2.*** *For every* $l \in 1..(K-1)$,
  · *If* $i \in 1..2^l$ *and* $j \in (2^l+1)..(2^{l+1}-1)$,

$$R_{i\ \ j} = \{(i, 2^l), (2^l+1, j)\} \tag{14}$$

$$\begin{aligned} R_{2N-j+1\ \ 2N-i+1} = \{&(2N-j+1, 2N-2^l), \\ &(2N-2^l+1, 2N-i+1)\} \end{aligned} \tag{15}$$

· *If* $i \in 2..2^l$ *and* $j \in (2^l+1)..(2^{l+1}-2)$, *and for every* $c \in 1..(2^{K-l}-1)$ *and* $d = c2^{l+1}$,

$$R_{i+d\ \ j+d} = \{(i+d, 2^l+d), (2^l+d+1, j+d)\} \tag{16}$$

$U_k$ $(k \in 1..2N)$ ***is defined by the following comprehension set:***

$$U_k = \{(i,j) \ \bullet\ i \in 1..2N \ \wedge\ j \in i..2N \ \wedge\ i \le k \ \wedge\ k \le j \ \wedge\ |R_{ij}| = 1\}$$

$|Z|$***, the number of variables*** $Z(i,j)$ ***of the design is the number of*** $R_{ij}$ ***of size*** $1$***:***

$$|Z| = \sum_{|R_{ij}|=1} 1$$

## Implementation in Haskell

The following Haskell [7] program implements the constructive method given in Definition 1.

This prototype implementation has been tested for N=$2^k$ being K less or equal 25.

Given an integer K, most functions compute information of the solutions of the $2^{K+1}$-RQP: |Z|, $U_i$ and $R_{ij}$.

```
pow2 :: Integer -> Integer
pow2 0 = 1
pow2 n = 2 * (pow2 (n-1))

a1 :: Integer -> [[(Integer,Integer)]]
a1 k = [[(i,i)] | i <- [1..(pow2 (k+1))]]

a2 :: Integer -> [[(Integer,Integer)]]
a2 k = [[(i, pow2 l)]
        | l <- [1 .. k],
          i <- [1 .. pow2 (l-1)]]
       ++
       [[(i+d, pow2 l + d)]
        | l <- [1 .. k],
          i <- [1 .. pow2 (l-1)],
          c <- [1 .. pow2 (k-l) - 1],
          let d = c * pow2 (l+1)]
       ++
       [[(i, pow2 l)]
        | l <- [1 .. k],
          r <- [1 .. l-2],
          i <- [pow2 l - pow2 (l-r) + 2 .. pow2 l - pow2 (l-r-1)]]
       ++
        [[(i+d, pow2 l + d)]
         | l <- [1 .. k],
         r <- [1 .. l-2],
         i <- [pow2 l - pow2 (l-r) + 2 .. pow2 l - pow2 (l-r-1)],
         c <- [1 .. pow2 (k-l) - 1],
         let d = c * pow2 (l+1)]

a3 :: Integer -> [[(Integer,Integer)]]
a3 k = [[(pow2 l + 1, j)]
        | l <- [1 .. k],
          j <- [3 * pow2 (l-1) + 1 .. pow2 (l+1)]]
       ++
       [[(pow2 l + 1 + d, j + d)]
        | l <- [1 .. k],
          j <- [3 * pow2 (l-1) + 1 .. pow2 (l+1)],
          c <- [1 .. pow2 (k-l) - 1],
          let d = c * pow2 (l+1)]
       ++
       [[(pow2 l + 1, j)]
        | l <- [1 .. k],
          r <- [1 .. l-2],
          j <- [pow2 l + 1 + pow2 l ‘div‘ (pow2 (r+1))
                ..pow2 l - 1 + pow2 l ‘div‘ pow2 r]]
       ++
       [[(pow2 l + 1 + d, j +d)]
        | l <- [1 .. k],
          r <- [1 .. l-2],
          j <- [pow2 l + 1 + pow2 l ‘div‘ (pow2 (r+1))
```

```
                ..pow2 l - 1 + pow2 l `div` pow2 r],
          c <- [1 .. pow2 (k-l) - 1],
          let d = c * pow2 (l+1)]

r1 :: Integer -> [[(Integer,Integer)]]
r1 k = a1 k ++ a2 k ++ a3 k

b1 :: Integer -> [[(Integer,Integer)]]
b1 k = [[(i,pow2 k),(pow2 k + 1,j)]
        | i <- [1 .. pow2 k],
          j <- [pow2 k + 1 .. pow2 (k+1)]]

b2 :: Integer -> [[(Integer,Integer)]]
b2 k = [[(i,pow2 l),(pow2 l + 1,j)]
        | l <- [1..k-1],
          i <- [1..pow2 l],
          j <- [pow2 l + 1..pow2 (l+1) - 1]]
       ++
       [[(pow2 (k+1) - j + 1, pow2 (k+1) - pow2 l),
         (pow2 (k+1) - pow2 l + 1,pow2 (k+1) - i + 1)]
        | l <- [1..k-1],
          i <- [1..pow2 l],
          j <- [pow2 l + 1..pow2 (l+1) - 1]]
       ++
       [[(i+d,pow2 l + d),(pow2 l + d + 1,j+d)]
        | l <- [1..k-1],
          i <- [2..pow2 l],
          j <- [pow2 l + 1..pow2 (l+1) - 1],
          c <- [1..pow2 (k-l) - 2],
          let d = c * pow2 (l+1)]

r2 :: Integer -> [[(Integer,Integer)]]
r2 k = b1 k ++ b2 k

r :: Integer -> [[(Integer,Integer)]]
r k = r1 k ++ r2 k

u :: Integer -> [(Integer,Integer)]
u k = [(i,j) | i <- [1 .. pow2 (k+1)],
               j <- [i .. pow2 (k+1)],
               i <= k, k <= j,
               (i,j) `elem` concat (a1 k ++ a2 k ++ a3 k)]

zCard :: Integer -> Integer
zCard k = fromIntegral (length (r1 k))
```

We can prove that given a N-RQP solution (Z,U,R) obtained by applying the method in Definition 1, we have:

1. The number of program variables required is

$$|Z| = N\log_2 N - 2N + 2\log_2 N + 2$$

2. The sum of costs of all update operations is

$$\frac{N^2}{2} - \frac{N}{2}\log_2 N + \frac{3N}{2} - 2$$

3. The sum of costs of all retrieve operations is

$$N^2 + N(3 - \log_2 N) - 2\log_2 N - 2$$

4. The average complexity of the Update and Retrieve operations is constant (this is a consequence of 2 and 3 above)

## Bibliography

[1] D.J. Volper, M.L. Fredman, *Query Time Versus Redundancy Trade-offs for Range Queries*, Journal of Computer and System Sciences 23, (1981) pp.355--365.

[2] W.A. Burkhard, M.L. Fredman, D.J.Kleitman, *Inherent complexity trade-offs for range query problems*, Theoretical Computer science, North Holland Publishing Company 16, (1981) pp.279--290.

[3] M.L. Fredman, *The Complexity of Maintaining an Array and Computing its Partial Sums*, J.ACM, Vol.29, No.1 (1982) pp.250--260.

[4] A. Herranz, A. Toni, *Digraphs Definition for an Array Maintenance Problem*, Preprint.

[5] A. Toni, *Lower Bounds on Zero-one Matrices*, Linear Algebra and its Applications, 376 (2004) 275--282.

[6] A. Toni, Complejidad y Estructuras de Datos para e*l problema de los rangos variables*, Doctoral Thesis, Facultad de Informática, Universidad Politécnica de Madrid, 2003.

[7] S. P. Jones, J. Hughes, *Report on the Programming Language Haskell 98. A Non-strict Purely Functional Language,* (February 1999).

## Authors' Information

***Adriana Toni*** *– Grupo de Validacion y Aplicaciones Industriales, Facultad de Informática, Universidad Politécnica de Madrid; 28660-Boadilla del Monte, Madrid, SPAIN; e-mail: atoni@fi.upm.es*

***Ángel Herranz Nieva*** *– Assistant Professor; Departamento de Lenguajes y Sistemas Informáticos; Facultad de Informática; Universidad Politécnica de Madrid; e-mail: aherranz@fi.upm.es*

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

# THE FUZZY GROUP METHOD OF DATA HANDLING WITH FUZZY INPUT VARIABLES

## Yuriy Zaychenko

***Abstract****: The problem of constructing forecasting models with incomplete and fuzzy input data is considered in this paper. For its solution Fuzzy Group Methods of Data Handing (FGMDH) with fuzzy inputs is suggested. The method enables to construct a forecasting fuzzy model using experimental data which are not distinct.*

*The method was implemented as software kit and experimental investigations of were carried out in the problem forecasting stock-prices at the Russian stock-exchange. The comparison of the suggested method with known methods: GMDH and fuzzy GMDH are also presented.*

***Keywords****: Group method of Data Handling, fuzzy, economic indexes, forecasting*

## Introduction

The problem of forecasting models constructing using experimental data in terms of fuzziness, when input variables are not known exactly and determined as intervals of uncertainty is considered in this paper. The fuzzy group method of data handling is proposed to solve this problem. The theory of this method was suggested and researched in [1-7]. As is well known, fuzzy GMDH allows constructing fuzzy models and has the following advantages:

1. The problem of optimal model finding is transformed to the problem of linear programming, which is always solvable;
2. There is interval regression model built as the result of method work out;
3. There is a possibility of adaptation of the obtained model.

The mathematical model of the problem mentioned above is built in this article and fuzzy GMDH with fuzzy inputs is elaborated in the paper. The corresponding program, which uses the suggested algorithm, was developed. And also the experimental researches and comparison of FGMDH with GMDH and neural nets in problems of stock prices forecasting was carried out and presented in this article.

## Math model of group method of data handling with fuzzy input data

### General view of FGMDH model with fuzzy input data

Let's consider a linear interval regression model:

$$Y = A_0 Z_0 + A_1 Z_1 + \ldots + A_n Z_n , \tag{1}$$

where $A_i$ are fuzzy numbers, which are described by threes of parameters $A_i = (\underline{A_i}, \breve{A}_i, \overline{A_i})$, where $\breve{A}_i$ – interval center, $\overline{A_i}$ – upper border of the interval, $\underline{A_i}$ - lower border of the interval, and $Z_i$ – also fuzzy numbers, which are determined by parameters $(\underline{Z_i}, \breve{Z}_i, \overline{Z_i})$, $\underline{Z_i}$ - lower border, $\breve{Z}_i$ - center, $\overline{Z_i}$ - upper border of fuzzy number.

Then $Y$ – output fuzzy number, which parameters are defined as follows (in accordance with L-R numbers multiplying formulas):
Center of interval:

$$\breve{y} = \sum \breve{A}_i * \breve{Z}_i ,$$

Deviation in the left part of the membership function:

$$\breve{y} - \underline{y} = \sum (\left|\breve{A}_i\right| * (\breve{Z}_i - \underline{Z}_i) + (\breve{A}_i - \underline{A_i}) * \left|\breve{Z}_i\right|) ,$$

And lower border of the interval:

$$\underline{y} = \sum (\breve{A}_i * \breve{Z}_i - \left|\breve{A}_i\right| * (\breve{Z}_i - \underline{Z}_i) - (\breve{A}_i - \underline{A_i}) * \left|\breve{Z}_i\right|) ,$$

Thus upper border of the interval

$$\overline{y} = \sum (\left|\breve{A}_i\right| * (\overline{Z}_i - \breve{Z}_i) + \left|\breve{Z}_i\right| * (\overline{A}_i - \breve{A}_i) + \breve{A}_i * \breve{Z}_i) .$$

For the interval model to be correct, the real value of input variable Y is needed to lay in the interval got by the method workflow.

So, the general requirements to estimation linear interval model are to find such values of parameters $(\underline{A_i}, \breve{A}_i, \overline{A_i})$ of fuzzy coefficients, which allow:

a) Observed values $y_k$ lay in estimation interval for $Y_k$;

b) Total width of estimation interval is minimal.

Input data for this task is $Z_k = [Z_{ki}]_i$ - input training sample, and also $y_k$ – known output values, $k = \overline{1, M}$ , *M* – the number of observation points.

There are two cases of fuzzy values membership function used in this work:

- Triangular membership functions
- Gaussian membership functions.

Quadratic partial descriptions were chosen:

$$f(x_i, x_j) = A_0 + A_1 x_i + A_2 x_j + A_3 x_i x_j + A_4 x_i^2 + A_5 x_j^2 .$$

## FGMDH with fuzzy input data for triangular membership function

### The form of math model for triangular MF

Let's consider the linear interval regression model:

$$Y = A_0 Z_0 + A_1 Z_1 + ... + A_n Z_n ,$$

Current task contains the case of symmetrical membership function for parameters *A<sub>i</sub>*, so they can be described via pair of parameters ( $a_i , c_i$ ).

$\underline{A}_i = a_i - c_i$ , $\overline{A}_i = a_i + c_i$ , $c_i$ – interval width, $c_i$ ≥ *0*, $z_i$ – also fuzzy numbers of triangular shape, which are defined by parameters $(\underline{Z_i}, \breve{Z}_i, \overline{Z_i})$ , $\underline{Z_i}$ - lower border, $\breve{Z}_i$ - center, $\overline{Z_i}$ - upper border of fuzzy number.

Then *Y* – fuzzy number, which parameters are defined as follows:

Center of the interval:

$$\breve{y} = \sum a_i * \breve{Z}_i ,$$

Deviation in the left part of the membership function:

$$\breve{y} - \underline{y} = \sum(a_i * (\breve{Z}_i - \underline{Z}_i) + c_i |\breve{Z}_i|) \text{, thus}$$

Lower border of the interval: $\underline{y} = \sum(a_i * \underline{Z}_i - c_i |\breve{Z}_i|)$

Deviation in the right part of the membership function:

$$\overline{y} - \breve{y} = \sum(a_i * (\overline{Z}_i - \breve{Z}_i) + c_i |\breve{Z}_i|) = \sum a_i \overline{Z}_i - a_i \breve{Z}_i + c_i |\breve{Z}_i| \text{, so}$$

Upper border of the interval: $\overline{y} = \sum(a_i * \overline{Z}_i + c_i |\breve{Z}_i|)$

For the interval model to be correct, the real value of input variable Y should lay in the interval got by the method workflow.

It can be described in such a way:

$$\begin{cases} \sum (a_i * \underline{Z}_{ik} - c_i |\breve{Z}_{ik}|) \le y_k \\ \sum (a_i * \overline{Z}_{ki} + c_i |\breve{Z}_{ik}|) \ge y_k , k = \overline{1, M} \end{cases}$$

Where $z_k = [z_{ki}]_i$ is input training sample, $y_k$ –known output values, $k = \overline{1, M}$ , *M* – number of observation points.

So, the general requirements to estimation linear interval model are to find such values of parameters $(a_i , c_i)$ of fuzzy coefficients, which enable:

a) Observed values $y_k$ lay in estimation interval for $Y_k$ ;

b) Total width of estimation interval is minimal.

These requirements can be redefined as a task of linear programming:

$$\min_{a_i, c_i} \sum_{k=1}^{M} (\sum (a_i * \overline{Z}_i + c_i |\breve{Z}_i|) - \sum (a_i * \underline{Z}_i - c_i |\breve{Z}_i|)), \tag{2}$$

under constraints:

$$\begin{cases} \sum (a_i * \underline{Z}_{ik} - c_i |\breve{Z}_{ik}|) \le y_k \\ \sum (a_i * \overline{Z}_{ki} + c_i |\breve{Z}_{ik}|) \ge y_k , k = \overline{1, M} \end{cases}. \tag{3}$$

**Formalized problem formulation in case of triangular membership functions**

Let's consider partial description

$$f(x_i, x_j) = A_0 + A_1 x_i + A_2 x_j + A_3 x_i x_j + A_4 x_i^2 + A_5 x_j^2 . \tag{4}$$

Rewriting it in accordance with the model (1) needs such substitution: $z_0 = 1$, $z_1 = x_i$, $z_2 = x_j$, $z_3 = x_i x_j$, $z_4 = x_i^2$, $z_5 = x_j^2$.

Then math model (2)-(3) will take the form

$$\begin{aligned} \min_{a_i, c_i} \; & (2Mc_0 + a_1 \sum_{k=1}^{M} (\overline{x}_{ik} - \underline{x}_{ik}) + 2c_1 \sum_{k=1}^{M} |\breve{x}_{ik}| + a_2 \sum_{k=1}^{M} (\overline{x}_{jk} - \underline{x}_{jk}) + 2c_2 \sum_{k=1}^{M} |\breve{x}_{jk}| + \\ & + a_3 \sum_{k=1}^{M} (|\breve{x}_{ik}|(\overline{x}_{jk} - \underline{x}_{jk}) + |\breve{x}_{jk}|(\overline{x}_{ik} - \underline{x}_{ik})) + 2c_3 \sum_{k=1}^{M} |\breve{x}_{ik} \breve{x}_{jk}| + 2a_4 \sum_{k=1}^{M} |\breve{x}_{ik}|(\overline{x}_{ik} - \underline{x}_{ik}) + , \\ & + 2c_4 \sum_{k=1}^{M} \breve{x}_{ik}^2 + 2a_5 \sum_{k=1}^{M} |\breve{x}_{jk}|(\overline{x}_{jk} - \underline{x}_{jk}) + 2c_5 \sum_{k=1}^{M} \breve{x}_{jk}^2 ) \end{aligned}$$

with the following conditions:

$$\begin{aligned} & a_0 + a_1 \underline{x}_{ik} + a_2 \underline{x}_{jk} + a_3 (-|\breve{x}_{ik}|(\breve{x}_{jk} - \underline{x}_{jk}) - |\breve{x}_{jk}|(\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik} \breve{x}_{jk}) + \\ & + a_4 (-2|\breve{x}_{ik}|(\breve{x}_{ik} - \underline{x}_{ik}) + \breve{x}_{ik}{}^2) + a_5 (2|\breve{x}_{jk}|(\breve{x}_{jk} - \underline{x}_{jk}) + \breve{x}_{jk}{}^2) - c_0 - c_1 |\breve{x}_{ik}| - \\ & - c_2 |\breve{x}_{jk}| - c_3 |\breve{x}_{ik} \breve{x}_{jk}| - c_4 \breve{x}_{ik}^2 - c_5 \breve{x}_{jk}^2 \le y_k \\ & a_0 + a_1 \overline{x}_{ik} + a_2 \overline{x}_{jk} + a_3 (|\breve{x}_{ik}|(\overline{x}_{jk} - \breve{x}_{jk}) + |\breve{x}_{jk}|(\overline{x}_{ik} - \breve{x}_{ik}) - \breve{x}_{ik} \breve{x}_{jk}) + a_4 (2|\breve{x}_{ik}|(\overline{x}_{ik} - \\ & - \breve{x}_{ik}) - \breve{x}_{ik}{}^2) + a_5 (2|\breve{x}_{jk}|(\overline{x}_{jk} - \breve{x}_{jk}) - \breve{x}_{jk}{}^2) + c_0 + c_1 |\breve{x}_{ik}| + c_2 |\breve{x}_{jk}| + c_3 |\breve{x}_{ik} \breve{x}_{jk}| + \quad , \\ & c_4 \breve{x}_{ik}^2 + c_5 \breve{x}_{jk}^2 \ge y_k \\ & c_l \ge 0, \;\; l = \overline{0,5} . \end{aligned}$$

As we can see, this is the linear programming problem, but there are still no limitations for non-negativity of variables $a_i$, so we need go to dual problem, introducing dual variables $\{\delta_k\}$ and $\{\delta_{k+M}\}$.

Write down dual problem:

$$\max (\sum_{k=1}^{M} y_k \cdot \delta_{k+M} - \sum_{k=1}^{M} y_k \cdot \delta_k), \tag{5}$$

under constraints:

$$\sum_{k=1}^{M}\delta_{k+M}-\sum_{k=1}^{M}\delta_{k}=0,$$

$$\sum_{k=1}^{M}\overline{x}_{ik}\cdot\delta_{k+M}-\sum_{k=1}^{M}\underline{x}_{ik}\cdot\delta_{k}=\sum_{k=1}^{M}(\overline{x}_{ik}-\underline{x}_{ik})$$

$$\sum_{k=1}^{M}\overline{x}_{jk}\cdot\delta_{k+M}-\sum_{k=1}^{M}\underline{x}_{jk}\cdot\delta_{k}=\sum_{k=1}^{M}(\overline{x}_{jk}-\underline{x}_{jk})$$

$$\begin{aligned}&\sum_{k=1}^{M}(|\breve{x}_{ik}|(\overline{x}_{jk}-\breve{x}_{jk})+|\breve{x}_{jk}|(\overline{x}_{ik}-\breve{x}_{ik})-\breve{x}_{ik}\breve{x}_{jk})\cdot\delta_{k+M}-\\&-\sum_{k=1}^{M}(-|\breve{x}_{ik}|(\breve{x}_{jk}-\underline{x}_{jk})-|\breve{x}_{jk}|(\breve{x}_{ik}-\underline{x}_{ik})+\breve{x}_{ik}\breve{x}_{jk})\cdot\delta_{k}=\\&=\sum_{k=1}^{M}(|\breve{x}_{ik}|(\overline{x}_{jk}-\underline{x}_{jk})+|\breve{x}_{jk}|(\overline{x}_{ik}-\underline{x}_{ik}))\end{aligned}\tag{6}$$

$$\sum_{k=1}^{M}(2|\breve{x}_{ik}|(\overline{x}_{ik}-\breve{x}_{ik})-\breve{x}_{ik}{}^{2})\cdot\delta_{k+M}-\sum_{k=1}^{M}(-2|\breve{x}_{ik}|(\breve{x}_{ik}-\underline{x}_{ik})+\breve{x}_{ik}{}^{2})\cdot\delta_{k}=\sum_{k=1}^{M}|\breve{x}_{ik}|(\overline{x}_{ik}-\underline{x}_{ik})$$

$$\sum_{k=1}^{M}(2|\breve{x}_{jk}|(\overline{x}_{jk}-\breve{x}_{jk})-\breve{x}_{jk}{}^{2})\cdot\delta_{k+M}-\sum_{k=1}^{M}(-2|\breve{x}_{ik}|(\breve{x}_{jk}-\underline{x}_{jk})+\breve{x}_{jk}{}^{2})\cdot\delta_{k}=\sum_{k=1}^{M}|\breve{x}_{jk}|(\overline{x}_{jk}-\underline{x}_{jk})$$

$$\sum_{k=1}^{M}\delta_{k+M}+\sum_{k=1}^{M}\delta_{k}\le 2M$$

$$\sum_{k=1}^{M}|\breve{x}_{ik}|\cdot\delta_{k+M}+\sum_{k=1}^{M}|\breve{x}_{ik}|\cdot\delta_{k}\le 2\sum_{k=1}^{M}|\breve{x}_{ik}|$$

$$\sum_{k=1}^{M}|\breve{x}_{jk}|\cdot\delta_{k+M}+\sum_{k=1}^{M}|\breve{x}_{jk}|\cdot\delta_{k}\le 2\sum_{k=1}^{M}|\breve{x}_{jk}|$$

$$\sum_{k=1}^{M}|\breve{x}_{ik}\breve{x}_{jk}|\cdot\delta_{k+M}+\sum_{k=1}^{M}|\breve{x}_{ik}\breve{x}_{jk}|\cdot\delta_{k}\le 2\sum_{k=1}^{M}|\breve{x}_{ik}\breve{x}_{jk}|,\tag{7}$$

$$\sum_{k=1}^{M}\breve{x}_{ik}^{2}\cdot\delta_{k+M}+\sum_{k=1}^{M}\breve{x}_{ik}^{2}\cdot\delta_{k}\le 2\sum_{k=1}^{M}\breve{x}_{ik}^{2}$$

$$\sum_{k=1}^{M}\breve{x}_{jk}^{2}\cdot\delta_{k+M}+\sum_{k=1}^{M}\breve{x}_{jk}^{2}\cdot\delta_{k}\le 2\sum_{k=1}^{M}\breve{x}_{jk}^{2}$$

$$\delta_{k}\ge 0,$$

$$\delta_{k+M}\ge 0,\ k=\overline{1,M}.$$

The task (5)-(7) can be solved using simplex-method. Having optimal values of dual variables $\{\delta_k\}$, $\{\delta_{k+M}\}$, we easily obtain the optimal values of desired variables $c_i$, $a_i$, $i = \overline{0,5}$, and also a desired fuzzy model for given partial description.

## Result of FGMDH with fuzzy input data workflow in RTS index forecasting

For estimation of efficiency of the suggested FGMDH method with fuzzy inputs the corresponding software kit was elaborated and numerous experiments of financial markets forecasting were carried out. Some of them are presented below.

### Forecasting of RTS index.

*Experiment 1. RTS index forecasting (opening price)*

In this experiment we used 5 fuzzy input variables, which represent stock prices of leading Russian energetic companies, which are included to the list of computations of RTS index:

LKOH – shares of "LUKOIL" joint-stock company,

EESR – shares of "РАО ЕЭС России"joint-stock company,

YUKO – shares of "ЮКОС" joint-stock company,

SNGSP – privileged shares of "Сургутнефтегаз" joint-stock company,

SNGS – common shares of "Сургутнефтегаз" joint-stock company.

Output variable is the RTS (opening price) index value of the same period (03.04.2006 – 18.05.2006).

Sample size – 32 values.

Training sample size – 18 values (optimal size of training sample for current experiment).

The following results were obtained:

1. For triangular membership function

a) For normalized input data

Criterion value for current experiment were: MSE = 0.055557

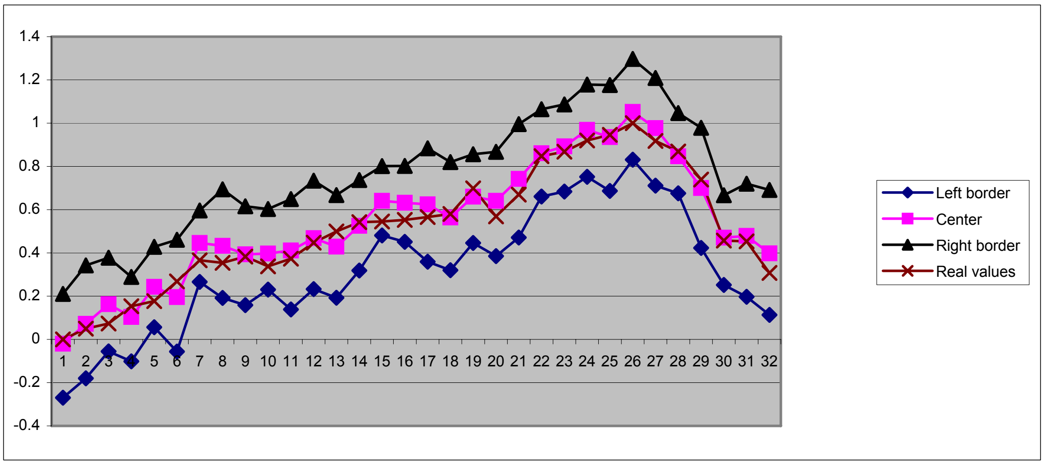


Fig.1. Experiment 1 results for triangular membership function and normalized values of input variables

2. For the case of Gaussian membership function (optimal level is α=0.8)

a) For normalized input data

Criterion values for this experiment were: MSE = 0.028013

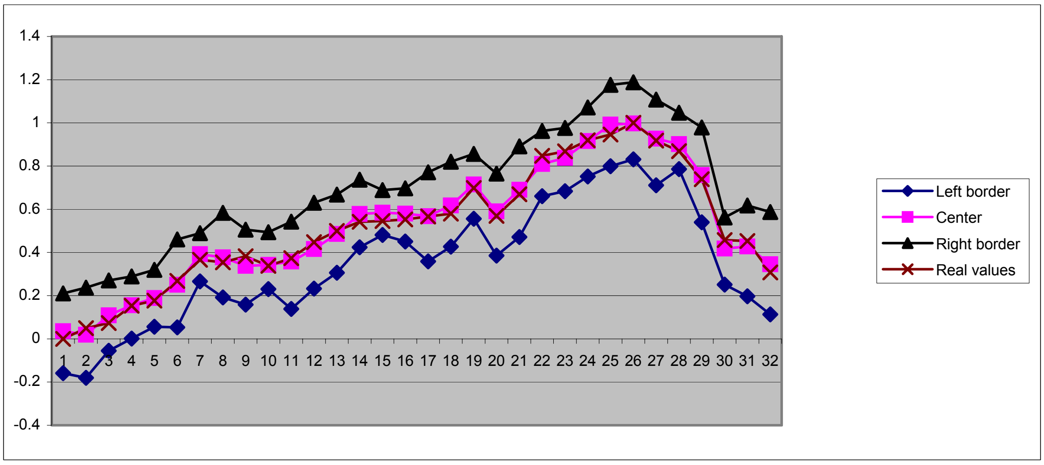


Fig.2. Experiment 1 result for Gaussian MF and normalized input data

b) for non-normalized input data:

Criterion values for current experiment were: MSE = 9.321461 MAPE = 0.4%

As we can see from the results of experiment 1, forecasting using triangular and Gaussian membership functions gives good results. Results of experiments with Gaussian MF are better than results of experiments with triangular MF.

For normalized data:

| | Triangular MF | Gaussian MF |
|---|---|---|
| MSE | 0.055557 | 0.028013 |

For non-normalized data:

| | Triangular MF | Gaussian MF |
|---|---|---|
| MSE | 18.48657 | 9.321461 |
| MAPE | 0.8% | 0.4% |

*Experiment 2. Forecasting of RTS index (closing price)*

This experiment uses the same input variables as the experiment 1 does.

Output variable is the value of RTS index (closing price) for the same period (03.04.2006 – 18.05.2006).

Sample size – 32 values.

Training sample size – 18 values (optimal size of training sample for current experiment).

The following results were obtained:

1. For triangular membership function

a) For normalized input data

Criterion value: MSE = 0.057379

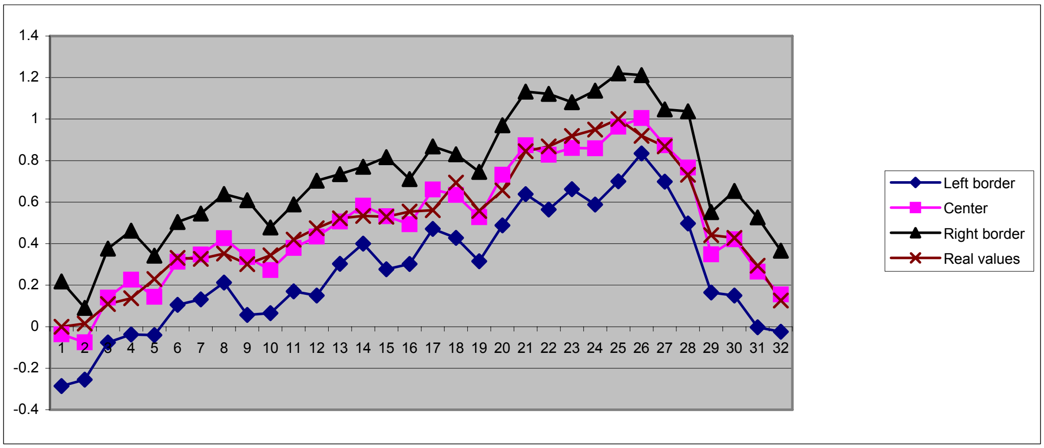


Fig.3. Experiment 2 result for triangular MF and normalized values of input variables

b) For non-normalized input data

Criterion values: MSE = 18.04394 MAPE =0.78%

1. For Gaussian membership function (optimal level α=0.85)

a) For normalized input data

Criterion value for current experiment was: MSE = 0.029582

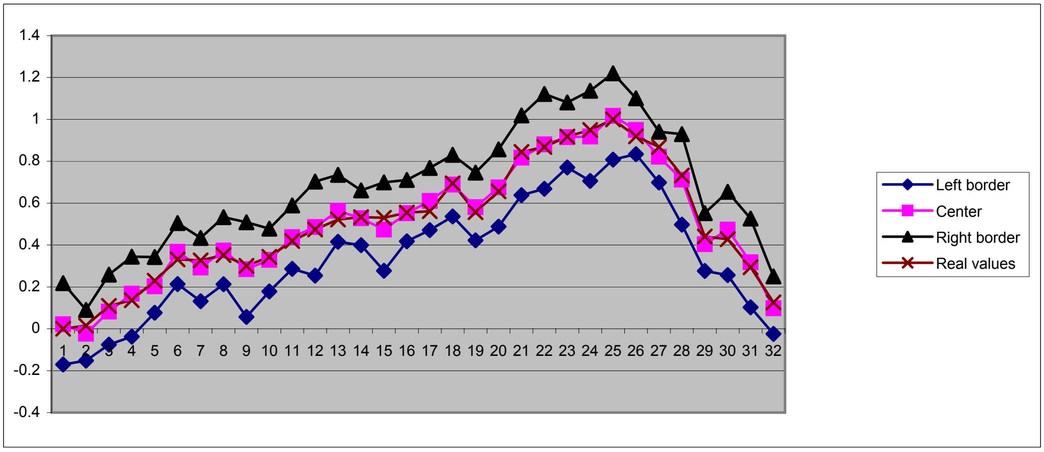


Fig.4. Experiment 2 result for Gaussian MF and normalized values of input variables

b) For non-normalized input data

Criterion values for this experiment: MSE = 9.302766; MAPE =0.37%.

As we can see from the results of experiment 2, forecasting using triangular and Gaussian membership functions gives good results. Results of experiments with Gaussian MF are better than results of experiments with triangular MF.

For normalized data:

| | Triangular MF | Gaussian MF |
|---|---|---|
| MSE | 0.057379 | 0.029582 |
| | | |

For non-normalized data:

| | Triangular MF | Gaussian MF |
|---|---|---|
| MSE | 18.04394 | 9.302766 |
| MAPE | 0.78% | 0.37% |

### Stock price forecasting

The following experiment uses stock prices of 4 leading energetic companies of Russia:

EESR – shares of “РАО ЕЭС России” joint-stock company,

YUKO – shares of “ЮКОС” joint-stock company,

SNGSP – privileged shares of “Сургутнефтегаз” joint-stock company,

SNGS – ordinary shares of “Сургутнефтегаз” joint-stock company.

Stock price of other company – “LUKOIL” joint-stock for the same period (03.04.2006 – 18.05.2006) was also forecasted.

Sample size – 32 values. Training sample size – 17 values (optimal size of training sample for this experiment).

The following results were obtained:

1. For triangular membership function. For normalized input data: Criterion value: MSE=0.056481

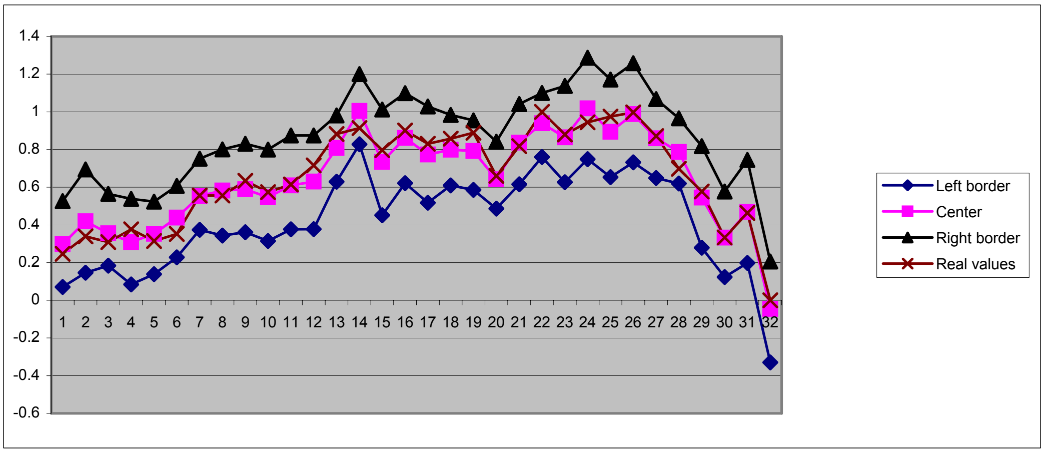


Fig.5. Experiment 4 results for triangular MF and normalized values of input variables

b) for non-normalized input data:

Criterion values: MSE = 0.914998; MAPE = 0.73%

2. For Gaussian membership function (optimal level of α=0.9)

a) For normalized input data:

Criterion value for this experiment: MSE = 0.030464

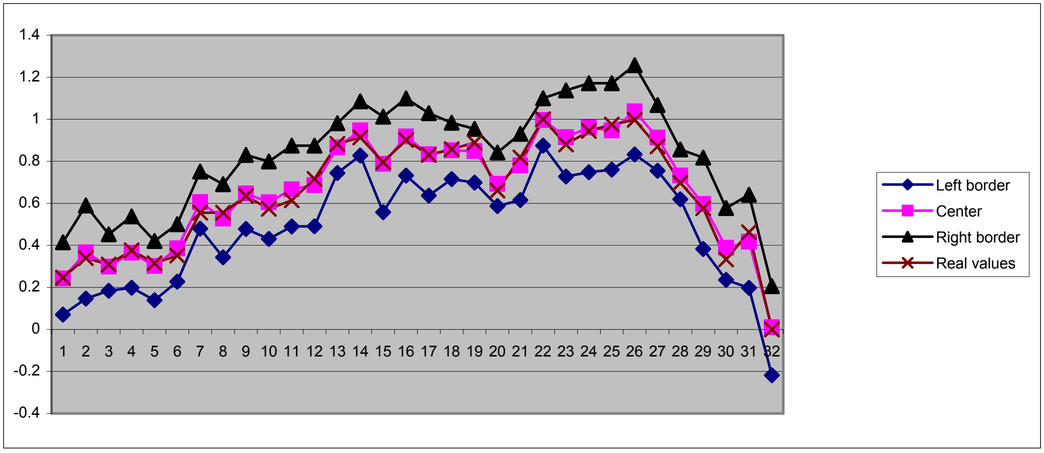


Fig.6. Experiment 4 results for Gaussian MF and normalized values of input variables

As we can see from the results of experiment 4, forecasting using triangular and Gaussian membership functions gives good results. Results of experiments with Gaussian MF are better than results of experiments with triangular MF.

For normalized data:

| | Triangular MF | Gaussian MF |
|---|---|---|
| MSE | 0.056481 | 0.030464 |
| | | |

For non-normalized data:

| | Triangular MF | Gaussian MF |
|---|---|---|
| MSE | 0.914998 | 0.493511 |
| MAPE | 0.73% | 0.33% |

## The comparison of GMDH, FGMDH and FGMDH with fuzzy input data

In the next experiments the comparison of the suggested method FGMDH with fuzzy inputs with known methods: classical GMDH and Fuzzy GMDH was performed

*Experiment 1. Forecasting of RTS index (opening price)*

Current experiment contains 5 fuzzy input variables, which are the stock prices of leading Russian energetic companies included into the list of RTS index calculation:

Output variable is the value of RTS index (opening price) of the same period (03.04.2006 – 18.05.2006).

Sample size – 32 values.

Training sample size – 18 values (optimal size of the training sample for current experiment).

The following results were obtained:

For normalized input when using Gaussian MF in group method of data handling with fuzzy input data:

For normalized values in GMDH and FGMDH:

MSE for GMDH = 0,1129737  MSE for FGMDH = 0,0536556

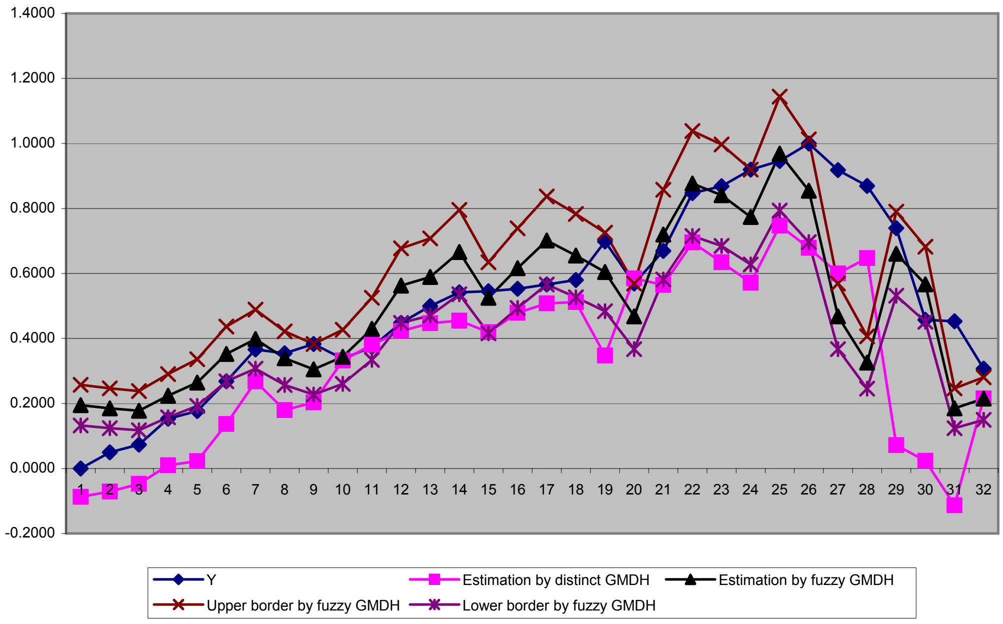


Fig.7. Experiment 1 results using GMDH and FGMDH

As the results of experiment 1 show, fuzzy group method of data handling with fuzzy input data gives more accurate result that GMDH with triangular membership function or Gaussian membership function. In case of triangular MF FGMDH with fuzzy data gives a little worse than FGMDH with Gaussian MF.

Table 1. MSE comparison for different methods of experiment 1

| | GMDH | FGMDH | FGMDH with fuzzy inputs, Triangular MF | FGMDH with fuzzy inputs, Gaussian MF |
|---|---|---|---|---|
| MSE | 0,1129737 | 0,0536556 | 0,055557 | 0,028013 |

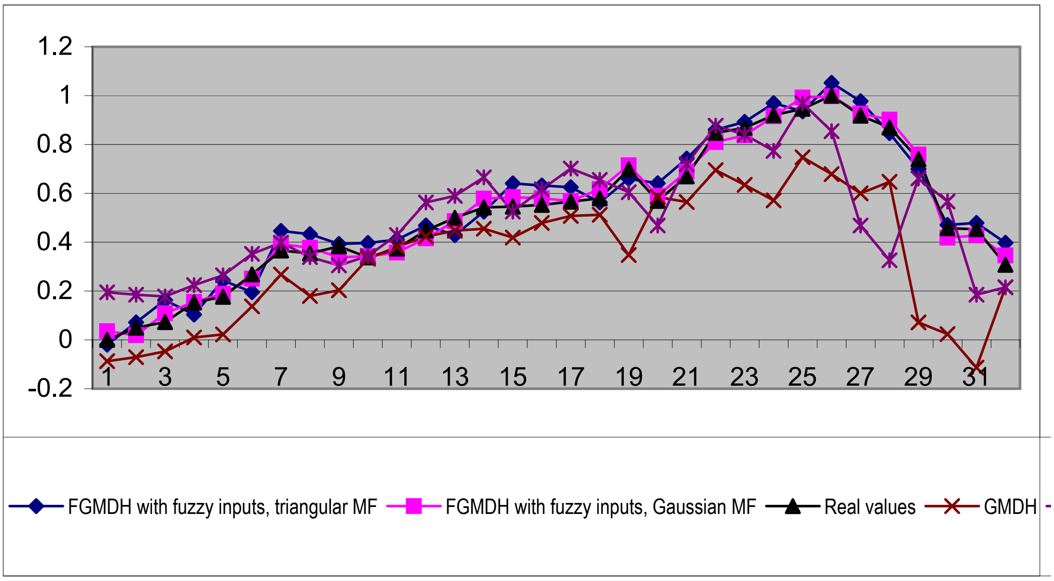


Fig. 8. GMDH, FGMDH (center of estimation), and FGMDH with fuzzy inputs (center of estimation) result comparison

*Experiment 2. RTS-2 index forecasting (opening price)*

Sample size – 32 values.

Training sample size – 19 values (optimal size of training sample for current experiment).

The following results were obtained:

For normalized input data when using triangular MF in group method of data handling with fuzzy input data: MSE = 0,061787

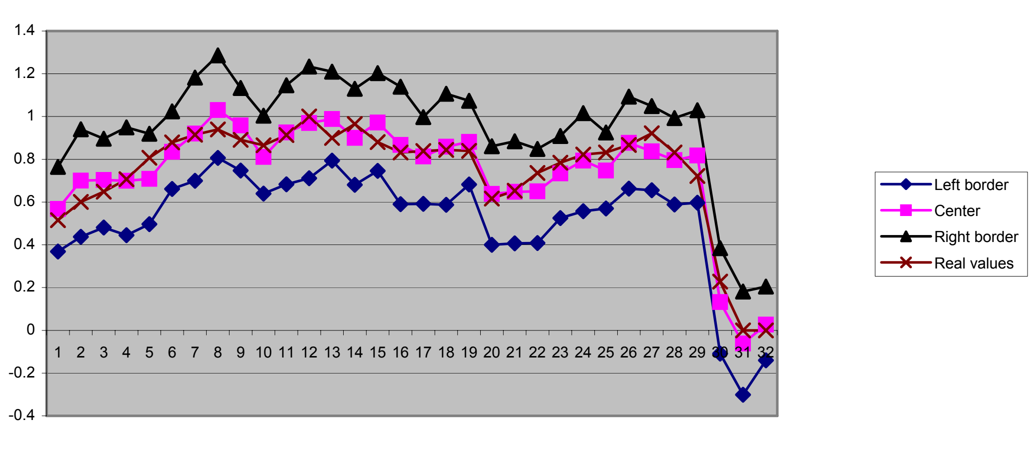


Fig.9. Experiment 2 results for triangular MF

For normalized input data using Gaussian MF in fuzzy group method of data handling with fuzzy input data:
For normalized values in GMDH and FGMDH method the following results were obtained:
MSE for GMDH = 0,051121;  MSE for FGMDH = 0,063035.

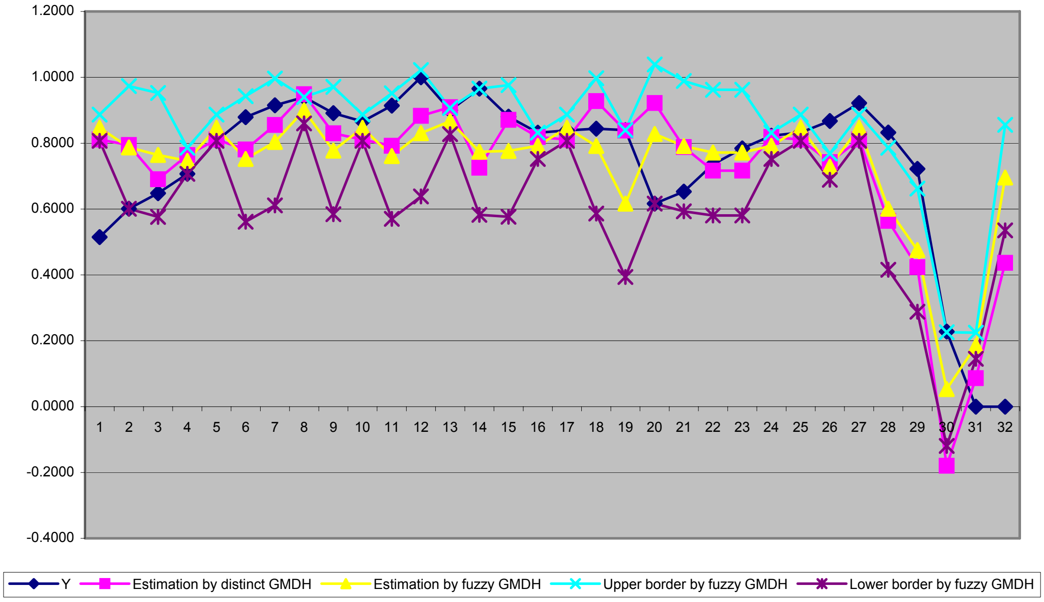


Fig.10. Experiment 2 results using GMDH and FGMDH

As the results of the experiment 2 show, fuzzy group method of data handling with fuzzy input data gives better result than GMDH and FGMDH in case of Gaussian membership function. In this example GMDH gives better results, than FGMDH and GMDH with fuzzy input data in case of triangular membership function.

Table 2. MSE of different methods of experiment 2 comparison

| | GMDH | FGMDH | FGMDH with fuzzy inputs, triangular MF | FGMDH with fuzzy inputs, Gaussian MF |
|---|---|---|---|---|
| MSE | 0,051121 | 0,063035 | 0,061787 | 0,033097 |

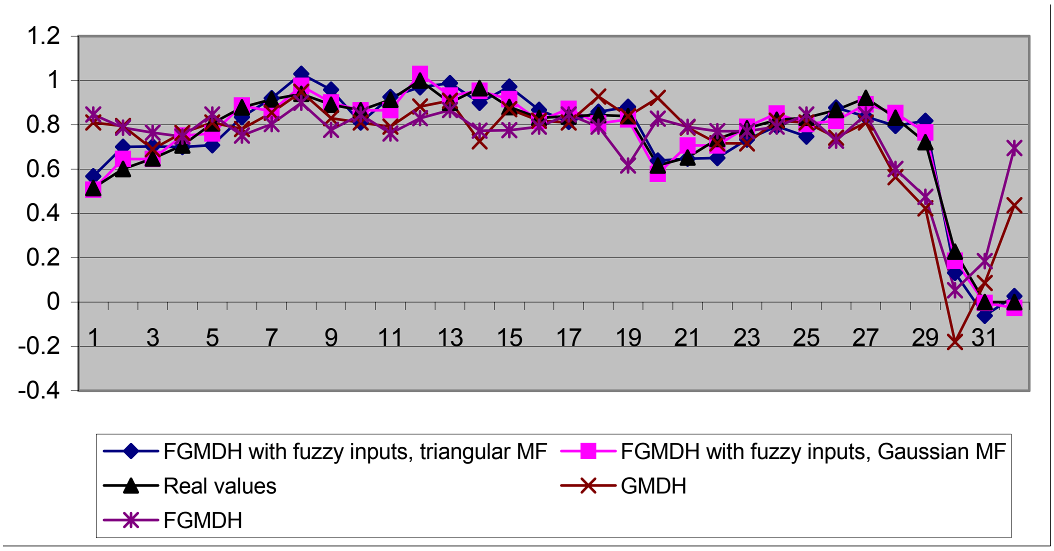


Fig.11. GMDH, FGMDH (center of estimation), and FGMDH
with fuzzy inputs (center of estimation) result comparison

## Conclusion

In this article new method of inductive modeling FGMDH with fuzzy inputs was suggested. This method represents the development of fuzzy GMDH when information is fuzzy and given in the form of uncertainty intervals. The mathematical model was constructed and corresponding algorithm was elaborated. The experimental results of application of the suggested method in the forecasting of market index and stock prices are presented and discussed. The comparison of the suggested method with classical GMDH and Fuzzy GMDH were performed and presented. The main advantages of the suggested method are following:

- It operates with fuzzy and uncertain input information and constructs the fuzzy model;
- The constructed model has minimal possible total width and in this sense it is optimal;
- For finding optimal model we solve corresponding linear programming problem which is always solvable for this task;
- We should not a priori set the form of a model the algorithm finds it itself using the ideas of evolution.

## References

1. Zaychenko Yu. "The Fuzzy Group Method of Data Handling and Its Application for Economical Processes Forecasting" - *Scientific Inquiry,* - Vol. 7, No.1, June, 2006 - p.83-96.
2. Zaychenko Yu. "Fuzzy method of inductive modeling in problems of macroeconomic indexes forecasting." *System researches and informational technologies,* #3 of 2003, p. 25-45.
3. Zaychenko Yu. P., and Zayetz I.O. "The synthesis and adaptation of fuzzy forecasting model based of self-organization method. *Science News of NTUU "KPI",* #2 of 2001.
4. Zaychenko Yu. P., and Zayetz I.O. "Research of different types of partial descriptions in problems of synthesis of fuzzy forecasting models", *Science works of Donetsk NTU,* vol. 47, p. 341-349.
5. Zaychenko Yu. P., Zayetz I.O., O.V. Kamotsky, O.V. Pavlyuk. Research of different kinds of membership functions of fuzzy forecasting models parameters in fuzzy group method of data handling. *USiM,* 2003, #2, p.56-67.
6. Zaychenko Yu. P. and Zayetz I.O. Comparative analysis of GMDH algorithms using different method of single-step adaptation of coefficients. *The NTUU Herald.*
7. Zaychenko Yu. P. Comparative analysis of forecasting models built using distinct and fuzzy GMDH with different algorithms of fuzzy forecasting models generation. *Materials of international seminar of inductive modeling IWIM 2005.*

## Author's Information

***Zaychenko Yuri** – professor NTUU "Kiev Polytechnic Institute", Institute of Applied System Analysis, Peremogy avenue 37, 03056, Kiev-56, Ukraine, phone: 8-044-2418693, e-mail: zaych@i.com.ua*

# IDENTIFICATION AND OPTIMAL CONTROL OF SYSTEM DESCRIBED BY QUASILINEAR PARABOLIC EQUATIONS

## Mahmoud Farag

***Abstract**: This paper treats the problem of control of a quasilinear parabolic equation with controls in the coefficients, in the boundary conditions and in the right side of the equation. The difference approximations problem is constructed. The minimization of the modified cost function is found by applying the penalty method combined with PQI method. The problem of identification of the unknown coefficients for a heat exchange process is solved numerically. Numerical results are reported.*

***Keywords**: Optimal control, parabolic equations, Finite difference method, Stability theory, PQI method.*

## 1. Introduction

An interesting and well investigated problem is the identification of coefficients in partial differential equations. In constract to this, the identification of nonlinear phenomena is less developed. This refers also to the nonlinear boundary conditions for the heat equation. The technical background consists in the identification of the heat exchange coefficient in our case. There are many papers dealing with the identification problem. Methods of solving of these problems are listed in the following:

1) Gradient or conjugate gradient methods, for example, [Farag, 2006].

2) The lagrangian method, for example, [Arada, 2003].

3) The regularization methods, for example, [Lur'e, 1995].

This paper treats the problem of control of a quasilinear parabolic equation with controls in the coefficients, in the boundary conditions and in the right side of the equation. The difference approximations problem is constructed. The minimization of the modified cost function is found by applying the penalty method combined with PQI method [El-Gendi,1995]. The problem of identification of the unknown coefficients for a heat exchange process is solved numerically. Numerical results are reported.

## 2. Control Problem of Heat Equation

Let D be a bounded domain of the N-dimensional Euclidean space $E_N$, $l$ , $T$ be given positive numbers, let $\Omega = \{(x,t) : x \in D , t \in\}$ . Let $V = \{v : v = v_1, v_2, \ldots, v_N \in E_N, \|v\|_{E_N} \le R\}$ where R > 0 are given numbers. We consider the heat exchange process described by the equation

$$(1) \qquad \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(\lambda(u,v)\frac{\partial u}{\partial x}\right) = f(x,t,u,v) \ , \ (x,t) \in \Omega$$

with initial and boundary conditions

$$(2) \qquad u(x,t)\big|_{t=0} = \phi(x) \ , \ x \in D$$

$$(3) \qquad \lambda(u,v)\left.\frac{\partial u}{\partial x}\right|_{x=0} = Y_0(t), \ \lambda(u,v)\left.\frac{\partial u}{\partial x}\right|_{x=l} = Y_1(t) \ , \ t \in [0,T],$$

where $\phi(x)$, $Y_0(t)$ , $Y_1(t) \in L_2(0,T)$. Besides, the functions $\lambda(u,v)$, $f(u,v)$ are continuous on $(u,v) \in [r_1, r_2] \times E_N$ , have continuous derivatives in u and $\forall\ (u,v) \in [r_1, r_2] \times E_N$ the derivatives $\frac{\partial \lambda(u,v)}{\partial u}, \frac{\partial f(u,v)}{\partial u}$ are bounded and $[r_1, r_2$ are given numbers.

On The set $V$ ,under the conditions (1)-(3) and additional restrictions

$$(4) \qquad \xi_0 \le \lambda(u,v) \le \mu_0 \ , \quad r_1 \le u(x,t) \le r_2$$

is required to minimized the functional

$$(5) \qquad J_\alpha(v) = \beta_0 \int_0^T [u(0,t) - y_0(t)]^2 + \beta_1 \int_0^T [u(l,t) - y_1(t)]^2 + \alpha\ \|v - \omega\|_{E_N}^2$$

where $y_0(t)$ , $y_1(t) \in L_2(0,T)$ are given numbers, $\alpha \ge 0, \beta_i \ge 0, i = \overline{0,2}$ , $\beta_0 + \beta_1 \ne 0$ are given numbers, $\omega = \{\omega_1, \omega_2, \ldots, \omega_N\} \in E_N$ is also given.

**Definition 2.1:** The solution of problem (1)-(4) is the function $u(x,t) \in V_2^{1,0}(\Omega)$ and satisfies the integral

$$(6)\int_0^l\int_0^T [u\frac{\partial \eta}{\partial t} - \lambda(u,v)\frac{\partial u}{\partial x}\frac{\partial \eta}{\partial x} + \eta\ f(x,t,u,v)\ ]\ dx\ dt$$

$$= -\int_0^l \phi(x)\ \eta(x,0)\ dx - \int_0^T Y_0(x)\ \eta(0,t)\ dt - \int_0^T Y_1(x)\ \eta(l,t)\ dt$$

$\forall\ \eta = \eta(x,t) \in W_2^{1,1}(\Omega)$ *and* $\eta(x,T) = 0.$

In [Farag, 2002], the existence and uniqueness of the solution of problem (1)-(5) are studied. Optimal control problems of the coefficients of differential equations do not always have solution [Tikhonov, 1974].

## 3. Modified Optimal Control Problem

It is well known that the penalty function methods are very effective techniques for solving constrained optimization problems via unconstrained problems. In recent years, these methods have been widely used to solve constrained optimal control problems. Applications of the exterior, interior and mixed penalty function methods in constrained optimal control problems can then be found. The inequality constrained problem (1) through (5) is converted to a problem without inequality constraints by adding a penalty function [Ibeijuba, 1983] to the objective (5), yielding the following $\Phi_{\alpha,s}(v,A_s)$ function:

$$(7)\qquad \Phi_{\alpha,s}(v,A_s) = J_\alpha(v) + P_s(v)$$

$$P_s(v) = A_s\int_0^l\int_0^T [\ F(u,v) + Q(u) + B(u)]\ dx\ dt$$

$$F(u,v) = [\ \max\ \{\xi_0 - \lambda(u,v),0\}]^2 + [\ \max\ \{\lambda(u,v) - \mu_0,0\}]^2$$

$$Q(u) = [\ \max\ \{r_0 - u(x,t;v),0\}]^2,\ B(u) = [\ \max\ \{u(x,t;v) - r_2,0\}]^2$$

where $A_s,\ s = 1,2,\ldots$ are positive numbers, $\lim_{s\to\infty} A_s = +\infty$ .

## 4. The Discrete Control Problem

Difference methods of solution of optimal control problems for partial differential equations are investigated comparatively small [Potapov,1978]and [Tagiev, 1982].

In this section, we shall find the discrete control problem for (7), (1)-(3) and two theorems prove the estimates of stability for the difference approximations problem (22)-(25) and an estimate on $v$ . From [Ladyzenskaya, 1973], we give the net norms for the net functions $[Z] = \{((Z_1)_i^j,\ldots,((Z_m)_i^j\},\ i = \overline{0,N},\ j = \overline{0,M}$ with m components: $\|Z\|_{L_2(\overline{\omega}_{h\tau})} = [h\tau\sum_{i=0}^{N-1}\sum_{j=1}^{M}(Z_i^j)^2\ ]^{\frac{1}{2}}$, $\max_j \|Z\|_{L_2(\overline{\omega}_h)} = [h\sum_{i=0}^{N-1}(Z_i^j)^2\ ]^{\frac{1}{2}}$, for the net functions $[Z] = \{((Z_1)_i,\ldots,((Z_m)_i\},i = \overline{0,N}$ the norm is $\|Z\|_{L_2(\overline{\omega}_h)} = [h\sum_{i=0}^{N-1}(Z_i)^2\ ]^{\frac{1}{2}}$ and also for the net functions $[Z] = \{((Z_1)^j,\ldots,((Z_m)^j\},\ j = \overline{0,M}$ the norm is $\|Z\|_{L_2(\overline{\omega}_\tau)} = [\tau\sum_{j=1}^{M}(Z_j)^2\ ]^{\frac{1}{2}}$ .

For discretization the optimal control problem (1)-(3),(7) in $\overline{\Omega}$ we introduce the net $\overline{\omega}_{h\tau} = \overline{\omega}_h \times \overline{\omega}_\tau$ where $\overline{\omega}_h = \{x_i = i\,h,\ i = \overline{0,N},\ N\ h = l\},\ \overline{\omega}_\tau = \{t_j = j\,\tau,\ i = \overline{0,M},\ M\tau = T\}$

Here and further for arbitrary net functions $u = u_i^j = u(x,t) = u(x_i, t_j), x = x_i \in \overline{\omega}_h, t = t_j \in \overline{\omega}_\tau$ adopt denotations [Ladyzenskaya, 1973]:

$$\hat{u} = u_i^{j+1} = u(x_i, t_{j+1}), u^* = u_i^{j-1} = u(x_i, t_{j-1}), u^- = u_{i-1}^j = u(x_{i-1}, t_j)$$

$$u^+ = u_{i+1}^j = u(x_{i+1}, t_j), u_x = \frac{u^+ - u}{h}, u_{\bar{x}} = \frac{u - u^-}{h}, u_t = \frac{\hat{u} - u}{\tau}, u_{\bar{x}} = \frac{u - u^*}{\tau}$$

The given functions in (6) approximate as follows:

$$\lambda_i^j = \frac{1}{h\tau} \int_{t_{j-1}}^{t_j} \int_{x_i}^{x_{i+1}} \lambda(u(x,t),v)\, dx\, dt\,,\ i = \overline{0, N-1}\,,\ j = \overline{1, M}$$

$$f_i^j = \frac{1}{h\tau} \int_{t_{j-1}}^{t_j} \int_{x_i}^{x_{i+1}} f(u(x,t),v)\, dx\, dt\,,\ i = \overline{0, N-1}\,,\ j = \overline{1, M}$$

$$\phi_i = \frac{1}{h} \int_{x_j}^{x_{i+1}} \phi(x)\ dx\ ,\ i = \overline{0, N-1},$$

$$(Y_0)^j = \frac{1}{\tau} \int_{t_{j-\frac{\tau}{2}}}^{t_{j+\frac{\tau}{2}}} Y_0(t)\ dt\ , (Y_1)^j = \frac{1}{\tau} \int_{t_{j-\frac{\tau}{2}}}^{t_{j+\frac{\tau}{2}}} Y_1(t)\ dt\ ,\ j = \overline{1, M}$$

The discrete analogy of the integral identity (6) writes in the form

$$(8)\quad h\,\tau \sum_{i=0}^{N-1} \sum_{j=1}^{M} u_i^j (\eta_i^j)_t - \sum_{i=0}^{N-1} \sum_{j=1}^{M} [-\lambda_i^j (u_i^j)_x (\eta_i^j)_x + f_i^j \eta_i^j] = -h \sum_{i=0}^{N-1} \phi_i\ \eta_i^0 - \tau \sum_{j=1}^{M} \eta_0^j (Y_0)^j - \tau \sum_{j=1}^{M} \eta_N^j (Y_1)^j$$

for any network function $\eta_i^j$, $\eta_i^M = 0$.

From [Samarski, 1992] and equality to zero the coefficients of $\eta_i^j$, we obtain the difference approximations problem for (1)-(3):

$$(9)\quad \begin{cases} (u_i^j)_{\bar{t}} - (\lambda_i^j (u_i^j)_{\bar{x}})_{\bar{x}} = f_i^j, i = \overline{0, N-1}, j = \overline{1, M} \\ u_i^0 = \phi_i, i = \overline{0, N-1} \\ -\ \lambda_0^j (u_0^j)_x - (Y_0)^j, j = \overline{1, M} \\ -\ \lambda_{N-1}^j (u_{N-1}^j)_x - (Y_1)^j = -h\ f_0^j - h\ (u_0^j)_{\bar{t}}, j = \overline{1, M} \end{cases}$$

But the functional (7) is approximated by the following way:

$$(10)\ DF(v) = \beta_0\ \tau \sum_{j=1}^{M}\ [u_0^j - (y_0)^j]^2\ + \beta_0\ \tau \sum_{j=1}^{M}\ [u_N^j - (y_1)^j]^2$$

$$+ h\tau \sum_{i=0}^{N-1} \sum_{j=1}^{M} [F(u_i^j, v) + Q(u_i^j) + B(u_i^j)]$$

A similar theorems 1,2 in [Farag,2004], the estimates of stability for the difference approximations problem (9) and an estimate on $V$ are given as follows:

**Theorem 4.1**

Suppose that that the all functions in the system (1)-(4) satisfy the above enumerated conditions. Besides, $\lambda(u,v),\ f(u,v)$ satisfy the Lipschits conditionon $V$ with constant $L$, $\forall\ (x,t) \in \Omega$, and for any $v \in V$. Then the estimates of stability for the difference approximations problem (9) are

$$
(11)\begin{cases}
\|u\|^2_{L_2(\bar{\Omega}_{h\tau})} \le C_2\ [\|\phi\|^2_{L_2(\bar{\Omega}_h)} + \|Y_0\|^2_{L_2(\bar{\Omega}_\tau)} + \|Y_1\|^2_{L_2(\bar{\Omega}_\tau)}] \\
\|u_x\|^2_{L_2(\bar{\Omega}_{h\tau})} \le C_2\ [\|\phi\|^2_{L_2(\bar{\Omega}_h)} + \|Y_0\|^2_{L_2(\bar{\Omega}_\tau)} + \|Y_1\|^2_{L_2(\bar{\Omega}_\tau)}] \\
\max\limits_j \|u^j\|^2_{L_2(\bar{\Omega}_h)} \le C_2\ [\|\phi\|^2_{L_2(\bar{\Omega}_h)} + \|Y_0\|^2_{L_2(\bar{\Omega}_\tau)} + \|Y_1\|^2_{L_2(\bar{\Omega}_\tau)}]
\end{cases}
$$

where $C_2$ is a constant depending only the constants in (1)-(4) and $L$ .

The proof of theorem 4.1 is then directly followed by theorem 1 in [Farag, 2004].

**Theorem 4.2**

Suppose that that the all functions in the system (1)-(4) satsify the above enumerated conditions. Besides, $\lambda(u,v),\ f(u,v)$ satisfy the Lipschits conditionon $V$ with constant $L$ , $\forall\ (x,t) \in \Omega$ and for any $v \in V$ . Then the stability estimation of the solution of difference approximations problem (9) on $v \in V$ are

$$
(12)\ h\sum_{i=1}^{N-1} (\delta u_i^j)^2 + h\tau \sum_{i=0}^{N-1} \sum_{j=0}^{M} (\delta u_i^j)^2 + h\tau \sum_{i=0}^{N-1} \sum_{j=0}^{M} (\delta u_i^j)_x^2 \le C_3\ \|\delta\, v\|^2_{E_N}
$$

where $C_3$ is a constant depending only the constants in (1)-(4) and $L$ .

The proof of theorem 4.2 is then directly followed by theorem 2 in [Farag, 2004].

## 5. Numerical Results

### 5.1 Numerical Approach

The outlined of the algorithm for solving OCP problem are as follows:

```
1- Given $it = 0$ , $\varepsilon_1 > 0$ , $\varepsilon_1 > 0$, $A_{it} > 0$ , $v^{it} \in V$ .
2- At each iteration $it$ do
   2.1 - Solve (1)-(3), then find $u(., v^{it})$ .
   2.2 - Minimize the functional (10) using PQI method.
   2.3 - Find optimal control $v_*^{it+1}$ .
   End do.
3- If $\|DF(v^{it+1}) - DF(v^{it})\| < \varepsilon_1$ , then Stop, else, go to Step 4.
4- Set $v^{it+1} = v^{it}$ , $A_{it+1} = \varepsilon_2\ A_{it}$ , $it = it + 1$ and go to Step 2.
```

### 5.2 Numerical Example

The numerical results were carried out for the following example of exact solution:

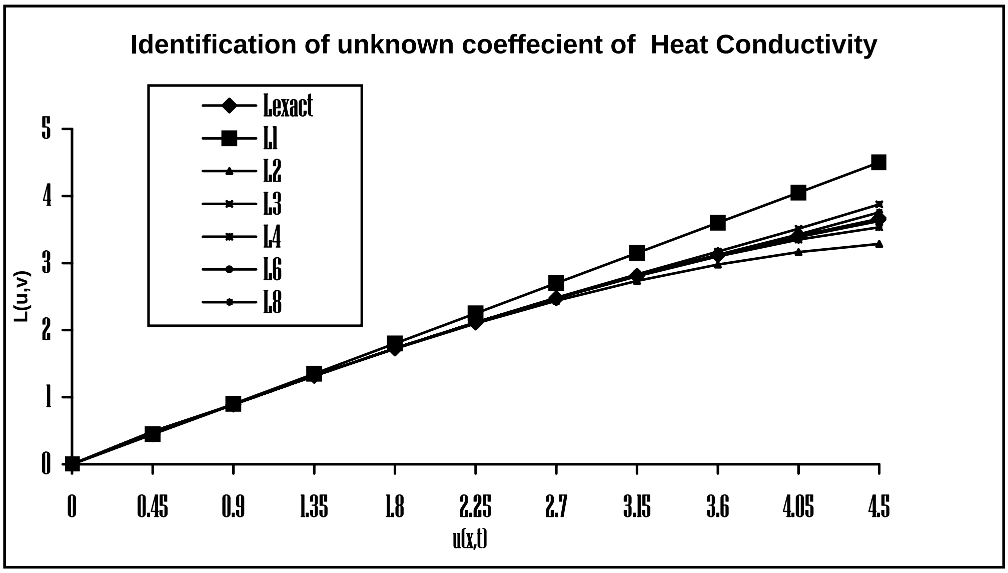
$$
u(x,t) = x + t,\ \lambda(u,v) = \frac{1}{1+u^2}\ ,\ x \in [0,0.8]\ , t \in [\ 0,\ 0.001]
$$

The iteration number $it$ for the function to be minimized $DF(v)$, the exact, approximate values of $\lambda(u,v)$ with $v^*$ tabulated in table 1 and the absolute error $\chi = \left| \dfrac{\lambda_{Exact} - \lambda_{Approx}}{\lambda_{Exact}} \right|$ also.

Table 1

| $it$ | $\lambda_{Exact}$ | $\lambda_{Approx}$ | $\chi = \left\lvert \dfrac{\lambda_{Exact} - \lambda_{Approx}}{\lambda_{Exact}} \right\rvert$ |
|---|---|---|---|
| 1 | .7352941E+00 | .1224296E+00 | .8334957E+00 |
| 2 | .7352941E+00 | .4115356E+00 | .4403116E+00 |
| 3 | .7352941E+00 | .4743363E+00 | .3549026E+00 |
| 4 | .7352941E+00 | .6568843E+00 | .1066374E+00 |
| 5 | .7352941E+00 | .7134509E+00 | .2970675E-01 |
| 6 | .7352941E+00 | .7150049E+00 | .2759337E-01 |

In Table 2, we report the number $NEF$ of function evaluations needed to attain the solution with accuracy on the modified function $DF(v)$ of the order $10^{-6}$. The above algorithm takes six iterations for decreasing $DF(v)$ to the value 0.8393609E-04. Knowing the computed optimal control values $v^*$ obtained by using the previous numerical algorithm, we can calculate the approximate values of the unknown coefficient $\lambda(u,v)$ can be represented in a series as $\lambda(u,v) = \sum_{i=1}^{k} u^k v_k$ . In the following Figure the curves denoted by $L_1\ , L_2\ , \ldots\ and\ \ L_{Exact}$ are the approximate and exact values of $\lambda(u,v)$. Obviously by increasing $k$ the coefficient $\lambda(u,v)$ will agree with precise ones.



## Acknowledgments

The author gratefully acknowledges Prof. Dr. Krassimir Markov, who made useful suggestions and remarks which helped to improve the paper.

## Bibliography

[Arada,2003]N. Arada, J.-P. Raymond and F. Troltzsch. On an augmented Lagrangian SQP method for a class of optimal control problems in Banach spaces, Comput. Optim and Appli.,vol. 15, 2003.

[El-Gendi,1995]T. M. El-Gendi, H. M. EL-Hawary, M. S. Salim and M. El-Kady. The computaional approach for optimal control problems of parabolic systems, J. Egypt. Math. Soc., vol. 3, 17--23(1995).

[Farag,1995]M. H. Farag. Application of the exterior penalty method for solving constrained optimal control problems,J. Math. and Phys. Soc. Egypt.(1995).

[Farag,2002]M. H. Farag. Necessary optimality conditions for constrained optimal control problems governed by parabolic equations, Journal of vibration and control ,V. 9, Issue 08, 2002.

[Farag,2004]M. H. Farag. A stability theorem for constrained optimal control problems, J. Computational Mathematics, V. 22(5),pp. 635-640,2004.

[Farag,2006]M. H. Farag and M. El-Monzery, Optimal control of a second order parabolic heat equation, Accepted for publication in Int. J. Infor. Theories and Appl.,Bulgaria,2006.

[Ibiejugba,1983]M. A. Ibiejugba. On the Ritz penalty method for solving the control of a diffusion equation,JOTA, 39(3),431-449(1983).

[Potapov,1978]M. M. Potapov. Difference approximations and the regularization of optimization problems for the systems of Goursat-Darboux type, Ser. Numer. and Keper., 2,17--26(1978).

[Ladyzhenskaya,1973]O. A. Ladyzhenskaya. Boundary value problems of mathematical physics,Nauka, Moscow, Russian(1973).

[Lur'e,1995]K. A. Lur'e. Optimal control in problems mathematical physics, Nauka, Moscow, 1975.

[Samarskii,1992]A. A. Samarskii. Introduction to numerical methods,Nauka, Moscow,1992.

[Tagiev,1982]R. K. Tagiev. Difference method of problems with controls in coefficients of hyperbolic equation, Cp. approx. methods and computer,109--119(1982).

[Tikhonov,1974]A. N. Tikhonov and N. Ya. Arsenin. Methods for the solution of incorrectly posed problems, Nauka, Moscow, Russian(1974).

[Yu,1998]W. Yu. A quasi-newton method in infinite-dimensional spaces and its application for solving a parabolic inverse problem, J. of Computational Mathematices, 16(4),305--318,1998.

## Author's Information

***M. H. Farag** - Department of Mathematics and Computer Science, Faculty of Education, Al Rustaq, Sultanate of Oman Or: Department of Mathematics, faculty of Science, El-Minia University, Egypt, e-mail: farag5358@yahoo.com*

# A COGNITIVE SCIENCE REASONING IN RECOGNITION OF EMOTIONS IN AUDIO-VISUAL SPEECH

**Velina Slavova, Werner Verhelst, Hichem Sahli**

***Abstract:** In this report we summarized the state-of-the-art of speech emotion recognition from the signal processing point of view. On the bases of multi-corporal experiment with machine-learned classifiers, the observation is made that the existing approaches for supervised machine learning lead to database dependent classifiers which can not be applied for multi-language speech emotion recognition without additional training because they discriminate the emotion classes following the used training language. As there are experimental results showing that Humans can perform language independent categorisation, we did a parallel between machine recognition and the cognitive process and tried to discover the sources of these divergent results. The analysis suggests that the main difference is that the speech perception allows extraction of language* *independent features although the language dependent features are incorporated in all levels of the speech signal and play a strong discriminative function in human perception. Based on several results in related domains, we have supposed that in addition, the cognitive process of emotion-recognition is based on categorisation, assisted by some hierarchical structure of the emotional categories, existing in the cognitive space of all humans. We propose a strategy for developing language independent machine emotion recognition, related to the identification of language independent speech features and to the use of additional information from visual (expression) features. The approach includes emphasizing the structure of the emotion categories in the cognitive space in order to reproduce the human 'strategy' for categorization.*

## Introduction

Traditional human machine interaction is normally based on passive instruments such as keyboard, mouse, etc. Emotion is one of the most important features of humans. Without the ability of emotion processing, computers and robots cannot communicate with humans in a natural way. It is therefore expected that computers and robots should process emotion and interact with human users in a natural way. Affective Computing and Intelligent Interaction is a key technology to enable computers to observe, understand and synthesize emotions, and to behave vividly. Affective computing aims at the automatic recognition and synthesis of emotions in speech, facial expressions, or any other biological communication channel (Picard, 1997). In fact, existing automatic speech recognition systems can benefit from the extra information that emotion recognition can provide (Ten Bosch, 2003; Dusan and Rabiner, 2005). In (Shriberg, 2005), the authors emphasize the importance of modelling non-linguistic information embedded in speech to better understand the properties of natural speech. Such understanding of natural speech is beneficial for the development of human-machine dialog systems. Several applications call for recognition only of the emotions in the speech, without processing the linguistic content. Such systems should be language independent.

During the last years the research concentrated in all these problems. We can sit the HUMAINE (Human-Machine Interaction Network on Emotion) a Network of Excellence in the EU's Sixth Framework Programme IST (Information Society Technologies). The thematic priority of HUMAINE aims to lay the foundations for European development of systems that can register, model and/or influence human emotional and emotion-related states and processes - 'emotion-oriented systems'. For the proposed here reasoning wore used several analyses and results of this research network, available on [http://emotion-research.net/].

## Automatic emotion recognition

Automatic recognition of emotions in speech aims at building classifiers (or models) for classifying emotions in unseen emotional speech. The data-driven approaches to the classification of emotions in speech use supervised machine learning algorithms (neural networks, support vector machines, etc.) that are trained on patterns of speech prosody. The training is performed with utterances or other speech instances, labelled with previously chosen set of emotions. Such labelled speech instances are taken from databases of emotional speech. Machine learned classifiers (ML-classifiers) can categorize other speech instances from the same database, according to the labels, used in the training procedure.

In general, the systems for speech analysis (speech recognition, speaker verification, emotion recognition) use techniques for *extraction* of *relevant* characteristics from the row signal. Concerning emotions, the relevant information is the *Prosody* (broadly determined as: *Intonation* – the way in which pitch changes over time, *Intensity* – the changes in intensity over time and *Rhythm* – segment's durations vs. time) and in the *Voice quality (*measured in spectral characteristics).

In Table 1 are given the features, used in one of the contemporary feature extraction approaches developed for Sony's robotic dog AIBO (Oudeyer, 2003). Table 2 illustrates another feature set, recently developed in VUB for the "segment based approach" (SBA) (Shami and Kamel, 2005). The size of the feature-vectors, provided as an input to the machine learning algorithm is practically not limited. One of the applied strategies for building a ML-classifier is to construct a feature vector with "everything which can be calculated" according to the reasoning that "the more information is collected from the row signal, the better it is". This strategy is often used in the practice. There exist classifiers with feature vectors of hundreds of values. The big length of the input vector reduces the performance of the classifier. The next step in this strategy is to discover the features which discriminate the speech data (to the training labels) and to discard the non-discriminative features.

Table 1. Feature set in the AIBO approach (Oudeyer, 2003).

| Acoustic features | Derived series of: | Statistics on the der. series, |
|---|---|---|
| -intensity | -minima, | -Mean, |
| -lowpass intensity | -maxima, | -maximum, |
| -highpass intensity | -durations between local extrema | -minimum, |
| -pitch | - the feature series itself | -range, |
| -norm of absolute | | -variance, |
| Vector derivative of the first 10 MFCC Components | | -median,<br>-first quartile, |
| *(MFCC - Mel-frequency cepstral coefficients)* | | -third quartile,<br>-inter-quartile range, |
| | | -Mean absolute value of the local derivative |

Table 2 Feature set in the Segment-based approach (SBA) (Shami and Verhelst, 2007)

| Pitch | Intensity | Speech Rate |
|---|---|---|
| -Variance | -Variance | -Sum of Ablute Delta MFCC |
| -Slope | -Mean | -Var. of Sum of Abs. Delta MFCC |
| -Mean | -Max | |
| -Range | | -Duration |
| -Max | | |
| -Sum of Abs Delta | | |

Speech science is already at a mature stage. Some studies focus on finding the most relevant acoustic features of emotions in speech as in (Fernandez and Picard, 2005; Cichosz and Slot, 2005). Other studies search for the best machine learning algorithm to use in constructing the classifier or investigate different classifier architectures. Lately, research has shifted towards investigating the proper time scale (utterances, segments) to use when extracting features as in (Shami and Kamel, 2005; Katz et al., 1996). Segment based approaches try to model the shape of acoustic contours more closely. There are also attempts to take into account phoneme-level prosodic and spectral parameters. (Lee S. et all., 2006(b), Lee, C.M. et all, 2004, Bulut et all 2005) All this efforts has lead to better and better ML-classifiers.

In all of the mentioned studies the classifiers were trained on one single speech corpus. It is known that ML-classifiers do not perform well on samples from other databases. There are no studies concerned with the problem of dependency of classifiers on the used speech corpora.

## Multi-corpora recognition

A recent study, conduced at VUB (Shami M., Verhelst W., 2007) treats the problem of multi-corpus training and testing of ML-classifiers. The study is based on the use of four emotional speech corpora: *Kismet, BabyEars* (both in American English), *Danish* (in Danish), and *Berlin* (in German). The four databases were grouped in two pairs: 1. Kismet-BabyEars pair, which contains infant directed affective speech, and 2. Berlin-Danish pair, containing adult directed emotional speech. The other difference between the two database pairs (DB-pairs) is in the length of the utterances (the infant-directed DB-pair contains shorter utterances).

Two approaches, corresponding to the two feature vectors (tables 1 and 2), were used - the segment based approach SBA and the utterance based approach AIBO. The two considered main questions have been: "When a classifier is trained to recognize a given emotion in one database, does it recognize the considered emotion in another database?" and "How does an ML-classifier perform if it is trained and tested on merged corpora, in other words – can it be generalize?" The "behaviour" of the classifiers described in Shami and Verhelst (2007) lead to several fundamental questions concerning the recognition of emotion in speech.

The speech entities in the four corpora contain speech instances for different sets of basic emotions, some of them overlapping. Table 3 and table 4 give the emotion labels (E-labels) and the numbers of speech instances labelled with them in each of the databases.

Table 3 Emotion Classes in Kismet and BabyEars databases (Shami and Verhelst, 2007)

| Kismet | | Baby Ears | |
|---|---|---|---|
| *Approval | 185 | *Approval | 212 |
| *Attention | 166 | *Attention | 149 |
| *Prohibition | 188 | *Prohibition | 148 |
| *Soothing | 143 | | |
| *Neutral | 320 | | |

Table 4 Emotion Classes in Berlin and Danish databases (Shami and Verhelst, 2007)

| Berlin | | Danish | |
|---|---|---|---|
| *Anger | 127 | *Angry | 52 |
| *Sadness | 52 | *Sad | 52 |
| *Happiness | 64 | *Happy | 51 |
| *Neutral | 78 | *Neutral | 133 |
| *Fear | 55 | *Surprised | 52 |
| *Boredom | 79 | | |
| *Disgust | 38 | | |

For the multi-corporal testing of classifiers, first the speech instances of the non-corresponding E-labels in each pair wore removed from the initial databases. In this way wore obtained "reduced" databases with only the common for the pair classes. The following experiments wore done:

*Between-corporal experiment:* Training on the one and testing on the other database of the pair. The results are not surprising: it seems that training on one database and testing on another database is not possible in general with the existing approaches.

*Integrated corpus experiment:* Merge databases into one "Integrated corpus" (for each pair).

*First condition:* Merge the classes from the corresponding E-label into a joint "common" class. For example, the instances from Kismet*Approval and from Baby Years*Approval wore fused a novel class: Integrated*Approval. The ML-classifiers wore trained and tested, on the fused classes. They "learned" them and "performed" the recognition task surprisingly well (classification accuracies: 74.60% for Kismet-Baby Ears and 72.2 % for Berlin-Danish[1]).

*Second condition:* Keep in the Integrated corpus the classes as they wore in the initial databases of the pairs. The ML-classifiers wore trained and tested in the integrated corpora on the old classes.

The classification accuracies obtained in the two "integrated" conditions wore similar: the accuracy of the classifiers in an "integrated corpus" could be seen as average of the accuracies in the one and in the other databases of the pair. So, the use of a heterogeneous corpus does not lead to a notable deterioration in classification accuracy. This is a very good practical result, as it is known that the less uniform the training corpus is, the less accurate the classifier is. And, on the other hand, a classifier learned using heterogeneous corpora is more robust. One important conclusion, given in this study, is that the existing approaches for classification of emotions in speech are efficient enough to construct a single classifier, based on larger training data from different corpora. From the practical point of view, the result gives a solution for building classifiers in integrated corpora with shared emotion classes.

[1] The results were also compared for differenf machhine-lerning algorithms, that is not given here

Here the results have been analysed from the point of view of another interesting founding, related to the representation of the emotion classes in the feature space. The result is seen in the Second “integrated” condition, were the emotion-classes have been preserved as they wore in the initial databases of the pairs.

Table 5. Confusion matrix of Berlin-Danish Integrated corpus (Shami and Verhelst 2007)

| A | B | C | D | E | F | G | H | ← | classified as |
|---|---|---|---|---|---|---|---|---|---|
| **74** | 1 | 2 | 1 | 0 | 0 | 0 | 0 | **A** | Berlin*Neutral |
| 3 | **36** | 0 | 25 | 0 | 0 | 0 | 0 | **B** | Berlin*Happy |
| 4 | 0 | **48** | 0 | 0 | 0 | 0 | 0 | **C** | Berlin*Sadness |
| 1 | 25 | 0 | **101** | 0 | 0 | 0 | 0 | **D** | Berlin*Anger |
| 0 | 0 | 0 | 0 | **106** | 2 | 17 | 8 | **E** | Danish*Neutral |
| 0 | 0 | 0 | 0 | 7 | **29** | 2 | 13 | **F** | Danish*Happy |
| 0 | 0 | 0 | 0 | 21 | 4 | **27** | 0 | **G** | Danish*Sad |
| 0 | 0 | 0 | 0 | 11 | 16 | 2 | **23** | **H** | Danish*Angry |

For Berlin-Danish integrated corpus, it came out that classifiers never “confuse” for example Berlin*Anger and Danish*Angry. The confusion matrix of Berlin-Danish pair is given in Table 5. It is seen that instances belonging to one of the databases are never “taken” as instances belonging to the other database. Automatic clustering (using the K-means clustering algorithm) showed that the same emotion-classes from the two databases are represented on different clusters and even that the entire databases don’t share any cluster.

For Kismet- BabyEars pair there was a small tendency of generalization over the emotions, as some instances of BabyEars wore “confused” with the equivalent emotion in Kismet (but never the inverse). Automatic clustering showed that the two databases share four (of the six) clusters and that when on one cluster there are classes from both databases, these classes represent one and the same emotion.

Why these results? This could be linked to the language in which the emotions are expressed. Or to the nature of the emotions - Kismet/BabyEars contains infant directed communicative intents, generally regarded as culture and language independent (Fernald, 1992). In any case, the question which arises at this point is related to the recognition accuracy (RA) of humans on this task.

## Comparison with cognitive processes

Human capacity to recognize emotions only from speech is reported in the literature between 60% and 85%, depending on the experiments, the emotion classes and other additional circumstances. For example, human listening recognition accuracy has been evaluated on 79% for stimuli from BabyEars database (Shami M., Verhelst W., 2006). On Danish database it is 67% (Engberg and Hansen, 1996). What about the vagueness of the different expressed emotions for the listeners? The reported in the literature experimental results show that, depending on the experiment, listeners recognize with unequal success the emotions Anger, Disgust, Fear, Happiness, Sadness, and Surprise, often supported as being basic for humans. For example, human RA is best for Anger and worse for Happiness in the experiment of Lee (Lee C.M. et all, 2004). In Danish database, human recognise best Sad and worse Happy. The abundance of such examples leads to the doubt that the target emotions are well expressed. One can also wonder have participants one and the same concept for the label “Sad”.

The important is that in almost all reported last year’s mono-corporal results, the recognition accuracy of the classifiers is comparable with the human categorization capacities for the samples, stored in the corresponding databases. The resemblances between the classifiers and the human evaluators within the same database go further: as it has been reported Shami and Verhelst (2007), the use of SBA approach on Danish database lead to a classifier which makes the same mistakes as humans. Listeners recognise best *Sad, the classifier does the same; listeners confuse *Surprise with *Happiness and *Neutral with *Sadness, the classifier does the same. From the modelling point of view, that means that the used feature-space is a good projection of the human cognitive space, which contains also models of acoustic parameters of speech emotions. The hope is that such kind of mapping will be available for the multi-corporal experiment. Unfortunately, that is not the case.

Suppose that the aim is to build a multilingual emotional classifier. The corpus should include labelled classes of speech instances from several databases. A classifier will "learn" Danish*Anger, German*Angry, Polish*Angriness etc. These classes could be fused in one class; the classifier will learn the image of this composed class and will become more robust. As it is demonstrated with the multi-corporal experiments, classifiers "learn" quite well the images of composed emotion-classes, represented on non intersecting clusters in the feature space. One may speculate and fuse Danish*Sad with Berlin*Happy to train the classifier on the novel class "Integrated*Potatoes". The expectation, looking at the confusion matrix of Berlin/Danish pair, is that the classifier will "learn" that class.

A known work in the domain of speech and emotion is the study of Klaus Scherer (Scherer K., 2000, Scherer et all 2001), reporting results (fig. 1) of a multi-language emotion encoding/decoding experiment. Scherer used a set of basic emotions: *{fear, joy, sadness, anger and neutral}* and tested human recognition accuracy on samples of emotional speech, containing content-free utterances composed of phonological units from many different Indo-European languages. That was done in nine countries, on three continents. In all cases human recognition accuracy was substantially better than chance and showed an overall accuracy of 66% across all emotions and countries, suggesting the existence of similar inference rules from vocal expression across cultures. This key-suggestion is widely accepted in the speech-emotion scientific domain. So, it turns out that there are common acoustic images of emotional speech in the human cognitive space, and they are applied with a good result even for utterances of a never heard or even invented[1] language.

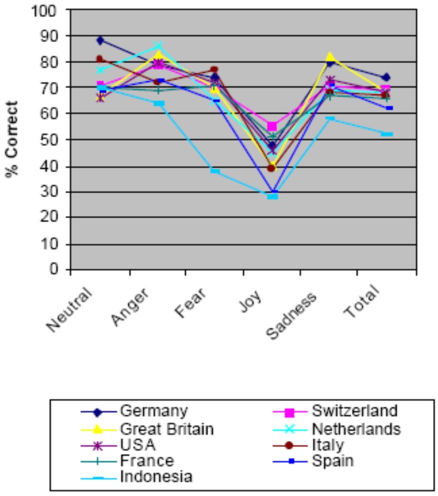


*Fig.1. Result for human recognition of speech-emotion across languages and cultures (Scherer K., 2000)*

Scherer's study found differences in the results across the countries: the highest accuracies wore obtained by native speakers of Germanic languages (Dutch and English), followed by Romanic languages (Italian, French, and Spanish). The lowest recognition rate was obtained for the only country studied that does not belong to the Indo-European language family, Indonesia.

Here a hypothesis could be made: the worse recognition result is obtained when using **only** the basic "perceptive" features which permit to categorize speech-emotions in the cognitive space.

The better recognition accuracy of the listeners from the other language groups can in this case be explained in two ways: 1. listeners perceive in the samples features which correspond to their perceptive features *in addition* of the basic perceptive features; 2. the emotion categories in the cognitive space of these listeners wore better fitting with the emotion-labels of the samples.

The results of the multi-corporal machine leaning experiment are not comparable to the results in the Scherer's experiment. Figure 2 illustrates an analogy between the machine recognition and the human recognition. Classifiers depend exclusively on the labelled training data and humans perform the task without be trained. It is clear that the used from humans perceptive features permit generalisation and categorization of the signal, but

[1] This is used also in the domain of synthesis of emotional speech – the produced speech is not in any language,

the features, extracted for the machine classifier do not allow that. If an ideal feature space could be employed, similar emotions belonging to different databases should be assigned to the same clusters, as humans do.

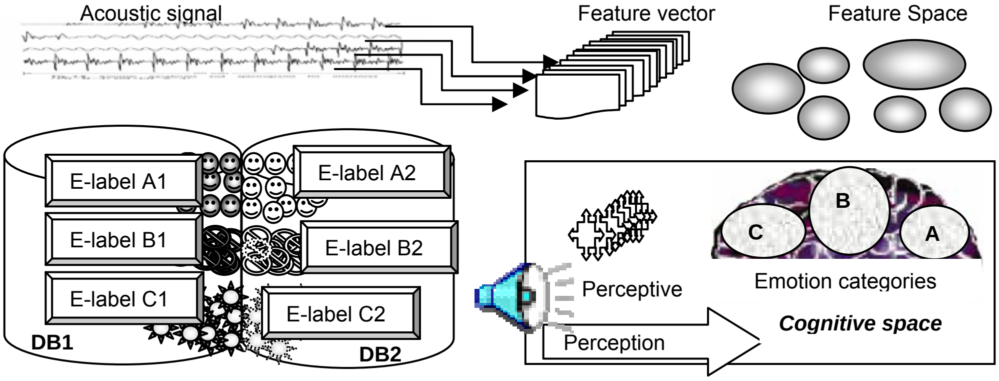


*Fig 2. Scheme of the analogy between cognitive process and machine recognition.*

Several atomic hypotheses could be made at this point. For example:

A. There is not enough emotion-relevant information captured by the feature vector.

B. There is language dependent information captured by the feature vector.

C. The perceptual features allow humans to categorize to more general categories. The cognitive space has a structure which permits them to path the sub-category, used in the proposed label.

Concerning the first two hypotheses, a lot of efforts have been made to ameliorate the feature extraction and to find relevant feature vectors. Acoustic correlates of specific emotional categories are investigated in terms of pitch, energy, temporal and spectral parameters, on suprasegmental, segmental and even on phoneme level. This is in aim to extract more and more emotion-relevant information (HUMAINE, 2004a). The question about the language dependence of the used features stays opened. The language dependent information is incorporated on all levels of the speech prosody. Newborns discriminate the different languages. Babies do that without relying on phonemic cues, but on the bases of rhythmic and intonational cues only (Ramus, F., 2002). We may expect that machine-learned classifiers do the same – they discriminate languages. So, the task is to present to the classifier only language independent information. A classic idea is to look for acoustic correlates in emotion in music, what corresponds a lot to Scherer's reasoning. There is a lot a research in this direction (Kim 2004; Kim et all 2004). But, the speech signal is much more complex. Haw could the language dependent and the language independent ingredients of the extracted features be separated, and how to do this on the suprasegmental, the segmental and, why not, on the phoneme level in order to take only the features with pure information about emotions only? Humans can do that. So, it should be possible to be done. Obviously, such a task demands a lot of particular research.

Hypothesis C. requires a separate approach. The C hypothesis explains the good performance of humans in speech emotion recognition with the structure of the cognitive space of emotion categories.

## Emotions

To study relations between speech and emotion, it is necessary to derive methods describing emotion. Although there have been numerous studies with regards to both the psychological and the engineering aspect of emotions, it is still not clear how to define and how to categorize human emotions. There are two basic approaches used.

The first approach is "discrete" (Fig. 3). Emotion categories are determined as entities with names and descriptions. Several theorists argue that a few emotions are basic or primary (Ekman, 1992; Izard, 1993). The emotions of anger, disgust, fear, joy, sadness, and surprise are often supported as being basic from evolutionary, developmental, and cross-cultural studies. That theoretical approach is convenient for the purposes of machine learning, as it provides directly labels for the training data. In speech emotion recognition, the attempts have so far concentrated on a small number of discrete, extreme emotions, in aim to obtain maximally distinguishable prosodic profiles.

*Fig. 3. Emotion categories -Labels*

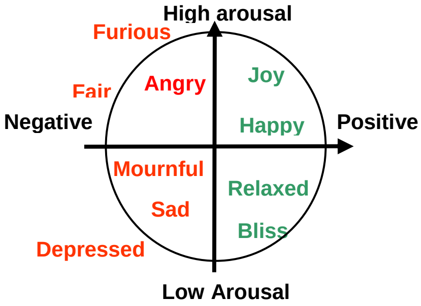


*Fig. 4. Emotion dimensions*

The other approach is “continuous”. The basic properties of the emotional states are described in a continuous space of “emotion dimensions” (fig. 4). The most frequently encountered emotion dimensions are activation (the degree of readiness to act) and evaluation (“valence” in terms of positive and negative). They provide a taxonomy allowing simple distance measures between emotions.

The central question for the experts in the field of speech emotion is: what has to be recognized is it *emotional categories and/or dimensions*. The performance of human participants and the performance of an automatic recognition system are totally dependent on the number and the degree of differentiation of the emotion categories/dimensions that have to be discriminated. The consensus of the experts from HUMAINE is that “labelling schemes based on traditional divisions of emotion into ‘primary’ or ‘basic’ categories is not relevant” (HUMAINE 2004, b). So, the task has turned to cluster the emotional states which have names in the continuous space. Several approaches have been developed in this purpose (Douglas-Cowie, et al 2003; Devillers et all 2005).

A large study was conducted within the international project AMI (Wan et all, 2005) to determine the most suitable emotion labels for the specific context of *meetings*. One of the contemporary labelling schemes *FeelTrace* (Cowie et all 2000), based on the above mentioned emotion dimensions, was used. A listing of 243 terms describing emotions was compiled from the lists of three research centres. These emotion-labels were first *clustered by meaning* by the project's experts. After that, participants from various companies and professions have evaluated the position of the separate emotions on the axes. Figure 5 gives the plot of the participant's evaluation of the emotions from one meaning-cluster.

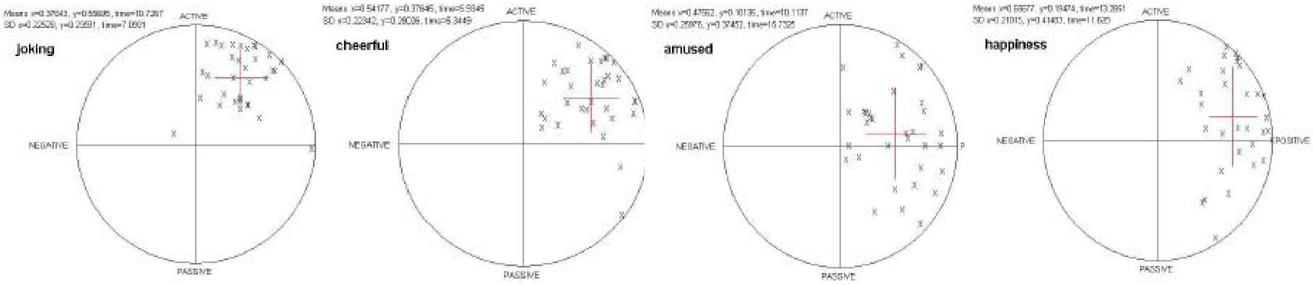


*Fig. 5. Results of landmark placement survey for joking, amused, cheerful and happiness (Wan et all, 2005)*

The first observation when analysing this experiment is that one can cluster emotion-names *by meaning*. The second observation is that the others agree on the same meaning-cluster, as they locate the emotions' names from the cluster on approximately the same place. The last observation is that the dispersion of participants' evaluations covers "semicircles" and quarters of the plane. One may suppose that in the cognitive space exist "generalized" categories, in correspondence of the clusters. In any case, the agreement of the participants on the meaning of the axes is evident. Where meaning and categories appear, there should be an attempt to analyse the cognitive processes underlying emotion.

## A Possible Cognitive Science Reasoning

At a first stage one should check if there are physiological phenomena, leading human beings to "innate" perception of the dimensions of emotion properties. Emotion-related biological changes are well documented. Recent studies (Kim 2004, Kim at all. 2004) also showed that parameters from measurements as cardiograms, encephalograms, respiration and skin conductivity, are highly correlated with the emotional dimensions. The study was conduced by provoking emotive states using music stimuli. As it is illustrated in fig. 6, on the Arousal axe there are two well distinguishable clusters, obtained when hearing songs inducing {joy and anger} for the left cluster and {sadness and bliss} for the right cluster.

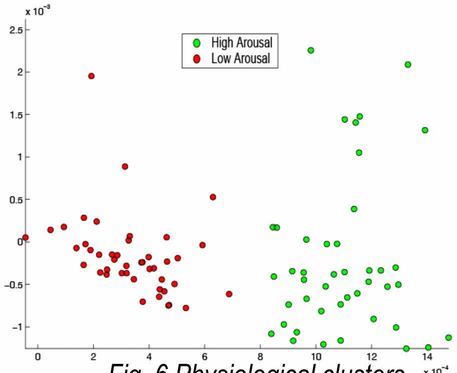


*Fig. 6 Physiological clusters on the Arousal axe (Kim et all 2004)*

So, there exists some innate knowledge about the emotion dimensions, as no-one learns how to feel when hearing music and what hard rhythm to have in this moment. The hypothesis that in the cognitive space exist "general" emotional categories is supported by the results, as the obtained physiological clusters correspond to quadrants of the plane on figure 4. The set of stimuli and the reactions suggest that humans distinguish such general categories.

These "general" categories do not obligatory have names. It is known in cognitive science that humans divide perceptual continuums intervals and than give names to the intervals. One example is the perception of colours and their names. The continuum of light frequencies is perceived in the same way by human being. But different cultures divide this continuum into intervals in different manner (and gave them names as "red" or "blue"). There are cultures in which the named-intervals for what we call "white" are nine and cultures which have only two names of colours for the entire spectrum.

The hypothesis that human perceive in emotive speech features that allows them to categorize to more general categories seems reliable. These categories do not necessarily have names in the language(s). But when presenting to someone a sample of positive active speech and the labels {angry, sad, happy, fear and neutral}, she will certainly decide that it is "happy".

The problem is how to shape the feature space of multilingual classifiers of emotions.

The most convenient for machines are taxonomies and tree structures. Imagine the plane arousal-valence is covered with specific emotions, as it shown in figure 7, for example with the labels E1 to E8 (This precise positioning is purely geometrical; the labels are just covering the quadrants and the neutral positions). Suppose the position of these labels correspond to precise emotions as Anger, Happiness etc. By the way, the names of those places can be determined.

Suppose these areas are leafs of a taxonomical tree. The upper level of the tree corresponds to general categories. As shown on figure 8, the taxonomic structure of general categories and more concrete emotions could be in two wais: 1. division to general categories depending on the arousal and to more concrete states following the valence – positive, neutral or negative (fig. 8.a); 2. division to general categories according to the valence and to concrete states following the arousal - positive, neutral or negative (fig. 8.b). As it is shown in figure 7(b), the 'general' level of classification could be useful for determining the tendency of the subject's behaviour.

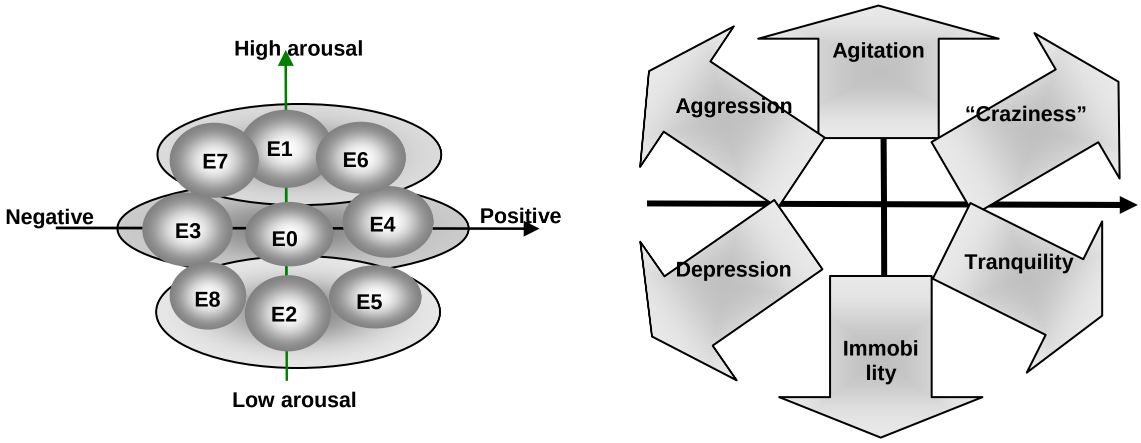


*Figure 7 (a). Lower level of the taxonomical trees; (b) Tendencies of behaviour in the emotions' space.*

Suppose that the set of language independent features, which leads to the classification to the general categories, is known. The proposal is to use the same strategy as humans seem to do. That leads to the following "algorithm":

- Take into account only language independent features, even if they are not too many.
- Classify to which general category belongs the speech signal.
- If we have information on the language, use additional information and classify to a leaf.

This strategy demands a kind of double classification – first to the general category and after that - to a leaf. But it avoids big mistakes. Such a classifier wouldn't need more and more data to be trained.

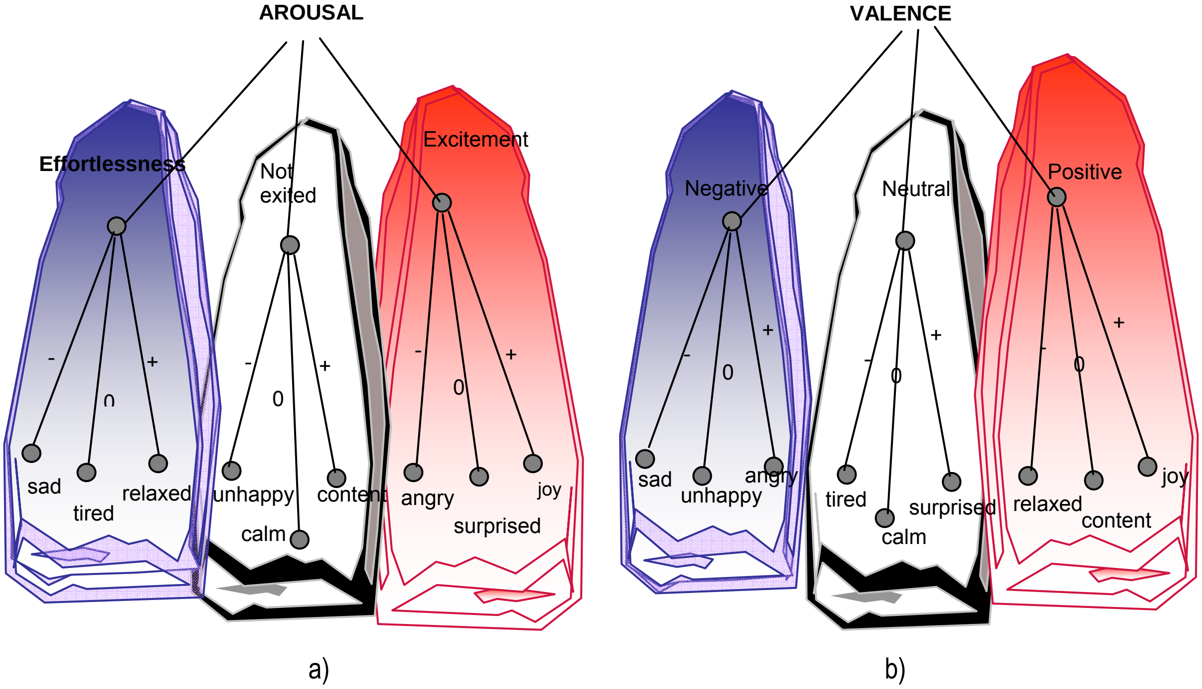


*Figure 8. Taxonomies of general emotional categories and less general emotions.*

From a general point of view, the capacities of a machine-learned classifier are never as perfect as human capacities for recognition and categorisation. It is obvious that the use of additional channels of information for the machine recognition, such as visual (expression) features, will be very helpful.

## Conclusion

The field of speech emotion recognition has attempts several promising results. However, the data-driven approaches lead to machine learned classifiers that are database dependent. The problem can be solved by means of merging emotion-speech corpora and training with more and more data.

The experimental results for human emotion recognition show that the underlying cognitive mechanisms allow language independent categorisation although the information about the used language is deeply involved in the speech signal. The analysis suggests also that the cognitive process uses some internal structure of the emotional categories, existing in the cognitive space.

We elaborate a general strategy for developing language independent emotion recognition, which does not need large amount of training samples in all languages. The proposed approach provides a basis for a future research and experimental work. The study should first consider the identification of language independent speech features and culture independent information from parallel modalities such as visual (expression) features. In a second step we would analyse several classifiers, by considering the general categories of emotion. Parallel to that we will investigate the relationship/dependencies between the emotion categories and language(s) for the classification of leafs (if necessary). A comparison with state-of-the art of automatic emotion classifiers will be made.

We are currently looking for funding for a research project. Several tracks are being considered, National, Bilateral and Europeen.

## Bibliography

Breazeal, C., Aryananda L., 2002. Recognition of Affective Communicative Intent in Robot-Directed Speech. In: Autonomous Robots, vol. 12, pp. 83-104.

Bulut M., Busso C., Yildirim S., Kazemzadeh A., Lee, C.M., Lee S., and Narayanan S., (2005) Investigating the role of phoneme-level modifications in emotional speech resynthesis. In Proc. of EUROSPEECH, Interspeech, Lisbon, Portugal, 2005

Cichosz, J., Slot, K., 2005. Low-dimensional feature space derivation for emotion recognition. In: Interspeech 2005, p.p.477-480, Lisbon, Portugal.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M., 2000 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time, ISCA Workshop on Speech & Emotion, Northern Ireland 2000, p. 19-26

Devillers L., Vidrascu L., Lamel L., (2005) Challenges in real-life emotion annotation and machine learning based detection, Neural Networks, Volume 18 , Issue 4, Elsevier Science Ltd. Oxford, UK, UK, 407 - 422

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. Speech Communication, Speech Communication, Elsevier Science Publishers B. V. Amsterdam, The Netherlands, 40, 33-66

Dusan, S., Rabiner, L., 2005. On Integrating Insights from Human Speech Perception into Automatic Speech Recognition. In: Interspeech 2005, Lisbon, Portugal.

Ekman P., "Are there basic emotions?" Psychological Review 99 (3), pp.550–553, 1992.

Engberg, I. S., Hansen, A. V., 1996. Documentation of the Danish Emotional Speech Database (DES). Internal AAU report, Center for Person Kommunikation, Denmark..

Fernald, A., 1992. Human maternal vocalizations to infants as biologically relevant signals: an evolutionary perspective. In: Barkow, J.H., Cosmides, L., Tooby, J. (Eds.), The Adapted Mind: Evolutionary Psychology and the Generation of Culture. Oxford University Press, Oxford.

Fernandez, R., Picard, R. W., 2005. Classical and Novel Discriminant Features for Affect Recognition from Speech. In: Interspeech 2005, p.p. 473-476, Lisbon, Portugal.

HUMAINE 2004b, « Theories and Models of Emotion », June 17-19, 2004, Work Group 3 – Synthesis, online available on http://emotion-research.net/

HUMAINE, 2004a, "Emotions and speech - Techniques, models and results Facts, fiction and opinions", Synteses of HUMAINE Workshop on Signals and signs (WP4), pr. by Noam Amir, Santorini, September 2004, online available on http://emotion-research.net/

Izard C., "Four systems for emotion activation: cognitive and noncognitive processes," Psychological Review 100, pp.68–90, 1993

Katz, G., Cohn, J., Moore, C., 1996. A combination of vocal F0 dynamic and summary features discriminates between pragmatic categories of infant-directed speech. In: Child Development, vol. 67, pp. 205-217.

Kim K. H., Bang S.W. and Kim S. R. (2004), Emotion recognition system using short-term monitoring of physiological signals, in: Journal of Medical and Biological Engineering and Computing, Springer Berlin / Heidelberg. Volume 42, Number 3 / May, 2004

Kim, Jonghwa, 2004, Sensing Physiological Information, Applied Computer Science, University of Augsburg, Workshop Santorini, HUMAINE WP4/SG3, 2004, online available on http://emotion-research.net/

Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S., "Emotion recognition based on phoneme classes", ICSLP, 2004.

Lee, S., Bresch E, Adams J., Kazemzadeh A., and Narayanan S., 2006 (a). A study of emotional speech articulation using a fast magnetic resonance imaging technique. In Proceedings of InterSpeech ICSLP, Pittsburgh, PA, Sept. 2006.

Lee, S., Bresch E, and Narayanan S., 2006 (b). An exploratory study of emotional speech production using functional data analysis techniques. In Proceedings of 7th International Seminar On Speech Production, Ubatuba, Brazil, pp. 525-532. December 2006.

Picard, R., 1997. "Affective Computing", MIT Press, Cambridge. 1997

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. Annual Review of Language Acquisition, 2, 85-115, 2002.

Rotaru, M., Litman, D., 2005. Using word-level pitch features to better predict student emotions during spoken tutoring dialogues. In: Interspeech 2005.

Scherer, K. R. "A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology," in : Proc. ICSLP 2000, Beijing, China, Oct 2000.

Scherer, Klaus R., Rainer Banse, and Harald G. Wallbott. 2001. "Emotion Inferences from Vocal Expression Correlate across Languages and Cultures," Journal of Cross-Cultural Psychology 32/1: 76-92.

Shami M., Kamel, M., 2005. Segment-based Approach to the Recognition of Emotions in Speech. In: IEEE Conference on Multimedia and Expo (ICME05), Amsterdam, The Netherlands.

Shami, M., Verhelst, W., 2006. Automatic Classification of Emotions in Speech Using Multi-Corpora Approaches. In: Proc. of the second annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2006), Antwerp, Belgium.

Shami M., Verhelst W., 2007; An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech, to appear in: Elsevier Editorial System(tm) for Speech Communication, 2007

Shriberg, E., 2005. Spontaneous speech: How people really talk and why engineers should care. In: Eurospeech 2005, Lisbon, Portugal.

Ten Bosch, L., 2003. Emotions, speech and the ASR framework. In: Speech Communication, vol. 40, no. 1–2,.213–225.

Wan V., Ordelman R., Moore J., Muller R., (2005) "AMI Deliverable Report, Describing emotions in meetings", internal project report, On line available http://www.amiproject.org/

## Authors' Information

***Velina Slavova*** *- New Bulgarian University, Department of Computer Science; vslavova@nbu.bg*

***Werner Verhelst*** *- Vrije Unversiteit Bruxels, Department of Electronics & Informatics; wverhels@etro.vub.ac.be*

***Hichem Sahli*** *Vrije Unversiteit Bruxels, Department of Electronics & Informatics; hsahli@etro.vub.ac.be*

# INTEGRAL TECHNOLOGY OF HOMONYMY DISAMBIGUATION IN THE TEXT MINING SYSTEM "LOTA"

Olga Nevzorova, Vladimir Nevzorov, Julia Zin'kina, Nicolay Pjatkin

***Abstract**: The article describes the integral technology of homonymy disambiguation, realized in "LOTA" textual document analysis system. The technology contains a totality of homonymy disambiguation methods as well as a scheme of their interaction..*

***Keywords**: natural language processing, functional homonymy, homonymy disambiguation..*

***ACM Classification Keywords**: H.3.1.Information storage and retrieval: linguistic processing*

## 1. Introduction

"LOTA" specialized system of textual documents processing is a system of Text Mining class. The system is intended for analyzing "The Logic of Work" specialized texts in Russian language, which describe the logic of work of a complex technical system in various functioning modes [Nevzorova et al., 2001]. The main task of the analysis is the extrication from the texts given of an informational model of the algorithms which solve a definite task in a particular problem situation$ and the control over structural and informational integrity of the chosen algorithm scheme.

The algorithm informational model includes:

- description of the input information flow (types of informational signals or a semantic description of the information flow with an indication of the information source - particular algorithm, particular measuring device);
- description of the processes of transforming input data into output data (acceptable way of problem solving);
- description of the output information flow (types of information signals or a semantic description of the information flow with an indication of the information receiving point).

Solution of the main task is provided by a complex of test processing technologies, which include:

- technologies of morphosyntactic analysis;
- technologies of semantic-syntactic analysis;
- technologies of interaction with applied ontology.

The indicated sum of technologies is formed on the basis of the central kernel, the applied ontology (further on, aviaontology), which supplies coordinated interaction of various program modules. Aviaontology conceptually describes the data domain of informationally maintaining different flight modes of anthropocentric systems [Dobrov et al., 2004]. Avaiontology is a notion net of the problem domain. The current ontology size is more than 1600 notions (about 5000 textual occurrences of the notions). Aviaontology belongs to the class of linguistic (lexical) ontologies and is intended for integration into various linguistic appendixes.

The program complex consists of three interacting subsystems: "Analyzer" subsystem of technical texts linguistic analysis, "OntoEditor+" subsystem and "Integrator" subsystem. The subsystem interaction is realized on the basis of "client-server" technology. Besides, subsystems act in different modes in various subtasks (server mode or client mode).

"OntoEditor+" software toolkit [Nevzorova et al., 2004] is a specialized database management system. The system is intended for manual editing of the ontologies kept in relational database in TPS format, as well as for user query service and outer programs. New system functionalities are provided by a "Linguistic toolbox" functional set, by means of which the integration of the applied ontology into linguistic appendixes is realized. The most typical tasks solved with the help of "OntoEditor+" system toolbox are: studying structural features of the applied ontology with the help of "OntoEditor+" system research toolbox; constructing the applied ontology linguistic frame; the task of covering the text with ontological entries; forming conclusions on the applied ontology; etc.

"Analyzer" subsystem realizes the main of linguistic text processing (graphematic, morphosyntactic and partly syntactic analysis). In this article, the integral technology of homonymy disambiguation will be investigated, which is aimed at functional, morphological and lexical homonymy disambiguation first of all.

"Integrator" subsystem solves external query extracting data from the text. External query structure contains algorithm informational model components. External query is interpreted in the course of interaction with "OntoEditor+" subsystem as a structure bound to the applied ontology. The extraction of informational model components is realized on the basis of identifying the elements of input text segment tree (interaction with "Analyzer" subsystem) with query structure elements (interaction with "OntoEditor+" subsystem).

## 2. Integral technology of homonymy disambiguation

"Logics" tests are real technical texts with complex syntactic structure. The texts contain many abbreviations, among them the author's ones, sentences with enumerations, homonymy of various types etc. The stage of linguistic analysis is supported by a range of standard linguistic resources, especially grammatical dictionary, formed on the basis of A.A. Zaliznyak's grammatical dictionary and substantially enlarged by adverbs and specialized vocabulary; standard abbreviations dictionary; collocation dictionaries etc. Under non-standard resources supporting linguistic text analysis are subsumed aviaontology linguistic frame and the indexed base of problem domain frequent collocations. Practically, these resources the main homonymy disambiguation technologies in the system, namely homonymy disambiguation on the basis of frequent collocations indexed base, as well as homonymy disambiguation on the basis of aviaontology linguistic frame.

In the last few years the group of authors was developing a universal technology of functional homonymy disambiguation on the basis of contextual rules method [Nevzorova et al., 2006]. The technology given is based on linguistic scrutiny of the homonyms' syntactic behavior and homonyms' grammatical characteristics specification. This scrutiny revealed a range of actual problems of Russian functional homonyms' lexicographic description. New statistical basis for homonymy classification and subtypes selection was proposed. The work upon developing a new Russian functional homonyms dictionary on the basis of corpus research was started. The contextual method of homonymy disambiguation is the basic one in the integral technology of homonymy disambiguation in "LOTA" system. However, practical tasks of the system revealed some important aspects of linguistic analysis, which stimulated the development of new homonymy disambiguation methods. Initially, the work related to obtaining quantitative and qualitative assessments of specialized textual base of system documents. The analysis revealed quantitative assessments and the distribution of homonyms into types. Another important assessment was obtained in the course of typical homonym contexts. This research revealed the degree of technical texts homonymy (on average, 15-20 % of homonyms); the frequent homonyms were listed, as well as their typical contexts. These results became a basis for new applied technologies of homonymy disambiguation.

Thus, the integral homonymy disambiguation technology developed in "LOTA" system includes the following methods:

- contextual method of functional homonymy disambiguation;
- method of functional, grammatical and lexical homonymy disambiguation based on the indexed collocation base;
- method of functional, grammatical and lexical homonymy disambiguation based on the ontology linguistic frame.

## 2.1. Contextual method of functional homonymy disambiguation

Contextual method of functional homonymy disambiguation comes to developing for every functional homonymy type a group of rules defining the syntactic context of the homonym disambiguation and forming the group control structure which defines the rule application order. In work [Nevzorova et al., 2006] the main benefits and drawbacks of this method were described, concrete structures of generalized rules for some functional homonymy types disambiguation were given. Contextual disambiguation method is applied at the stage of morphosyntactic text analysis, as rather frequently a syntactic method of building homogeneous groups is used in homonymy disambiguation. Disambiguating a homonym outside the group allows considering not only the local homonym context, but also the more remote one. Assuredly, it is one of the main benefits of the method. In the integral technology, the contextual method is a basic one and is applied last among the homonymy disambiguation methods.

## 2.2. Method of functional homonymy disambiguation based on the indexed collocation base

In order to realize this method, an integrated program technology of building up the homonym contexts base index was developed. The program technology developed includes the modules of creating and leading the homonym index, the module of coordinating the index base with the main linguistic resource (grammatical dictionary) and the mechanism of solving input queries on disambiguation (search) of typical homonymy contexts in the input text on the basis of homonyms index. The technology developed was realized on the basis of main "OntoEditor +" and "Analyzer" subsystems interaction.

"OntoEditor+" system, in order to provide effective integration into linguistic appendixes, supports a group of interconnect protocols of informational exchange with outer system program modules and outer dictionary databases, supplying client-server mode work. The functional, morphological and lexical homonymy disambiguation in the input texts is realized on the basis of a mechanism of recognizing the homonym contexts fixed in the indexed context base. Three main mechanisms of enlarging the indexed functional homonyms context base were developed:

- manual input and editing data on typical homonym contexts;
- import of typical homonym contexts from a textual file prepared in a special format of data representation;
- import of typical homonym contexts discovered by special search mechanism of the "Analyzer" subsystem.

This mechanism is organized as a query to the "Analyzer" subsystem, with "OntoEditor+" subsystem transferring to it a textual corpus where the search is being carried out. While processing the "Analyzer" subsystem transfers into "OntoEditor+" subsystem the information about the homonym contexts discovered. This is written either into the homonym index, or in the automatic mode, or in the mode of a dialogue with the operator. The special feature of the dialogue mode is the self-study mode, which is realized using the event diary mechanism. Depending on the settings, the diary records some important events in the system, for example, information change in the homonym index or the interaction with the "Analyzer" subsystem. In the self-study mode the sequence of earlier generated dialogues is saved and controlled, which provides the generation of unique dialogues only and homonymy disambiguation without repetition.

Based on the experimental text collection, as well as a range of linguistic resources, among them the most substantial being the National Corpus of the Russian Language (www.ruscorpora.ru), and Russian Associative Dictionary in 2 volumes (Karaulov Yu.N., Cherkasova N.V. etc. - M.: OOO "Izd.-vo Astrel'", 2002) a base of functional homonymy disambiguating collocation was built (about 30000 collocations currently). A program module was developed, which supplies the disambiguating collocation items generation according to their models in the course of forming a functional homonyms base index. The collocation model defines the functional or lexical homonym disambiguating context. Currently, about 2000 collocation models were formed on the basis of the abovementioned resources. The collocation model consists of two parts. In the first part, the collocation component word forms are represented (as a rule, binary or ternary), the second part contains code parameters of the word form inner description according to the grammatical dictionary, along with the position and the distance of the disambiguating word form in relation to the homonym.  Such model allows generating all disambiguating contexts, which are differed by the disambiguating word form. For example, the Russian collocation model 'relatively short' *'otnositel'no korotkij'* (Russian functional homonym *otnositel'no* disambiguated as an adverb) is expanded by the whole of the Russian adjective '*korotkij'* paradigm. Statistical analysis of the collocation model types allowed revealing the most frequent types where the following models belong:

- homonym (adverb/short form of adjective) + verb (disambiguating word form), for example,  'act effectively' (*'effectivno dejstvova') ;*

- homonym (noun/adjective) + noun, for example,  'close combat' ( *'blizhnij boj')*.

The homonymy disambiguation method based on the collocation models is effectively used in disambiguating complex homonymy cases, for example for 'this/it' 'eto','all' 'vse' homonyms there were compiled more than 200 disambiguating collocations.

## 2.3. Method of homonymy disambiguation based on the ontology linguistic frame

"OntoEditor+" subsystem linguistic toolbox provides integrating the ontology into various appendixes related to text processing. The linguistic toolbox realizes the functions of text corpus download; automatical statistics leading on different corpus objects; functions of pre-syntactic text processing (sentence segmentation, abbreviation recognition, homonymy disambiguation based on the special protocols of interaction with  outward lexicographical resources); forming the ontology linguistic frame; recognizing the applied ontology terms in the input text (cover task). The coupling of the ontological and linguistic (grammatical) resources is realized through the mechanisms of the ontology linguistic frame. The ontology linguistic frame is created with the help of the developed program toolbox, through which grammatical information about ontological concepts and their textual forms is fixed. Each ontological entry (как a composite term as a rule) is supplied by the corresponding grammatical information; in this process the corresponding homonymy (functional, morphological, lexical) is disambiguated. Grammatical information is transferred into the "OntoEditor+" subsystem from the "Analyzer" subsystem on the basis of special protocols of interaction. Functional, morphological, lexical homonymy disambiguation is completed on the basis of special dialogues with an expert linguist. Special procedures check the word forms in a term entry on the consistency of their grammatical characteristics. Lexical information reliability control is also carried out.  Reliability control traces the changes both in the grammatical dictionary and the ontology. Allowing for the complexity and numerous stages of the abovementioned procedures, a master of linguistic frame construction was developed in the "Ontoeditor+" subsystem; the master is called out by a command from the main menu.

The mechanism of homonymy disambiguation on the basis of the ontology linguistic frame is related to solving the task of recognizing ontological entries in the text (text covering task). For each recognized ontological entry containing a homonym, the information about the grammatical characteristics of this homonym in the context of the ontological entry is transferred. The method allows disambiguating functional, morphological and lexical homonymy.

## 2.4. The interaction of homonymy disambiguation methods

The integral technology of homonymy disambiguation includes three abovementioned methods of homonymy disambiguation. The interaction of methods in solving the task of homonymy disambiguation is provided by the interaction of the main subsystems of the "LOTA" system. "OntoEditor+" subsystem provides the realization of homonymy disambiguation method based on collocations and the one based on the ontology linguistic frame. In the development of these methods, the engineering approach is used, which allows selecting the typical frequent language cases, which are actively used in technical language. Initially, in the course of homonymy disambiguation based on these methods, general and special system knowledge is used, which is stored in various databases. "Analyzer" subsystem provides the realization of homonymy disambiguation method based on contextual rules, so the linguistic system knowledge is used. This method is a universal one, not depending on the specific problem domain. In the current version it supplies disambiguation accuracy of not less than 95 %. However, there exist some types of functional homonymy which are too complex for this method, for example, "conjunction/particle" type. The disambiguation of such homonymy is often possible only after the completion of a full syntactic analysis.

"OntoEditor+" and "Analyzer" subsystems interaction is realized on the basis of special interconnect protocols of interaction. In the course of the integral technology application, the homonymy disambiguation is carried out in two stages. On the first stage, the "Analyzer" subsystem (client) transfers a query on input text homonymy disambiguation to the "OntoEditor+" subsystem (server). The "OntoEditor+" subsystem returns the information about the disambiguated homonyms based on its own methods to the "Analyzer" subsystem. On the second stage, the "Analyzer" subsystem disambiguates the rest of the homonyms on the basis of contextual rules.

## 3. Conclusion

The integral technology of homonymy disambiguation is effectively used at the stage of pre-syntactic analysis in "LOTA" system. Essentially, the integral technology is a combination of engineering and linguistic approach to the solution of the given task. The integral technology projection is based upon the processes of coordinated interaction of different language levels, first of all the ontological level (providing the system model of knowledge about the world) with different language levels (morphological and syntactic). In the system, there was realized an effective mechanism of various subsystems interaction, which supply the realization of different methods in the integral technology.

## 4. Acknowledgements

The work has been completed with partial support of Russian Foundation Basic Research (grant № 05-07-90257).

## 5. Bibliography

[Nevzorova et al., 2001]. Nevzorova O.A., Fedunov B.E. Sistema analiza tehnicheskih tekstov "LoTA": osnovnye koncepcii i proektnye reshenija. // Izv. RAN. Teorija i sistemy upravlenija.– 2001. № 3. S. 138-149. In Russian.

[Dobrov et al., 2004] Dobrov B.V., Lukashevich N.V., Nevzorova O.A., Fedunov B.E. Metody i sredstva avtomatizirovannogo proektirovanija prikladnoj ontologii // Izvestija RAN. Teorija i sistemy upravlenija. M.: 2004. № 2. S. 58-68. In Russian.

[Nevzorova et al., 2006]. Nevzorova O..A., Zin'kina JU.V., Pjatkin N.V. Metod kontekstnogo razreshenija funkcional'noj omonimii: analiz primenimosti // Trudy mezhd. konf. Dialog'2006. M., Nauka, 2006. S. 399 – 402. In Russian.

[Nevzorova et al., 2004]. Nevzorova O.A., Nevzorov V.N. Sistema vizual'nogo proektirovanija ontologij "OntoEditor": funkcional'nye vozmozhnosti i primenenie //IX nacional'naja konferencija po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2004. M.: Fizmatlit, 2004. T. 3. S.937-945. In Russian.

## Authors' Information

***Olga Nevzorova*** – *Research Institute of Mathematics and Mechanics, Tatar State al University of Humanities and Pedagogiks, Kazan, Russia; e-mail: olga.Nevzorova@ksu.ru*

***Vladimir Nevzorov*** – *Kazan State Technical University, Russia; e-mail: nevzorov@mi.ru*

***Julia Zin'kina*** – *Kazan State University, Russia; e-mail: zjuliv@mail.ru*

***Nicolaj Pjatkin*** – *Research Institute of Mathematics and Mechanics, Kazan, Russia; e-mail: nikolaip@mail.ru*

# MATRIX: AN INCREMENTAL ALGORITHM FOR INFERRING IMPLICATIVE RULES FROM EXAMPLES BASED ON GOOD CLASSIFICATION TESTS

Xenia Naidenova

***Abstract:*** *A new incremental algorithm MATRIX is proposed for inferring implicative logical rules from examples. The concept of a good diagnostic test for a given set of positive examples lies in the basis of this algorithm.*

***Keywords:*** *learning logical rules from examples, machine learning, inductive inference, good diagnostic test*

## Introduction

Our approach to machine learning problems is based on the concept of a good diagnostic (classification) test. This concept has been advanced firstly in the framework of inferring functional and implicative dependencies from relations (Naidenova and Polegaeva, 1986). But later the fact has been revealed that the task of inferring all good diagnostic tests for a given set of positive and negative examples can be formulated as the search of the best approximation of a given classification on a given set of examples and that it is this task that all well known machine learning problems can be reduced to (Naidenova, 1996).

This paper is organized as follows. The concept of a good diagnostic test is defined and the problem of inferring all good diagnostic tests for a given classification on a given set of examples is formulated. The next section contains the description of a mathematical model underlying algorithms proposed. Then we give a decomposition of learning algorithms into subtasks that allows to costruct incremental procedure for good tests generating. The concepts of an essential value and an essential example are also introduced and an incremental learning algorithm MATRIX is described.

## The Concept of a Good Classification Test

A good diagnostic test for a given set of examples is defined as follows. Let $R$ be a table of examples and $S$ be the set of indices of examples belonging to $R$. Let $R(k)$ and $S(k)$ be the set of examples and the set of indices of examples from a given class $k$, respectively.

Denote by $FM = R/R(k)$ the examples of the classes different from class $k$. Let $U$ be the set of attributes and $T$ be the set of attributes values (values, for short) each of which appears at least in one of the examples of $R$. Let $n$ be the number of examples of $R$. We denote the domain of values for an attribute $Atr$ by $dom(Atr)$, where $Atr \in U$.

By $s(a)$, $a \in T$, we denote the subset $\{i \in S$: $a$ appears in $t_i$, $t_i \in R\}$, where $S = \{1, 2, .., n\}$. Following (Cosmadakis, et al., 1986), we call $s(a)$ the interpretation of $a \in T$ in $R$. It is possible to say that $s(a)$ is the set of indices of all the examples in $R$ which are covered by the value $a$.

Since for all $a, b \in dom(Atr)$, $a \neq b$ implies that the intersection $s(a) \cap s(b)$ is empty, the interpretation of any attribute in $R$ is a partition of $S$ into a family of mutually disjoint blocks. By $P(Atr)$, we denote the partition of $S$ induced by the values of an attribute $Atr$. The definition of $s(a)$ can be extended to the definition of $s(t)$ for any collection $t$ of values as follows: for $t$, $t \subseteq T$, if $t = a_1 a_2 ... a_m$, then $s(t) = s(a_1) \cap s(a_2) \cap ... \cap s(a_m)$.

**Definition 1**. A collection $t \subseteq T$ ($s(t) \neq \varnothing$) of values, is a diagnostic test for the set $R(k)$ of examples if and only if the following condition is satisfied: $t \not\subset t^*$, $\forall\, t^*$, $t^* \in FM$ (the equivalent condition is $s(t) \subseteq S(k)$).

To say that a collection $t$ of values is a diagnostic test for the set $R(k)$ is equivalent to say that it does not cover any example belonging to the classes different from $k$. At the same time, the condition $s(t) \subseteq S(k)$ implies that the following implicative dependency is true: 'if $t$, then $k$.

It is clear that the set of all diagnostic tests for a given set $R(k)$ of examples (call it '$DT(k)$') is the set of all the collections $t$ of values for which the condition $s(t) \subseteq S(k)$ is true. For any pair of diagnostic tests $t_i$, $t_j$ from $DT(k)$, only one of the following relations is true: $s(t_i) \subseteq s(t_j)$, $s(t_i) \supseteq s(t_j)$, $s(t_i) \approx s(t_j)$, where the last relation means that $s(t_i)$ and $s(t_j)$ are incomparable, i.e. $s(t_i) \not\subset s(t_j)$ and $s(t_j) \not\subset s(t_i)$. This consideration leads to the concept of a good diagnostic test.

**Definition 2**. A collection $t \subseteq T$ ($s(t) \neq \varnothing$) of values is a good test for the set $R(k)$ of examples if and only if the following condition is satisfied: $s(t) \subseteq S(k)$ and simultaneously the condition $s(t) \subset s(t^*) \subseteq S(k)$ is not satisfied for any $t^*$, $t^* \subseteq T$, such that $t^* \neq t$.

Good diagnostic tests possess the greatest generalization power and give a possibility to obtain the smallest number of implicative rules for describing examples of a given class $k$.

## The Characterization of Classification Tests

Any collection of values can be irredundant, redundant or maximally redundant.

**Definition 3**. A collection $t$ of values is irredundant if for any value $v \in t$ the following condition is satisfied: $s(t) \subset s(t/v)$.

If a collection $t$ of values is a good test for $R(k)$ and, simultaneously, it is an irredundant collection of values, then any proper subset of $t$ is not a test for $R(k)$.

**Definition 4**. Let $X \to v$ be an implicative dependency which is satisfied in $R$ between a collection $X \subseteq T$ of values and the value $v$, $v \in T$. Suppose that a collection $t \subseteq T$ of values contains $X$. Then the collection $t$ is said to be redundant if it contains also the value $v$.

If $t$ contains the left and the right sides of some implicative dependency $X \to v$, then the following condition is satisfied: $s(t) = s(t/v)$. In other words, a redundant collection $t$ and the collection $t/v$ of values cover the same set of examples. If a good test for $R(k)$ is a redundant collection of values, then some values can be deleted from it and thus obtain an equivalent good test with a smaller number of values.

**Definition 5**. A collection $t \subseteq T$ of values is maximally redundant if for any implicative dependency $X \to v$ which is satisfied in $R$ the fact that $t$ contains $X$ implies that $t$ also contains $v$.

If $t$ is a maximally redundant collection of values, then for any value $v \notin t$, $v \in T$ the following condition is satisfied: $s(t) \supset s(t \cup v)$. In other words, a maximally redundant collection $t$ of values covers the number of examples greater than the collection $(t \cup v)$ of values. If a diagnostic test for a given set $R(k)$ of examples is a good one and it is a maximally redundant collection of values, then by adding to it any value not belonging to it we get a collection of values which is not a good test for $R(k)$.

Any example $t$ in $R$ is a maximally redundant collection of values because for any value $v \notin t$, $v \in T$ $s(t \cup v)$ is equal to $\varnothing$.

For example, in Table 1 the collection '*Blond Bleu*' is a good irredundant test for class 1 and simultaneously it is maximally redundant collection of values. The collection '*Blond Embrown*' is a test for class 2 but it is not good test and simultaneously it is maximally redundant collection of values. The collection '*Embrown*' is a good irredundant test for class 2. The collection '*Red*' is a good irredundant test and the collection '*Tall Red Bleu*' is a maximally redundant and good test for class 1. Any example $t$ in $R$ is a maximally redundant collection of values because for any value $v \notin t$, $v \in T$ $s(t \cup v)$ is equal to $\varnothing$.

*Table* - 1. Example 1 of Data Classification. (This example is adopted from (Ganascia, 1989)).

| Index of example | Height | Color of hair | Color of eyes | Class |
|---|---|---|---|---|
| 1 | Low | Blond | Bleu | 1 |
| 2 | Low | Brown | Bleu | 2 |
| 3 | Tall | Brown | Embrown | 2 |
| 4 | Tall | Blond | Embrown | 2 |
| 5 | Tall | Brown | Bleu | 2 |
| 6 | Low | Blond | Embrown | 2 |
| 7 | Tall | Red | Bleu | 1 |
| 8 | Tall | Blond | Bleu | 1 |

*Note to Table 1 and all the following tables: the values of attributes must not be considered as the words of English language, they are the abstract symbols only .*

## An Approach for Constructing Good Irredundant Tests

Let R, T, s(t), t $\subseteq$ T be as defined earlier. We give the following propositions the proof of which can be found in (Naidenova, 1999).

**Proposition 1.**

*The intersection of maximally redundant collections of values is a maximally redundant collection.*

**Proposition 2.**

*Every collection of values is contained in one and only one maximally redundant collection with the same interpretation.*

**Proposition 3.**

*A good maximal redundant test for R(k) either belongs to the set R(k) or it is equal to the intersection of q examples from R(k) for some q, $2 \leq q \leq nt$, where nt is the number of examples in R(k).*

One of the possible ways for searching for good irredundant tests for a given class of examples is the following: first, find all good maximally redundant tests; second, for each good maximally redundant test, find all good irredundant tests contained in it. This is a convenient strategy as each good irredundant test belongs to one and only one good maximally redundant test with the same interpretation.

It should be more convenient in the following considerations to denote the set *R*(*k*) as *R*(+) (the set of positive examples) and the set *R*/*R*(*k*) as *R*(-) (the set of negative examples). We will also denote the set *S*(*k*) as *s*(+).

## The Duality of Good Diagnostic Tests

In the definition 2, we used correspondences of Galois *G* on $S \times T$ and two relations $S \to T$, $T \to S$ (Ore, 1944), (Riguet, 1948). Let $s \subseteq S$, $t \subseteq T$. We define the relations as follows:

$$S \to T: t(s) = \{\text{intersection of all } t_i: t_i \subseteq T, i \in s\} \text{ and } T \to S: s(t) = \{i: i \in S, t \subseteq t_i\}.$$

Extending *s* by an index *j** of some new example leads to receiving a more general feature of examples:

$$(s \cup j^*) \supseteq s \text{ implies } t(s \cup j^*) \subseteq t(s).$$

Extending *t* by a new value *A* leads to decreasing the number of examples possessing the general feature ‘*tA*’ in comparison with the number of examples possessing the general feature ‘*t*’:

$$(t \cup A) \supseteq t \text{ implies } s(t \cup A) \subseteq s(t).$$

We introduce the following generalization operations (functions): generalization_of(*t*) = *t′* = *t*(*s*(*t*)); generalization_of(*s*) = *s′* = *s*(*t*(*s*)).

As a result of the generalization of $s$, the sequence of operations $s \rightarrow t(s) \rightarrow s(t(s))$ gives that $s(t(s)) \supseteq s$. This generalization operation gives all the examples possessing the feature $t(s)$. As a result of the generalization of $t$, the sequence of operations $t \rightarrow s(t) \rightarrow t(s(t))$ gives that $t(s(t)) \supseteq t$. This generalization operation gives the maximal general feature for examples the indices of which are in $s(t)$.

## The Definition of Good Diagnostic Tests as dual objects

We implicitly used two generalization operations in all the considerations of diagnostic tests. Now we define a diagnostic test as a dual object, i.e. as a pair $(SL, TA)$, $SL \subseteq S$, $TA \subseteq T$, $SL = s(TA)$ and $TA = t(SL)$.

**Definition 6**. Let $PM = \{s_1, s_2, \ldots, s_m\}$ be a family of subsets of some set $M$. Then $PM$ is a Sperner system (Sperner, 1928) if the following condition is satisfied: $s_i \not\subset s_j$ and $s_j \not\subset s_i$, $\forall(i,j)$, $i \neq j$, $i, j = 1, \ldots, m$.

**Definition 7**. To find all *Good Maximally Redundant Tests* (GMRTs) for a given class $R(k)$ of examples means to construct a family $PS$ of subsets $s_1, s_2, \ldots, s_{np}$ of the set $S$ such that:

1) $s_j \subseteq S(k)$, $\forall j = 1, \ldots, np$;

2) $PS$ is a Sperner system;

3) each $s_j$ is a maximal set in the sense that adding to it the index $i$ of the example $t_i$ such that $i \notin s_j$, $i \in S$ implies $s(t(s_j \cup i)) \not\subset S(k)$. Putting it in another way, $t(s_j \cup i)$ is not a test for the class $k$, so there exists such example $t^*$, $t^* \in R(-)$ that $t(s_j \cup i) \subseteq t^*$.

The set of all GMRTs is determined as follows: $\{t\colon t(s_j), s_j \in PS, \forall j, j = 1, \ldots, np\}$.

Let $R$ be a table of examples and $S$, $T$ are defined as before. Let MUT be the set of all dual objects, that is, the set of all pairs $(s, t)$, $s \subseteq S$, $t \subseteq T$, $s = s(t)$ and $t = t(s)$. This set is partially ordered by the relation '$\leq$', where $(s, t) \leq (s^*, t^*)$ is satisfied if and only if $s \subseteq s^*$ and $t \supseteq t^*$.

The set $\Psi = (\text{MUT}, \cup, \cap)$ is an algebraic lattice, where operations $\cup$, $\cap$ are defined for all pairs $(s^*, t^*)$, $(s, t) \in \text{MUT}$ in the following way (Wille, 1992):

$$(s^*, t^*) \cup (s, t) = ((s^* \cup s), (t^* \cap t)),$$

$$(s^*, t^*) \cap (s, t) = ((s^* \cap s), (t^* \cup t)).$$

The unit element and the zero element are $(S, \varnothing)$ and $(\varnothing, T)$, respectively.

Inferring good tests is reduced to inferring for any element $(s^*, t^*) \in \text{MUT}$ all the elements nearest to it in the lattice with respect to the ordering $\leq$, that is, inferring all $(s, t)$, that $(s^*, t^*) \leq (s, t)$ and there does not exist any $(s^{**}, t^{**})$ such that $(s^*, t^*) \leq (s^{**}, t^{**}) \leq (s, t)$, or inferring all $(s, t)$, that $(s^*, t^*) \geq (s, t)$ and there does not exist any $(s^{**}, t^{**})$ such that $(s^*, t^*) \geq (s^{**}, t^{**}) \geq (s, t)$. Inferring the chains of lattice elements ordered by the inclusion relation lies in the foundation of generating all types of diagnostic tests:

(1) $s_0 \subseteq \ldots \subseteq s_i \subseteq s_{i+1} \subseteq \ldots \subseteq s_m$ $(t(s_0) \supseteq t(s_1) \supseteq \ldots \supseteq t(s_i) \supseteq t(s_{i+1}) \supseteq \ldots \supseteq t(s_m))$,

(2) $t_0 \subseteq \ldots \subseteq t_i \subseteq t_{i+1} \subseteq \ldots \subseteq t_m$ $(s(t_0) \supseteq s(t_1) \supseteq \ldots \supseteq s(t_i) \supseteq s(t_{i+1}) \supseteq \ldots \supseteq s(t_m))$.

We will use only the chain (1) for inferring good diagnostic tests.

## Decomposition of Good Classification Tests Inferring into Subtasks

Now we consider some decompositions of the problem that provide the possibility to restrict the domain of searching, to predict, in some degree, the number of tests, and to choose tests with the use of essential values and/or examples. We consider three kinds of subtasks: for a given set of positive examples

1) given a positive example $t$, find all GMRTs contained in $t$;

2) given a non-empty collection of values $X$ (maybe only one value A) such that it is not a test, find all GMRTs containing $X$;

3) given a non-empty collection of values $X$ (maybe only one value A) such that it is not a test and a positive example $t$, $X \subseteq t$, find all GMRTs containing $X$ and simultaneously contained in $t$.

## Forming the Subtasks

**The subtask of the first kind**. We introduce the concept of an example's projection proj($R$)[$t$] of a given positive example $t$ on a given set $R$(+) of positive examples. The proj($R$)[$t$] is the set $Z$ = {z: ($z$ is non empty intersection of $t$ and $t'$) & ($t' \in R$(+)) & ($z$ is a test for $R$(+))}.

If the proj($R$)[$t$] is not empty and contains more than one element, then it is a subtask for inferring all GMRTs that are in $t$. If the projection contains one and only one element equal to $t$, then $t$ is a GMRT.

**The subtask of the second kind.** We introduce the concept of an attributive projection proj($R$)[$A$] of a given value $A$ on a given set $R$(+) of positive examples.

The projection proj($R$)[$A$] = {$t$: ($t \in R$(+)) & ($A$ appears in $t$)}. Another way to define this projection is: proj($R$)[$A$] = {$t_i$: $i \in$ ($s(A) \cap s(+)$)}. If the attributive projection is not empty and contains more than one element, then it is a subtask of inferring all GMRTs containing a given value $A$. If $A$ appears in one and only one example, then $A$ does not belong to any GMRT different from this example.

Forming the projection of $A$ makes sense if $A$ is not a test and the intersection of all positive examples in which $A$ appears is not a test too, i.e. $s(A) \not\subseteq s(+)$ and $t' = t(s(A) \cap s(+))$ is not a test for a given set of positive examples.

*Denote* the set {$s(A) \cap s(+)$} *by* splus($A$). Generally, we can consider the projection proj($R$)[$X$], $X \subseteq T$.

**The subtask of the third kind.** Now we introduce proj($R$)[$t$ x $A$] = proj($R$)[$A$ x $t$], where $A \subseteq t$ . In order to construct this projection the folowing steps are implemented: 1) to select all examples $t_i \in R$(+) containing $A$; 2) In each selected example, it is necessary to take only the values which appear in $t$. The result is the following:

Z = {z: $z = t \cap t_i, \neq \emptyset$, $t_i \in R(+)$, $A \subseteq t_i$, $A \subseteq t$, ($z$ is a test for $R$(+))}.Generally, we can consider the projection proj($R$)[$t$ x $X$] = proj($R$)[$X$ x $t$], where $X \subseteq t$ .

To make the operation of forming a projection perfectly clear we construct the projection proj($R$)[ $t_2$ x *Brown*] on the examples of the second class where $t_2$ = '*Low Brown Bleu*' (Table 1). This projection includes $t_2$ and the intersections of $t_2$ with the other positive examples of Class 2, containing the value '*Brown*', i.e. with the examples $t_3$, $t_5$ (Table 3).

In order to check whether an element of the projection is a test or not we use the function to_be_test($t$) in the following form: to_be_test($t$) = if $s(t) \subseteq s(+)$ then *true* else *false*, where $s$(+) is the set of indices of positive examples, $s(t)$ is the set of indices of all positive and negative examples containing $t$. If $s$(-) is the set of indices of negative examples, then $S = s(+) \cup s(-)$ and $s(t) = \{i: t \subseteq t_i, i \in S\}$.

The subtask turns out to be very simple because the intersection of all the rows of the projection is a test for the second class: $t$({2,3,5}) = '*Brown*', $s$(*Brown*) = {2,3,5} and {2,3,5} $\subseteq s(+)$.

*Table - 3.* The Intersections of Example $t_2$ with the Examples of Class 2. ($t_2$ x *Brown*)

| Index of example | Height | Color of hair | Color of eyes | test? |
|---|---|---|---|---|
| 2 | Low | Brown | Bleu | Yes |
| 3 | | Brown | | Yes |
| 5 | | Brown | Bleu | Yes |

## Reducing the Subtasks

The following theorem gives the foundation for reducing projections of any kind. The proof of this theorem can be found in (Naidenova et al., 1995).

**THEOREM 1.**

Let *A* be a value from *T*, *X* be a maximally redundant test for a given set *R*(+) of positive examples and *s*(*A*) ⊆ *s*(*X*). Then *A* does not belong to any maximally redundant good test for *R*(+) different from *X*.

Deleting values from a projection can imply deleting rows stopping to be tests. Deleting rows from a projection can imply deleting values satisfying the condition of the Theorem 1. To illustrate the way of reducing projections, we consider another partition of the rows of Table 1 into the sets of positive and negative examples as shown in Table 4.

*Table* - 4. The Example 2 of a Data Classification.

| Index of example | Height | Color of hair | Color of eyes | Class |
|---|---|---|---|---|
| 1 | Low | Blond | Bleu | 1 |
| 2 | Low | Brown | Bleu | 1 |
| 3 | Tall | Brown | Embrown | 1 |
| 4 | Tall | Blond | Embrown | 2 |
| 5 | Tall | Brown | Bleu | 2 |
| 6 | Low | Blond | Embrown | 2 |
| 7 | Tall | Red | Bleu | 2 |
| 8 | Tall | Blond | Bleu | 2 |

Let s(+) be equal to {4,5,6,7,8}. The value ‘Red’ is a test for positive examples because s(Red) = splus(Red) = {7}. Delete ‘Red’ from the projection. The value ‘Bleu’ is not a test because s(Bleu) = {1,2,5,7,8}. But splus(Bleu) = {5,7,8} and t(splus(Bleu)) = ‘Tall Bleu’ is a test for Class 2. Delete ‘Bleu’ from examples of Class 2 as shown in Table 5. Now the rows $t_5$ and $t_7$ are not tests for Class 2 and they can be deleted.

*Table* - 5. The Example of a projection reduced.

| Index of example | Height | Color of hair | Color of eyes | Class | Test? |
|---|---|---|---|---|---|
| 1 | Low | Blond | Bleu | 1 | Yes |
| 2 | Low | Brown | Bleu | 1 | Yes |
| 3 | Tall | Brown | Embrown | 1 | Yes |
| 4 | Tall | Blond | Embrown | 2 | Yes |
| 5 | Tall | Brown |  | 2 | No |
| 6 | Low | Blond | Embrown | 2 | Yes |
| 7 | Tall |  |  | 2 | No |
| 8 | Tall | Blond |  | 2 | Yes |

## Choosing Values and Examples for the Formation of Subtasks

**Definition 8**. Let *t* be a collection of values that is a test for a given set of positive examples. We say that the value *A* in *t* is essential if (*t*/*A*) is not a test for a given set of positive examples.

Generally, we are interested in finding the maximal subset *sbmax*(*t*) ⊂ *t* such that *t* is a test but *sbmax*(*t*) is not a test for a given set of positive examples. Then *sbmin*(*t*) = *t*/ *sbmax*(*t*) is the minimal set of essential values in *t*.

**Definition 9**. Let *s* be a subset of indices of positive examples; assume also that *t*(*s*) is not a test. The example $t_j$, *j* ∈ *s* is to be said an essential one if *t*(*s*/*j*) proves to be a test for a given set of positive examples.

Generally, we are interested in finding the maximal subset *sbmax*(*s*) ⊂ *s* such that *t*(*s*) is not a test but *t*’ = *t*(*sbmax*(*s*)) is a test for a given set of positive examples. Then *sbmin*(*s*) = *s*/*sbmax*(*s*) is the minimal set of indices of essential examples in *s*.

In order to construct the projection proj(*R*)[$t_i$ x *A*] = proj(*R*)[*A* x $t_i$] for a subtask of the third kind it is very convenient to take $t_i$ and A such that *i* ∈ *sbmin*(*splus*(*A*)) and *A* ∈ *sbmin*($t_i$).

## An Approach for Searching for Essential Values

Let t be a test for positive examples. Construct the set of intersections {t ∩ t': t' ∈ R(-)}. It is clear that these intersections are not tests for positive examples. Take one of the intersections with the maximal number of values in it. The values complementing the maximal intersection in t is one of the minimal sets of essential values in t.

Return to Table 6. Exclude the value 'Red' (we know that 'Red' is a test for Class 2) and find the minimal subsets of essential values for $t_4$, $t_5$, $t_6$, $t_7$, and $t_8$. The result is in Table 6.

*Table* - 6. The Essential Values for the Examples *$t_4$, $t_5$, $t_6$, $t_7$, and $t_8$.*

| Index of example | Height | Color of hair | Color of eyes | Subsets of essential values | Class |
|---|---|---|---|---|---|
| 1 | Low | Blond | Bleu | | 1 |
| 2 | Low | Brown | Bleu | | 1 |
| 3 | Tall | Brown | Embrown | | 1 |
| 4 | Tall | Blond | Embrown | {Blond} | 2 |
| 5 | Tall | Brown | Bleu | {Bleu}, {Tall} | 2 |
| 6 | Low | Blond | Embrown | {Embrown} | 2 |
| 7 | Tall | | Bleu | {Tall}, {Bleu} | 2 |
| 8 | Tall | Blond | Bleu | {Tall} | 2 |

## An Approach for Searching for Essential Examples

Let *STGOOD* be the partially ordered set of elements *s* satisfying the condition that *t*(*s*) is a good maximally redundant test (GMRT) for *R*(+). We can use the set *STGOOD* to find indices of essential examples in some subset *s** of indices for which *t*(*s**) is not a test. Let *s** = {$i_1$, $i_2$, … , $i_q$}. Construct the set of intersections {*s** ∩ *s'*: *s'*∈ *STGOOD*}. Any obtained intersection *s** ∩ *s'* corresponds to a test for positive examples. Take one of the intersections with the maximal number of indices. The subset of *s** complementing in *s** the maximal intersection is one of the minimal sets of indices of essential examples in *s**. For instance, *s** = {2,3,4,7,8}, *s'* = {2,3,4,7}, *s'* ∈ *STGOOD*, hence 8 is the index of essential example $t_8$ in *s**.

In the beginning of inferring GMRTs, the set *STGOOD* is empty. The procedure with the use of which a quasi-maximal subset of *s** that corresponds to a test is obtained has been described in (Naidenova, 2005).

## MATRIX – an Algorithm for Incremental Inferring Good Maximally Redundant Diagnostic Tests

Incremental learning is necessary when a new portion of training examples becomes available over time. Suppose that each new example comes with the indication of its class membership. The following actions are necessary with the arrival of a new example:

- Check whether it is possible to perform generalization of some existing GMRTs for the class to which the new example belongs (class of positive examples), i.e., whether it is possible to extend the set of examples covered by some existing GMRTs or not.

- Infer all the GMRTs contained in the new example.

- Check the validity of the existing GMRTs for negative examples, and if it necessary:

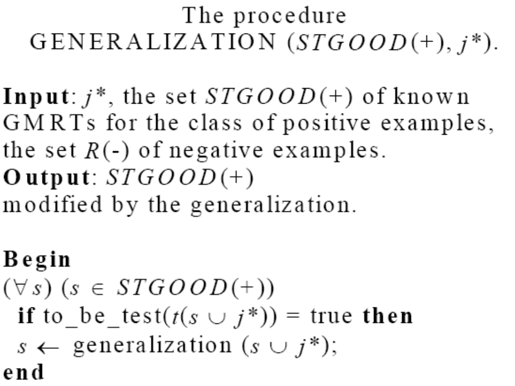
Modify tests that are not valid (test for negative examples is not valid if it is included in a positive example, i.e., in other words, it accepts an example of positive class).

Thus the process of inferring all the GMRTs is divided into the subtasks that conform to three acts of reasoning:

- Pattern recognition or using already known rules (tests) for determining the class membership of a new positive example and generalization of these rules (deductive reasoning and increasing the inductive base of already existing knowledge). This act is performed by the procedure GENERALIZATION (STGOOD, j*) (Figure 1).

- Inferring new rules (tests) that are included in a new positive example. This act can be reduced to the subtask of the first kind or to the subtask(s) of the third kind.

- Correcting rules (tests) of alternative (negative) classes that accept a new positive example (deductive and inductive diagnostic reasoning to modify knowledge). This act can be reduced to the subtask of the second kind or the subtask(s) of the third kind.

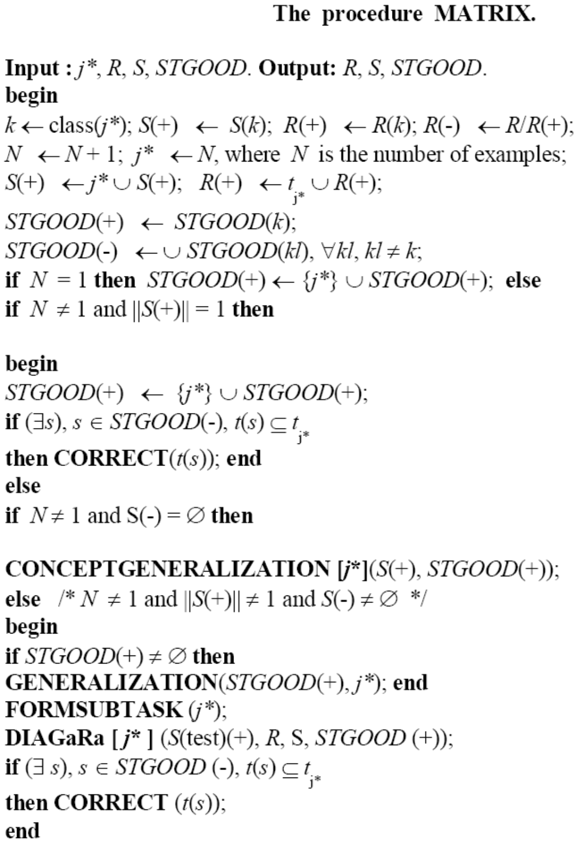


```
The procedure
GENERALIZATION ($STGOOD(+), j^*$).

Input: $j^*$, the set $STGOOD(+)$ of known
GMRTs for the class of positive examples,
the set $R(-)$ of negative examples.
Output: $STGOOD(+)$
modified by the generalization.

Begin
($\forall s$) ($s \in STGOOD(+)$)
  if to_be_test($t(s \cup j^*)$) = true then
  $s \leftarrow$ generalization ($s \cup j^*$);
end
```

*Figure 1.* The Procedure for Generalizing the Existing GMRTs.



```
The procedure MATRIX.

Input : $j^*$, $R$, $S$, $STGOOD$. Output: $R$, $S$, $STGOOD$.
begin
$k \leftarrow$ class($j^*$); $S(+) \leftarrow S(k)$; $R(+) \leftarrow R(k)$; $R(-) \leftarrow R/R(+)$;
$N \leftarrow N + 1$; $j^* \leftarrow N$, where $N$ is the number of examples;
$S(+) \leftarrow j^* \cup S(+)$; $R(+) \leftarrow t_{j^*} \cup R(+)$;
$STGOOD(+) \leftarrow STGOOD(k)$;
$STGOOD(-) \leftarrow \cup STGOOD(kl), \forall kl, kl \neq k$;
if $N = 1$ then $STGOOD(+) \leftarrow \{j^*\} \cup STGOOD(+)$; else
if $N \neq 1$ and $\|S(+)\| = 1$ then

begin
$STGOOD(+) \leftarrow \{j^*\} \cup STGOOD(+)$;
if $(\exists s), s \in STGOOD(-), t(s) \subseteq t_{j^*}$
then CORRECT($t(s)$); end
else
if $N \neq 1$ and $S(-) = \varnothing$ then

CONCEPTGENERALIZATION [$j^*$]($S(+)$, $STGOOD(+)$);
else   /* $N \neq 1$ and $\|S(+)\| \neq 1$ and $S(-) \neq \varnothing$  */
begin
if $STGOOD(+) \neq \varnothing$ then
GENERALIZATION($STGOOD(+)$, $j^*$); end
FORMSUBTASK ($j^*$);
DIAGaRa [ $j^*$ ] ($S$(test)(+), $R$, S, $STGOOD$ (+));
if $(\exists s), s \in STGOOD(-), t(s) \subseteq t_{j^*}$
then CORRECT ($t(s)$);
end
```

*Figure 2.* The Incremental Procedure MATRIX

All these subtasks can be solved by DIAGaRa, the basic procedure for inferring GMRTs (Naidenova, 2005).

We must consider four possible situations that can take place when a new example comes to the learning system: 1) the data base is empty; 2) the data base contains only examples of the class to which a new example belongs; 3) The data base contains only examples of the negative class with respect to a new example; 4) the data base contains examples both of the positive and the negative classes. Case 2 conforms to the generalization process taking into account only the similarity relation between examples of the same class. This problem is known in the literature as inductive inference of generalization hypotheses or unsupervised generalization. An algorithm for solving this problem in the framework of a mathematical model based on Galois's connections can be found in (Kuznetzov, 1993). Let CONCEPTGENERALIZATION [j*](S(+), STGOOD(+)) be the procedure of generalization of positive examples in the absence of negative examples.

The algorithm MATRIX for inferring GMRTs is presented in Figure 2. CORRECT (*t*) is the procedure of modifing test *t*, FORMSUBTASK(*j*) is the procedure of forming subtasks.

## Appendix: An Example of Using the Algorithm MATRIX

The data in Table 7 are intended for processing by the incremental learning procedure MATRIX. This table is adopted from (Quinlan and Rivest, 1989). The sets STGOOD(1) and STGOOD(2) in Tables 8 accumulate the collections of indices that correspond to the GMRTs for the examples of Class 1 and Class 2, respectively, at each step of the algorithm. Only one new example is added at each step of the procedure.

*Table - 7.* The Data for Processing by the Incremental Procedure MATRIX

| Index of example | Outlook | Temperature | Humidity | WindY | Class |
|---|---|---|---|---|---|
| 1 | Sunny | Hot | High | No | 1 |
| 2 | Sunny | Hot | High | Yes | 1 |
| 3 | Overcast | Hot | High | No | 2 |
| 4 | Rain | Mild | High | No | 2 |
| 5 | Rain | Cool | Normal | No | 2 |
| 6 | Rain | Cool | Normal | Yes | 1 |
| 7 | Overcast | Cool | Normal | Yes | 2 |
| 8 | Sunny | Mild | High | No | 1 |
| 9 | Sunny | Cool | Normal | No | 2 |
| 10 | Rain | Mild | Normal | No | 2 |
| 11 | Sunny | Mild | Normal | Yes | 2 |
| 12 | Overcast | Mild | High | Yes | 2 |
| 13 | Overcast | Hot | Normal | No | 2 |
| 14 | Rain | Mild | High | Yes | 1 |

*Table – 8.* The Records of Step-by-Step Results of the Incremental Procedure MATRIX.

| j* | class (J*) | STGOOD(1), STGOOD(2) | j* | class (J*) | STGOOD(1), STGOOD(2) |
|---|---|---|---|---|---|
| {1} | 1 | STGOOD(1):{1} | {8} | 1 | STGOOD(1):{1,2,8}, {6} |
| {2} | 1 | STGOOD(1):{1,2} | {9} | 2 | STGOOD(2):{3,7},{4,5},{5,9} |
| {3} | 2 | STGOOD(2):{3} | {10} | 2 | STGOOD(2):{3,7},{4,5,10},{5,9,10} |
| {4} | 2 | STGOOD(2):{3}, {4} | {11} | 2 | STGOOD(2):{3,7},{4,5,10}, {5,9,10},{10,11},{9,11} |
| {5}, | 2 | STGOOD(2):{3},{4,5} | {12} | 2 | STGOOD(2):{3,7,12},{4,5,10}, {5,9,10},{10,11},{9,11},{11,12} |
| {6} | 1 | STGOOD(1):{1,2},{2,6} | {13} | 2 | STGOOD(2):{3,7,12,13},{4,5,10},{5,9,10,13},{10,11},{9,11},{11,12} |
| {7} | 2 | STGOOD(2):{3,7},{4,5}<br>STGOOD(1):{1,2}, {6} | {14} | 1 | STGOOD(1):{1,2,8}, {6,14}<br>STGOOD(2):{3,7,12,13},{4,5,10}, {5,9,10,13},{10,11},{9,11} |

Table 9 contains all the GMRTs obtained for the examples of Class 1 and Class 2: TGOOD(*j*) = {*t*(*s*): *s* ∈ STGOOD(*j*)}.

This application of the algorithm MATRIX did not require any call for the procedure DIAGARA. Only one address was necessary to the procedure CONCEPTGENERALIZATION[*j*] in the beginnig of inferring GMRTs. Only two addresses were necessary to the procedure CORRECT(*t*(*s*)) after the arrival of Example 7 and Example 14. All the subtasks were very simple and allowed reading the tests directly without calling for the procedure DIAGaRa.

*Table* - 9. The Sets TGOOD(1) and TGOOD(2) Produced by the Procedure MATRIX

| TGOOD(1) | TGOOD(2) |
|---|---|
| Sunny High | Rain No |
| Rain Yes | Normal No |
| - | Mild Normal |
| - | Sunny Normal |
| - | Overcast |

## Bibliography

[Cosmadakis et al., 1986] S. Cosmadakis, P. C. Kanellakis, N. Spyratos, “Partition Semantics for Relations”, *Journal of Computer and System Sciences*, Vol. 33, No. 2, pp.203-233, 1986. [Demetrovics and Vu, 1993] J. Demetrovics and D. T. Vu, “Generating Armstrong Relation Schemes and Inferring Functional Dependencies from Relations”, *International Journal on Information Theory & Applications*, Vol. 1, No. 4, pp.3-12, 1993.

[Ganascia, 1989] J.- Gabriel. Ganascia, “EKAW - 89 Tutorial Notes: Machine Learning”, *Third European Workshop on Knowledge Acquisition for Knowledge-Based Systems*, Paris, France, pp. 287-296, 1989.

[Kuznetsov, 1993] S. O. Kuznetsov, “Fast Algorithm of Constructing All the Intersections of Finite Semi-Lattice Objects”, *Proceedings of VINITI*, Series 2, No. 1, pp. 17-20, 1993.

[Naidenova, 1996] X. A. Naidenova, “Reducing Machine Learning Tasks to the Approximation of a Given Classification on a Given Set of Examples”, *Proceedings of the 5-th National Conference at Artificial Intelligence*, Kazan, Tatarstan, Vol. 1, pp. 275-279, 1996.

[Naidenova, 1999] X. A. Naidenova, “The Data-Knowledge Transformation”, in: “*Text Procesing and Cognitive Technologies”, Paper Collection*, editor Solovyev, V. D., - Pushchino, Russia, Vol. 3, pp. 130-151, 1999.

[Naidenova, 2005] X. A. Naidenova, “DIAGARA : An Incremental Algorithm for Inferring Implicative Rules from Examples”, International Journal ‘Information Theories & Applications’, Vol. 12, pp. 171-186.

[Naidenova et al., 1995] X. A. Naidenova, M. V. Plaksin, V. L. Shagalov, “Inductive Inferring All Good Classification Tests”, *Proceedings of International Conference “Knowledge-Dialog-Solution*”, Jalta, Ukraine, Vol. 1, pp.79-84, 1995b.

[Naidenova and Polegaeva, 1986] X. A. Naidenova, J. G. Polegaeva, “An Algorithm of Finding the Best Diagnostic Tests”, *The 4-th All Union Conference “Application of Mathematical Logic Methods”*, Theses of Papers, Mintz, G; E, Lorents, P. P. (Eds), Institute of Cybernetics, National Acad. of Sciences of Estonia, Tallinn, Estonia, pp. 63-67, 1986.

[Ore, 1944] O. Ore, “*Galois Connexions*”, Trans. Amer. Math. Society, Vol. 55, No. 1, pp. 493-513, 1944.

[Quinlan and Rivest,1989] J. R. Quinlan, and R. L. Rivest, “Inferring Decision Trees Using the Minimum Description Length Principle”, Information and Computation, Vol. 80, No. 3, pp. 227-248, 1989.

[Riguet, 1948] J. Riguet, “Relations Binaires, Fermetures, Correspondences de Galois”, *Bull. Soc. Math*., France, Vol. 76., No 3, pp.114-155, 1948.

[Sperner, 1928] E. Sperner, “Eine satz uber Untermengen einer Endlichen Menge”. *Mat.* Z., Vol. 27, No. 11, pp. 544-548, 1928.

[Wille, 1992] R. Wille, “Concept Lattices and Conceptual Knowledge System”, *Computer Math. Appl*., Vol. 23, No. 6-9, pp. 493-515, 1992.

## Author's Information

***Naidenova Xenia Alexandrovna -*** *Military medical academy, Saint-Petersburg, Stoikosty street, 26-1-248, e-mail:* ***naidenovaxen@gmail.com***

# SEMANTIC MODELLING FOR PRODUCT LINES ENGINEERING

**Mikhail Roshchin, Peter Graubmann, Valery Kamaev**

***Abstract****: The aim of our work is to present solutions and a methodical support for automated techniques and procedures in domain engineering, in particular for variability modelling. Our approach is based upon Semantic Modelling concepts, for which semantic description, representation patterns and inference mechanisms are defined. Thus, model-driven techniques enriched with semantics will allow flexibility and variability in representation means, reasoning power and the required analysis depth for the identification, interpretation and adaptation of artefact properties and qualities.*

***Keywords****: Variability Modeling, Semantic Modelling, Product Line Engineering, MDA.*

## Problem Statement

Let us assume that we require a software system that is specifically tailored to rely on our needs; that is valid and consistent within the reality of the environment and involved domains. But the cost issue plays an important role, and the development of specific and generic products is not that cost-effective as we expect. For reduction of costs, software engineering aims of an increasing reuse by collecting and composing artefacts and assets, components and products into complex systems and new applications. Also, the ideas and concepts of families of systems and product lines are formalized for easier way of future artefact implementation.

Behind the system composition process and derivation of new product implementation based on reuse, there is a heavy and massive layer of computing model-based procedures. Therefore models are considered to be interchangeable and valid for particular task and requirements. Model-driven engineering introduces models together with techniques for system design and artefact adaptation into business process and software lifecycle. At the same time, domain engineering provides with deep understanding of the targeted domain and its specifics, and variability modelling specifies commonalities, variants and features, their relations and restrictions, for the whole product family of systems realized and presented as models.

But, due to the high diversity of modelling techniques, distinctions between models of different aspects, domain-dependent and company-specific knowledge and specifications, the reuse is still difficult and non-trivial. The lack of formal semantics for MDAs [Greenfield, 2004], domain and variability models and requirements engineering, affects with the impossibility of pragmatic and cost-effective solution for automated reasoning techniques. The absence of well-established semantic model does not allow us to provide self-configuring techniques, consistency verification procedures and advanced selection of valid artefacts.

Domain engineering has been proved to handle a high priority share in the entire model-driven engineering, but the state of the art shows that the lack of formal semantics and proper tool support for automated reasoning have hindered the development in this area. So far, the knowledge representation techniques based on semantics are being developed in isolation from software engineering activities, in particular from feature and variability modelling. Existing semantic approaches are not aligned with the entire modelling process, and need an advanced review on conceptual level for the proper role and place of formal methods within existing software engineering streams.

No doubts, that model-driven architecture, domain engineering, variability and feature models are perfect approaches themselves. But there is an urgent need to enrich them with formal methods of knowledge representation and benefit from that in the near future [Assmann, 2003].

Here we focus just on variability modelling, assuming that our approach can be used in a wider range, in particular for MDE and domain engineering. It is shown how semantic modelling can handle and support variability modelling, and how software engineering will benefit from that.

## Semantic Modelling Approach

The need for variability modelling and its role within the scope of domain engineering in the software development area are obvious and generally accepted. Variability modelling becomes necessary when we derive new specifications for further artefact implementation from the set of commonalities and variants related to particular system family. Also, it is important for describing dynamical behaviour of systems. We take a variability model as proposed by [Buehne, 2005]. But still, there are open questions and issues, mentioned by different research institutes and software communities, which have hindered the expected development in the field of knowledge reuse.

The automation in general is based on a set of specific methods and procedures, that allow us to substitute the human participation with some formal algorithms. The design automation needs assistance in making decisions and solving problems in analyzing requirements from customers, and mapping them onto our product family description – variability model. But applying selection procedures to variability model is not sufficient. The project manager has to be aware of existing components, which are ready for reuse. Thus component repository and its participation in a decision procedure play an important role (see 0).

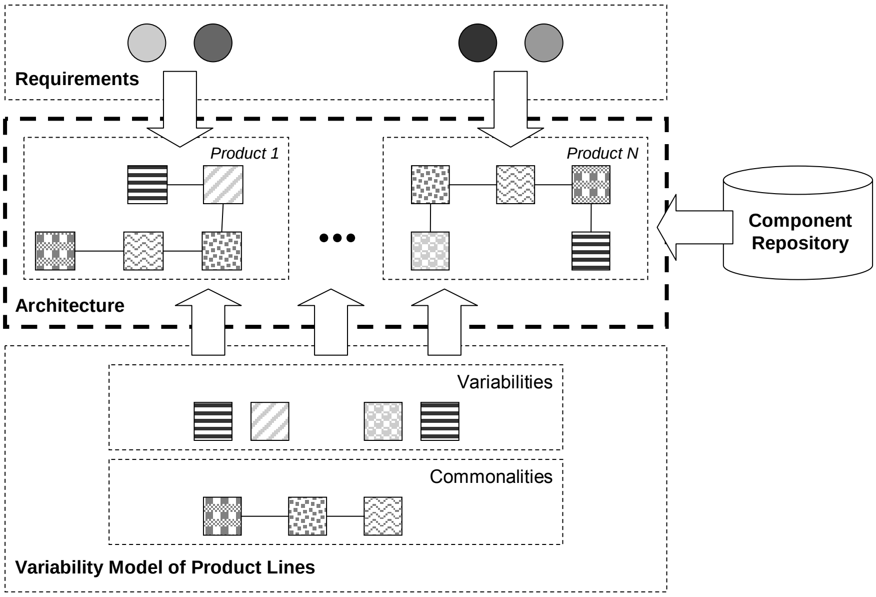


Figure 1. Software Design: from Requirements to

Our goal is to provide proper methods and tool support for formally expressing, processing and analyzing models and variants. We need to introduce formal semantics and appropriate automated reasoning techniques. Based on that, we achieve explicit consideration of environmental, behavioural and business model aspects, interoperability of the diversity of components. Semantic modelling allows acquisition, interpretation and adaptation of different variability models into one decision process.

Our Semantic Modelling approach presented in [Graubmann, 2006] is based on two concepts, which are significant for the whole procedure and aligned with requirements to semantics. These concepts are Logic-on-Demand and Triple Semantic Model (see 0).

The Triple Semantic Model Concept

Our Semantic Model is based on the principles of the Triple Semantic Model concept, which aims in defining a distributed computing model for the whole lifecycle of variability model and to provide mechanisms to distinguish between different entities represented within that model. It consists of three levels: the Ontology Level, the Dynamic Annotation Level, and the Annotation Level. The ontologies on the *Ontology Level* are intended to provide a general framework, in most cases based on a specific application domain, to describe any kind of product line and related information. Since ontologies enforce proper definitions of the concepts in the application domain, they also play an essential role in standardising the definitions of component or service properties [0], requirements and interfaces with respect to their domain. Ontologies hold independently from actual circumstances, the situation in the environment or the actual time. However, such dependencies from actual, dynamically changing circumstances do have an important influence in the compositional approach. Hence, rules determining how to cope with this dynamicity have to be provided if one has to include it into the reasoning. They are specified on the *Dynamic Annotation Level*: Dynamic annotations play the role of mediators between the ontology and the static semantic annotations that describes the artefact variants and features, and in particular its requirements with respect to reuse and composition. It becomes possible to express behavior variants, and options depending on dynamic features, and it enables the reasoning about particular situations and dynamically changing lifecycle conditions. The *Annotation Level* comprises the static descriptions of the properties and qualities of artefacts.

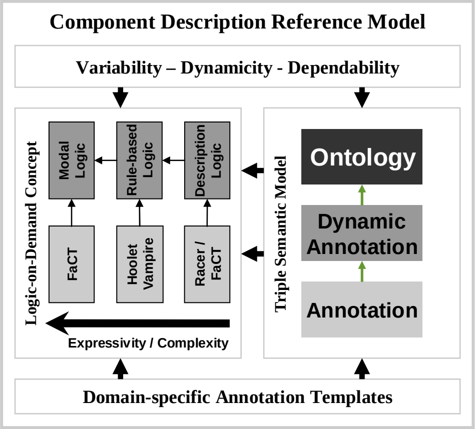


Figure 2. Semantic Model

The Logic-on-Demand Concept

Semantic modelling of products and families involves a large variety of information from different application domains and of various categories, like terms and definitions, behaviour rules, probability relations, and temporal properties. Thus, it seems to be the obvious to choose the most expressive logical formalism that is capable to formulate and formalise the entire needed information. But, doing so very likely results in severe decidability problems.

Our semantic modelling approach, based on the concept of Logic-on-Demand (LoD), is supposed to overcome the problems of complexity of formal semantics by accommodating the expressivity of the proposed ontology languages to the varying needs and requirements, in particular with respect to decidability. The main purpose of the LoD concept is to provide an adequate and adaptive way that is based on uniform principles for describing all the notions, relations and rules, the behaviour and anything else that proves necessary during the component or service annotation process. To achieve this, LoD means to define a basic logical formalism that is adequate and tailored to the application domain and to incorporate additional logic formalisms and description techniques with further expressivity as optional features that can be used whenever needed. These additional formalisms share notions and terms with the basic formalism which will be grounded syntactically in OWL and semantically in the description logics.

Thus, semantic modelling is applied for both formal description of Variability Models in Product Line Engineering and software components. The meta-model of variability description can be easily obtained by substitution of nodes and edges on modelling graph by classes and property relations from Description Logic. An instances of the classes represent specific notions and features from product family description.

A brief sketch of the component or service selection and composition process according to the Triple Semantic Model now comprises the following steps:

- Requirements on a component or service to be integrated into the system are collected. They serve as selection criteria when candidates are checked.
- The dynamic annotation and the (static) annotation of the candidate component/service are used to create an annotation that is valid in the given situation and time.
- This annotation is analysed and compared with the initial requirements.
- If the result shows that the component fits, it may be integrated (what may include the generation of data transformations in order to adapt the interfaces).

So, software engineers and system developers have to define their specific view on the concrete component/service and they naturally formulate this information in the terminology of the domain or system family to which the component/service belongs. If the annotating is done properly, we have the complete information about the component/service properties. Due to the Logic-on-Demand concept, this information is available not only for the developers but also presented in a form that is readable for automated acquisition and adaptation tools and thus, it allows reasoning and derivation of additional information.

The validation of the approach includes three aspects:

- Evaluation of the applied formal semantics with respect to sufficiency and decidability. The work on Logic-on-Demand concept is still in progress. Our intention is to avoid complexity issues and to guarantee adequate system response time.
- Feasibility issue. So far, we implement proposed techniques in a prototype tool. It covers the whole lifecycle of semantic modelling – starting out from defining semantic patterns and domain-specific information and eventually providing fully automated composition techniques based on semantic models.
- Estimating the additional cost and time for semantic modelling according an approach. Do a creation of semantic models and an implementation of formal methods and techniques really pay off in software engineering? This question touches a most important issue of our work and will be investigated in concordance the prototype tool development.

## Conclusion

Introducing a well-structured semantic modelling procedure for variability modelling provides with flexibility of representation means and methods. It allows correct (self) configuration and composition of different shares among the whole set of domain pieces during the entire modelling process, by taking into account behavioural, environmental and business aspects. Improved acquisition, interpretation and adaptation techniques allow to increase reuse among different domains and system families. Formal methods in modelling support automated derivation of an executable and sufficient model for further system or artefact implementation based on semantic mapping of requirements criteria and the given set of features and variants.

Our approach proposes an annotation process and its semantic extensions through knowledge-based techniques as the basis for semantic modelling. The Component Description Reference Model (semantic model for software components) structures the annotation process and introduces flexibility with respect to the description mechanisms what allows for a trade-off between expressivity and complexity and the selection of the appropriate reasoning tools. It is based on the Logic-on-Demand concept which means to achieve a proper compromise between existing semantic approaches and it proposes a hybrid knowledge-based solution for annotating software components. By introducing the Triple Semantic Model concept we allow an integration of means to adequately express dynamicity and variability into an modelling process.

There are, however, still open questions. We continue to work on automatic mapping of different ontologies from heterogeneous environments and knowledge application domains, on integration of different logic formalisms for component and service description, and on the mutual adaptation of problem solvers based on different logics and inference algorithms, to name but a few of the themes to be tackled in the future. We also will particularly focus on tool support for the proposed techniques to demonstrate the expected benefits, and we will later on integrate the techniques into a software development environment.

## Bibliography

[Graubmann, 2006] Peter Graubmann, Mikhail Roshchin, "Semantic Annotation of Software Components", Accepted for 32th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), Component-Based Software Engineering Track, 2006

[Greenfield, 2004] Jack Greenfield, Keith Short, *Software Factories*. Wiley Publishing, 2004

[Pahl, 2002] Claus Pahl, “Ontologies for Semantic Web Components,” ERCIM News No 51, Oct. 2002. http://www.ercim.org/publication/Ercim_News/enw51/pahl.html

[Assmann, 2003] Uwe Assmann, Steffen Zschaler, Gerd Wagner, *Ontologies, Meta-Models and the Model-Driven Paradigm,* 2003

[Buehne, 2005] Stan Buehne, Kim Lauenroth, Klaus Pohl: *Modelling Requirements Variability across Product Lines.* Proceedings of the 2005 13th IEEE International Conference on Requirements Engineering (RE '05), IEEE

## Authors' Information

***Roshchin Mikhail** - PhD Student of Volgograd State Technical University, working in collaboration with CT SE, Siemens AG; e-mail: roshchin@gmail.com*

***Graubmann Peter –** Senior Engineer, CT SE, Siemens AG; e-mail: peter.graubmann@siemens.com*

***Kamaev Valery –** Prof., Chair Head for CAD Technologies of Volgograd State Technical University, e-mail: kamaev@cad.vstu.ru*

# A FRAMEWORK FOR FAST CLASSIFICATION ALGORITHMS

**Thakur Ghanshyam, R.C.Jain**

***Abstract**: Today, due to globalization of the world the size of data set is increasing, it is necessary to discover the knowledge. The discovery of knowledge can be typically in the form of association rules, classification rules, clustering, discovery of frequent episodes and deviation detection. Fast and accurate classifiers for large databases are an important task in data mining. There is growing evidence that integrating classification and association rules mining, classification approaches based on heuristic, greedy search like decision tree induction. Emerging associative classification algorithms have shown good promises on producing accurate classifiers. In this paper we focus on performance of associative classification and present a parallel model for classifier building. For classifier building some parallel-distributed algorithms have been proposed for decision tree induction but so far no such work has been reported for associative classification.*

**Keywords:** *classification, association, and data mining.*

## 1.Introduction

Data mining algorithms task is discovering knowledge from massive data sets. Building classifiers is one of the core tasks of data mining. Classification generally involves two phases, training and test. In the training phase the rule set is generated from the training data where each rule associates a pattern to a class. In the test phase the generated rule set is used to decide the class that a test data record belongs to. Traditionally, greedy search techniques such as decision trees [8] and others are used to develop classifiers. Decision Tree Induction approaches have been preferred to other traditional techniques due to the generation of small rule set and transparent classifiers. Transparent classifier means that rules are straightforward and simple to understand unlike some opaque classifiers such as one generated by neural networks where interpretation of rules is difficult. Greedy techniques in decision tree construction approaches tend to minimize overlapping between training data records to generate small rule sets. However small rule sets have some disadvantages. Greedy techniques may achieve global optimality if the problem has a optimal substructure. A novel technique associative classification based on association rule mining searches globally for all rules that satisfy minimum support and count thresholds [5].

## 2. Work already done in the field.

Several methods for improving the efficiency of all approach have been proposed [4,5,7,9,10] based on a recursive method for constructing a decision tree. In associative classification the classifier model is composed of a particular set of association rules, in which consequent of each rule is restricted to classification class attribute. The experiments in [4,5,7] show that this approach achieves higher accuracy than traditional approaches. Many sequential algorithms have been proposed for associative classification [4,5,7,9,10]. However associative classification suffers from efficiency due to the fact that it often generates a very large number of rules in association rule mining and it also takes efforts to select high quality rules from among them [7].

Since data mining is mostly applied on databases, which are very large, to improve the performance parallel algorithms are needed. Many parallel approaches have been given for association rule mining [11] and decision tree classifiers [3], no parallel algorithm has been proposed for associative classification. In this paper a model is proposed for parallel approach of associative classification for significant performance improvement. We propose a parallel model for CBA [5] algorithm.

## 3.Proposed methodology

*Parallel Approaches for Data Mining:*

Since data mining is frequently applied over large datasets, performance of algorithms is of concern. Exploiting the inherent parallelism of data mining algorithms provide a direct solution to their performance lift. A classification of different approaches to parallel processing for data mining is presented in [3].

Parallel Approaches:

1.Task Parallelism

1. Divide & Conquer
2. Task Queue

Task-parallel algorithms assign portions of the search space to separate processors. The task parallel approaches can again be divided in two groups. The first group is based on a Divide and Conquer strategy that divides the search space and assigns each partition to a specific processor.

2.Data Parallelism

1. Record Based
2. Attribute Based

The second group is based on a task queue that dynamically assigns small portions of the search space to a processor whenever it becomes available. Data-parallel, approaches distribute the data set over the available processors. Data-parallel approaches are in two directions. A partitioning based on records will assign non-overlapping sets of records to each of the processors. Alternatively a partitioning of attributes will assign sets of attributes to each of the processors. Attribute-based approaches are based on the observation that many algorithms can be expressed in terms of primitives that consider every attribute in turn. If attributes are distributed over multiple processors, these primitives may be executed in parallel. Many other issues on parallel processing of data mining with respect to Association Rule mining have been presented in [11].

The main challenges include synchronization and communication minimization, workload balancing, finding good data layout, data decomposition, and disk I/O minimization.

The parallel design space spans three main components:

1. the hardware platform,
2. the type of parallelism,
3. the load-balancing strategy.

Two dominant approaches for using multiple processors have emerged:

**Distributed memory (where each processor has a private memory)**

In distributed-memory (DMM) architecture, each processor has its own local memory and independent hard disk, which only that processor can access directly. For a processor to access data in the local memory of another processor, message passing must send a copy of the desired data elements from one processor to the other. A distributed memory, message-passing architecture cures the scalability problem by eliminating the bus, but at the expense of programming simplicity

**Shared memory (where all processors access common memory).**

Shared-memory (SMP) architecture has many desirable properties. Each processor has direct and equal access to all the system's memory. Parallel programs are easy to implement on such a system. A different approach to multiprocessing is to build a system from many units, each containing a processor and memory. Although shared memory architecture offers programming simplicity, a common bus's finite bandwidth can limit scalability. Load balancing strategies entail static or dynamic approaches. Static load balancing initially partitions work among the processors using a heuristic cost function, no subsequent data or computation movement is available. Dynamic load balancing seeks to address this by taking work from heavily loaded processors and reassigning it to lightly loaded ones. Computation movement also entails data movement, because the processor responsible for a computational task needs the data associated with that task. Dynamic load balancing thus incurs additional costs for work and data movement, and also for the mechanism used to detect whether there is an imbalance. However, dynamic load a balancing is essential if there is a large load imbalance or if the load changes with time. Dynamic load balancing is especially important in multi-user environments with transient loads and in heterogeneous platforms, which have different processor and network speeds. These kinds of environments include parallel servers and heterogeneous clusters, meta-clusters, and super-clusters (the so called grid platforms that are becoming common today).There are various approaches that can be applied to parallel processing of data mining algorithms. Cost measures for various parallel data mining strategies to predict their computation, data access and communication performance are presented in [6].

### *I. Associative Classification*

Associative Classification is an integrated framework of Association Rule Mining (ARM) and Classification. Focusing on a special subset of association rules whose right-hand-side is restricted to the classification class attribute does the integration. This subset of rules is referred as the Class Association Rules (CARs). CBA (Classification Based on Associations) [5] is a sequential approach of building associative classifier. CBA consists of two parts, a rule generator (called CBA-RG), which is based on algorithm Apriori for finding association rules in [2], and a classifier builder (called CBA-CB). CBA approach is described below. Assuming given dataset is a normal relational table, which consists of *N* cases described by *l* distinct attributes. These *N* cases have been classified into *q* known classes. For Associative Classification it is assumed that in training data set all continuous attributes (if any) have been discretized as a preprocessing step. For all attributes, all the possible values are mapped to a set of consecutive positive integers. With these mappings, a data case can be treated as a set of (*attribute*, *integer-value*) pairs and a class label. Each (*attribute*, *integer-value*) pair is called an *item.* Let *D* be the dataset. Let *I* be the set of all items in *D*, and *Y* be the set of class labels. We say that a data case $d \in D$ contains $X \subseteq I$, a subset of items, if $X \subseteq d$. A classification rule (CAR) is an implication of the form $X \rightarrow y$, where $X \subseteq I$, and $y \in Y$. A rule $X \rightarrow y$ holds in *D* with confidence *c* if *c*% of cases in *D* that contain *X* are labeled with class *y*. The rule $X \rightarrow y$ has support *s* in *D* if *s%* of the cases in *D* contain *X* and are labeled with class *y*.

### *Rule Generator CBA-RG*

The key operation of CBA-RG is to find all *ruleitems* that have support above *minsup*. A *ruleitem* is of the form: <*condset*, y>, where condset is a set of items, $y \in Y$ is a class label. The support count of the condset (called *condsupCount*) is the number of cases in *D* that contain the *condset*. The support count of the *ruleitem* (called *rulesupCount*) is the number of cases in *D* that contain the *condset* and are labeled with class *y*. Each *ruleitem* basically represents a rule: $condset \rightarrow y$, whose support is (*rulesupCount* / |*D*|) *100%, where |*D*|

is the size of the dataset, and whose confidence is(*rulesupCount* / *condsupCount*)*100%. *Ruleitems* that satisfy minsup are called *frequent ruleitems*, while the rest are called *infrequent ruleitems*. For example, the following is a ruleitem: <{(A, 1), (B, 1)}, (class,1)>, where A and B are attributes. If the support count of the *condset* {(A, 1), (B, 1)} is 3, the support count of the *ruleitem* is 2, and the total number of cases in *D* is 10, then the support of the *ruleitem* is 20%, and the confidence is 66.7%. If *minsup* is 10%, then the ruleitem satisfies the *minsup* criterion. We say it is frequent. For all the *ruleitems* that have the same *condset*, the *ruleitem* with the highest confidence is chosen as the possible rule (PR) representing this set of *ruleitems*. If there are more than one *ruleitem* with the same highest confidence, we randomly select one *ruleitem*. For example, we have two *ruleitems* that have the same *condset*:

1. <{(A, 1), (B, 1)}, (class: 1)>.
2. <{(A, 1), (B, 1)}, (class: 2)>.

Assume the support count of the *condset* is 3. The support count of the first *ruleitem* is 2, and the second *ruleitem* is 1. Then, the confidence of *ruleitem* 1 is 66.7%, while the confidence of *ruleitem* 2 is 33.3% With these two *ruleitems*, we only produce one PR (assume |*D*| = 10): (A, 1), (B, 1)_(class, 1) [*support* = 20%, *confidence*= 66.7%]. If the confidence is greater than *minconf*, we say the rule is *accurate*. The set of *class association rules* (CARs) thus consists of all the PRs that are both frequent and accurate.

The CBA-RG algorithm generates all the frequent *ruleitems* by making multiple passes over the data. In the first pass, it counts the support of individual *ruleitem* and determines whether it is frequent. In each subsequent pass,

it starts with the seed set of *ruleitems* found to be frequent in the previous pass. It uses this seed set to generate new possibly frequent *ruleitems*, called *candidate ruleitems*. The actual supports for these candidate *ruleitems* are calculated during the pass over the data. At the end of the pass, it determines which of the *candidate ruleitems* are actually frequent. From this set of frequent *ruleitems*, it produces the rules (CARs). Let *k-ruleitem* denote a *ruleitem* whose *condset* has *k* items. Each element $F_k$ of this set is of the following form: <(*condset*, *condsupCount*), (*y*, *rulesupCount*)>.

The CBA-RG algorithm is given in Figure 1.

```
$F_1$ = {large 1-ruleitems};
$CAR_1$ = genRules($F_1$);
for ($k$ = 2; $F_{k-1} \neq \phi$ ; $k$++) do
  $C_k$ = candidateGen($F_{k-1}$);
  for each data case $d \in D$ do
    $C_d$ = ruleSubset($C_k$, $d$);
    for each candidate $c \in C_d$ do
      c.condsupCount++;
      if d.class = c.class then c.rulesupCount++;
    end
  end
  $F_k$ = {$c \in C_k$ | c.rulesupCount >=minsup};
  $CAR_k$ = genRules($F_k$);
end
CARs = $\cup_k CAR_k$;
```

Figure 1

```
R = sort(R); // sort on precedence $\succ$
for each rule $r \in R$ in sequence do
   temp = $\phi$ ;
   for each case $d \in D$ do
      if d satisfies the conditions of r then
         store d.id in temp and mark r if it correctly
               classifies d;
   if r is marked then
      insert r at the end of C;
      delete all the cases with the ids in temp from D;
      selecting a default class for the current C;
      compute the total number of errors of C;
   end
end
Find the first rule p in C with the lowest total number
of errors and drop all the rules after p in C;
Add the default class associated with p to end of C;
return C (our classifier).
```

Figure 2. The CBA-RG : M1 algorithm

### *II.Building a Classifier CBA-CB*

To produce the best classifier out of the whole set of rules, a heuristic approach is used. A total order on the generated rules is defined. For more details on CBA approach readers are referred to [5]. This is used in selecting the rules for classifier.

Definition: Given two rules, $r_i$ and $r_j$, $r_i \succ r_j$ (also called $r_i$ precedes $r_j$ or *ri* has a higher precedence than $r_j$) if *1)*. the confidence of $r_i$ is greater than that of $r_i$, or *2)*. their confidences are the same, but the support of $r_i$ is greater than that of $r_j$, or 3. both the confidences and supports of $r_i$ and $r_j$ are the same, but $r_i$ is generated earlier than $r_j$ ;

Let *R* be the set of generated CARs and *D* the training data. The basic idea of the algorithm is to choose a set of high precedence rules in *R* to cover *D*.

Our classifier is of the following format: <$r_1,r_2,r_3,\ldots,r_n$, *default_ class* >, where $r_i \in R$, $r_a \succ r_b$ if $b > a$. *default_class* is the default class. In classifying an unseen case, the first rule that satisfies the case will classify it. If there is no rule that applies to the case, it takes on the default class as in C4.5. A pseudo code of algorithm M1 for building such a classifier is shown in Figure 2.

***III.Parallel and Distributed Associative Classification***

CBA is an associative classification algorithm that uses an Apriori based approach to mine CARs and produces a subset of these CARs after pruning to form a classifier. We adapt here popular CBA algorithm discussed in section 3 to present our approach of parallel associative classification.For both phases of associative classification, rule generation phase and classifier builder phase our approach is based on Distributed Memory Systems, Record Based data parallelism and uses Static Load Balancing. The above configuration of parallel approach suits to the inherent parallel nature of existing serial approach. Three parallel versions of Apriori are given in [1] on shared nothing architecture. We adapt count distribution algorithm of ARM mining for mining of CARs in associative classification and present parallel version of CBA-M1 for classifier building. Count distribution approach has minimized communication among the processors. For CARs mining the training data set is partitioned among *P* processors. Each processor works on its local partition of the database and performs same set of instructions to mine CARs that have global min support and confidence. Later when all CARs are found, same partitions of training set are used in respective nodes and pruning process based on coverage analysis is applied in parallel to generate reduced set of CARs, to form classifier.Our approach simply achieves load balancing if training data sets are sufficiently randomly distributed over different processors to avoid any data skew.

| | |
|---|---|
| $C_k$ | Set of candidate *k-ruleitems* |
| $F_k$ | Set of frequent *k-ruleitems* |
| $D$ | Training Dataset |
| $D_i$ | Local Training Dataset on $i^{th}$ Processor |
| $P_i$ | $i^{th}$ Processor |
| $R$ | Set of generated *CARs* |

*Figure 3. Notations*

It can simply be inferred
$|D| = N$ and number of processors = $p$
$|D_i| = |D|/p$ (approximately), $i=1,2,\ldots p$

The necessary communication among processors is through message broadcasting. There has been no need of dynamic load balancing as the Associative Classifier builder task does not involve multi-user environment with changing training data sets during learning. As in ARM in associative classification too the static load balancing is inherent in the partitioning of the database among processors because training data sets have been made available in a homogeneous environment. Parallel versions adapted from CBA for CAR generation and classifier builder are presented below.

Pseudo codes given in Figure 4 use notations given in Figure 3.

```
do in parallel
  k = 2;
  while (Fk-1 ≠ ϕ ) do
    for each processor Pi (i=1..p) do
      Ck = candidateGen(Fk-1);
      for each data case d∈ D do
        Cd = ruleSubset(Ck, d);
        // compute local support for ruleitems
        for each candidate c∈ Cd do
          c.condsupCount++;
          if d.class = c.class then c.rulesupCount++
        end
      end
      exchange local Ck counts with all other processors
                                        and synchoronize;
      for each c∈ Ck  c.condsupCount=∑Pi(i=1..p)c.condsupCount;
        c.rulesupCount =∑ Pi (i=1..p)c.rulesupCount;
      end
      Fk = {c∈ Ck | c.rulesupCount >=minsup};
      CARk = genRules(Fk);
      k++;
  end while
  CARs =∪ CARk
End parallel
```

*Figure 4*

*Parallel CAR generation phase*

At the time of generating partitions $D_i$ for processors $P_{i,}$ $F_1$ and hence $CAR_1$ can be generated and distributed to distributed processors along with data partitions. Hence algorithm begins with a seed of F1. During partitioning *class_distribution* i.e. number of data cases for each of the classes will also be computed and distributed to processors to be used during class builder phase. Pseudo code of parallel rule generation algorithm is presented in figure 4.

In Parallel CAR generation algorithm each processor independently and in parallel generates identical $C_k$ for $k > 1$ and calculates local counts and broadcast these to all other processors. At this step processors synchronize and wait for all other processors to compute and broadcast their local counts. Summing local counts of all processors global counts for $C_k$ are computed. Each processor now computes $F_k$ and generates CARs from $F_k$. Process is iterated for next *k* till until $F_k$ is empty. At the end of rule generation algorithm each processor has complete set of CARs.

## 4.Cost Measures For Proposed Model

All this information exchanged is integer valued and its volume is very small.

Cost estimate of sequential CBA approach based on cost models in [6] is as follows. Global structure of both rule generation and class builder algorithms is a loop building more accurate concepts from those of the previous iterations. Suppose loop in rule generation algorithm and classifier builder algorithms executes $k_{s1}$ ,$k_{s2}$ times and builds $\Omega_1$, ,$\Omega_2$ concepts respectively. Total size *D* is *N* and number of attributes in *D* is *l.* The cost estimate of the sequential CBA algorithm can be given by formula

$$Cost_{seq} = k_{s1} [\ STEP(N*l, \Omega_1\ ) + ACCESS\ (N*l)\ ]\ \ + k_{s2} [\ STEP(N*l, \Omega_2\ ) + ACCESS\ (N*l)\ ]$$

Where *STEP* gives the cost of single iteration of the loop, and *ACCESS* is the cost of accessing the data set once.

If the CBA is performed in parallel version of rule generation algorithm and classifier builder algorithms and requires $k_{a1}$ ,$k_{a2}$ iterations respectively with number of *p* processors, formula for the cost can be given by

$$Cost_{par} = k_{a1} [STEP(N*l/p, \Omega_1\ )+ACCESS\ (N*l/p) + C_{e1}] + k_{a2} [\ STEP(N*l/p, \Omega_2\ )+ACCESS\ (N*l/p)+C_{e2}]$$

$C_{e1}$, $C_{e2}$ are total cost of communication and information exchange between the processors.

It can be reasonably assumed that

$STEP(N*l/p, \Omega_1\ ) = STEP(N*l, \Omega_1\ )/p$

and

$ACCESS\ (N*l/p) = ACCESS\ (N*l)/p$

So, we get significant *p*-fold speedup in executing parallel version except cost of overheads. In our model overhead cost is small as information exchanged is integer valued and its volume is very small.

## 5.Conclusion

In this paper the focus was on the performance of classifier builder approach known as associative classification. We proposed a model to show that associative classification task can be performed in parallel on distributed memory systems to achieve a significant performance lift. We have presented parallel versions for both ruled generation and class builder phase of sequential CBA algorithm for load balancing we have distributed almost equal number of data sets randomly on each of the local processors to avoid data skew ness.

## 6.References

1. R. Agrawal and J. C. Shafer, "Parallel Mining of Association Rules", *IEEE Transactions On Knowledge And Data Engineering*, pages 962–969, 1996.
2. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In *Proc. of the Int. Conf. on Very Large Databases*, SanDiago, Chile, pages 487–499, 1994.
3. J. Chattratichat, "Large Scale Data Mining: Challenges and Responses," *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, 1997, pp 143-146.
4. W. Li, J. Han and J. Pei, "CMAR: Efficient classification based on multiple class-association rules. In *Proc. of the Int. Conf. on Data Mining*, pages 369–376, 2001.

5. B. Liu, W. Hsu and Y. Ma, "Integrating classification and association rule mining", In *Knowledge Discovery and Data Mining*, pages 80–86, 1998.
6. D. B. Skillicorn, "Strategies for parallel data mining", *IEEE Concurrency*, vol. 7, No. 4,1999.
7. X. Yin and J. Han, "CPAR: Classification based on Predictive Association Rules", In *Proc. of the Int. Conf. on Data Mining*, SDM. SIAM, 2003.
8. F. Thabtah, P. Cowling and Y. Peng, "MCAR: Multi-class Classification based on Association Rule Approach. In *Proceeding of third IEEE International Conference on Computer Systems and Applications*, Cairo, Egypt, pages 1-7, 2005.
9. F. Thabtah, P. Cowling and Y. Peng, "MMAC: A New Multi-class, Multi-label Associative Classification Approach", In *Proceeding of fourth IEEE International Conference on Data Mining* (ICDM '04), Brighton, UK, pages 217-224, Nov. 2004.

## Author's Information

Thakur S. Ghanshyam, Dr. R.C.Jain – Department of Computer Application; Samrat Ashok Technological Institute; Vidisha(M.P.),INDIA; e-mail: ghanshyamthakur@gmail.com

# USING THE AGGLOMERATIVE METHOD OF HIERARCHICAL CLUSTERING AS A DATA MINING TOOL IN CAPITAL MARKET[1]

Vera Marinova–Boncheva

***Abstract:*** *The purpose of this paper is to explain the notion of clustering and a concrete clustering method- agglomerative hierarchical clustering algorithm. It shows how a data mining method like clustering can be applied to the analysis of stocks, traded on the Bulgarian Stock Exchange in order to identify similar temporal behavior of the traded stocks. This problem is solved with the aid of a data mining tool that is called XLMiner™ for Microsoft Excel Office.*

***Keywords***: *Data Mining, Knowledge Discovery, Agglomerative Hierarchical Clustering.*

***ACM Classification Keywords***: *I.5.3 Clustering*

## Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally are time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining consists of analysis of sets of supervised data with the aim of finding unexpected dependencies or to be generalized in a new way that is understandable and useful for owners of the data. There is a great deal of data mining techniques but we differentiate two of them like classification and clustering as supervised and unsupervised learning from data. [2]

[1] This work was supported by the Ministry of Education and Science under Project № MU-MI-1601/2006

## The Analysis of Clustering

Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Cluster Analysis, also called data segmentation, has a variety of goals. They all relate to grouping or segmenting a collection of objects (also called observations, individuals, cases, or data rows) into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered that depends on the data and the application. Different types of similarity measures may be used to identify classes (clusters), where the similarity measure controls how the clusters are formed. Some examples of values that can be used as similarity measures include distance, connectivity, and intensity. [4]

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- constrained - based clustering;
- interpretability and usability. [7]

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value. A hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering uses a completely probabilistic approach. [5, 6]

There are a number of problems with clustering. Among them:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of “distance” (for distance-based clustering);
- if an obvious distance measure doesn’t exist we must “define” it, which is not always easy, especially in multi-dimensional spaces;
- the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

Clustering is a method that is applicable in many fields like:

- Marketing: finding groups of customers with similar behavior when it is given a large database of customer data containing their properties and past buying records;
- Biology: classification of plants and animals given their features;
- Libraries: book ordering;
- Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- City-planning: identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: clustering observed earthquake epicenters to identify dangerous zones;
- WWW: document classification; clustering weblog data to discover groups of similar access patterns.

## Hierarchical Clustering

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to N clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the N objects into groups, and divisive methods, which separate N objects successively into finer groupings. Agglomerative techniques are more commonly used, and this is the method implemented in the free version of XLMiner™ which is the Microsoft Office Excel add-in. [1]

If it is given a set of N items to be clustered and a N*N distance (or similarity) matrix then the basic process of agglomerative hierarchical clustering can be done iteratively following these four steps:

1. Start by assigning each item to a cluster. Let the distances (similarities) between the clusters are the same as the distances (similarities) between the items they contain;
2. Find the closest (most similar) pair of clusters and merge them into a single cluster;
3. Compute distances (similarities) between the new cluster and each of the old clusters;
4. Repeat step 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be different because of the varieties in the definition of the distance (or similarity) between clusters:

- Single linkage clustering (nearest neighbor technique) – here the distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group is considered i.e. the distance between two clusters is given by the value of the shortest link between clusters. At each stage the two clusters for which the distance is minimum are merged;
- Complete linkage clustering (farthest neighbor) – is the opposite of the single linkage i.e. distance between groups is defined as the distance between the most distant pair of objects, one from each group. At each stage the two clusters for which the distance is minimum are merged;
- Average linkage clustering – the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group. At each stage the two clusters for which the distance is minimum are merged;
- Average group linkage clustering – with this method, groups once formed are represented by their mean values for each variable, that is their mean vector and inter-group distance is defined in terms of distance between two such mean vectors. At each stage the two clusters for which the distance is minimum are merged. In this case, those two clusters are merged such that the newly formed cluster, on average, will have minimum pairwise distances between the points in it;
- Ward's hierarchical clustering – Ward (1963) proposed a clustering procedure seeking to form the partitions $P_n,...,P_1$ in a manner that minimizes the loss associated with each grouping and to quantify that loss in a form that is readily interpretable. At each step the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in "information loss" are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion. [3]

Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. By cutting the dendrogram at a desired level clustering of the data items into disjoint groups is obtained. [1]

Major weakness of agglomerative clustering methods is that:

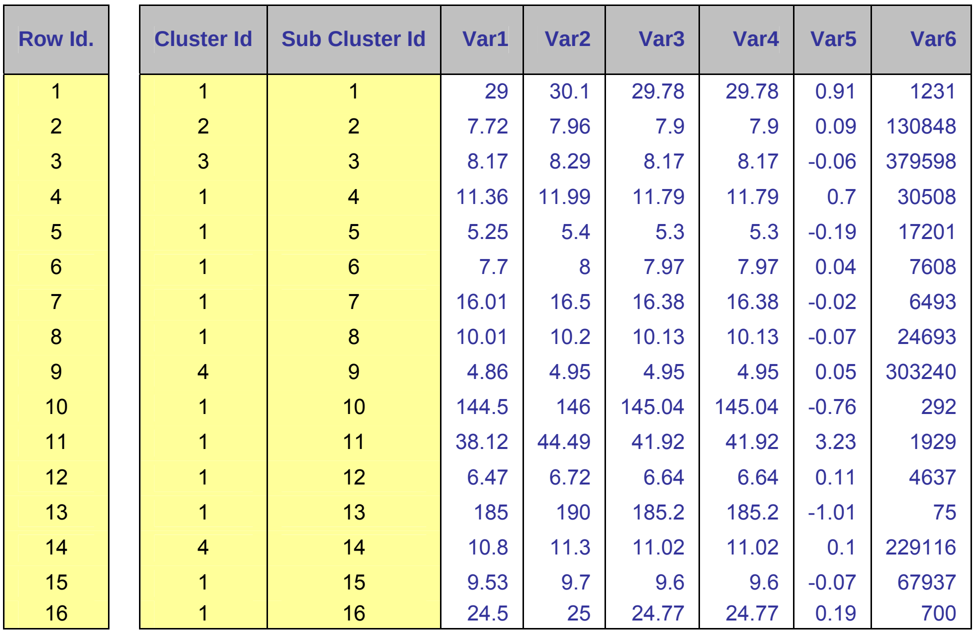
- they do not scale well and time complexity is at least $O(n^2)$, where $n$ n is the number of total objects;
- they can never undo what was done previously.

## Clustering of Stocks, traded on the Official Market of BSE

As inputs we have taken data for 16 stocks from the Bulgarian Stock Exchange in a single day. (Table 1) These data are listed on the Internet address: http://www.econ.bg/capital.html. It contains information for each stock as the code and the name of the company, the nominal, prices (low, high, last, medium), the change in price in comparison to the previous day and the traded amount of this kind of stock.

| **company code** | **nominal** | **prices** | | | | **change** | **amount** |
|---|---|---|---|---|---|---|---|
| | | low | high | last | medium | | |
| CENHL | 1 | 29 | 30.1 | 29.78 | 29.78 | 0.91 | 1231 |
| SFARM | 1 | 7.72 | 7.96 | 7.9 | 7.9 | 0.09 | 130848 |
| CCB | 1 | 8.17 | 8.29 | 8.17 | 8.17 | -0.06 | 379598 |
| PETHL | 1 | 11.36 | 11.99 | 11.79 | 11.79 | 0.7 | 30508 |
| DOVUHL | 1 | 5.25 | 5.4 | 5.3 | 5.3 | -0.19 | 17201 |
| IHLBL | 1 | 7.7 | 8 | 7.97 | 7.97 | 0.04 | 7608 |
| ALBHL | 1 | 16.01 | 16.5 | 16.38 | 16.38 | -0.02 | 6493 |
| GAZ | 1 | 10.01 | 10.2 | 10.13 | 10.13 | -0.07 | 24693 |
| PET | 1 | 4.86 | 4.95 | 4.95 | 4.95 | 0.05 | 303240 |
| ORGH | 1 | 144.5 | 146 | 145.04 | 145.04 | -0.76 | 292 |
| HVAR | 1 | 38.12 | 44.49 | 41.92 | 41.92 | 3.23 | 1929 |
| SEVTO | 1 | 6.47 | 6.72 | 6.64 | 6.64 | 0.11 | 4637 |
| ODES | 1 | 185 | 190 | 185.2 | 185.2 | -1.01 | 75 |
| CHIM | 1 | 10.8 | 11.3 | 11.02 | 11.02 | 0.1 | 229116 |
| MONBAT | 1 | 9.53 | 9.7 | 9.6 | 9.6 | -0.07 | 67937 |
| KTEX | 1 | 24.5 | 25 | 24.77 | 24.77 | 0.19 | 700 |

Table 1. Information about stocks, traded on the Official Market of Bulgarian Stock Exchange



| Row Id. | Cluster Id | Sub Cluster Id | Var1 | Var2 | Var3 | Var4 | Var5 | Var6 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 29 | 30.1 | 29.78 | 29.78 | 0.91 | 1231 |
| 2 | 2 | 2 | 7.72 | 7.96 | 7.9 | 7.9 | 0.09 | 130848 |
| 3 | 3 | 3 | 8.17 | 8.29 | 8.17 | 8.17 | -0.06 | 379598 |
| 4 | 1 | 4 | 11.36 | 11.99 | 11.79 | 11.79 | 0.7 | 30508 |
| 5 | 1 | 5 | 5.25 | 5.4 | 5.3 | 5.3 | -0.19 | 17201 |
| 6 | 1 | 6 | 7.7 | 8 | 7.97 | 7.97 | 0.04 | 7608 |
| 7 | 1 | 7 | 16.01 | 16.5 | 16.38 | 16.38 | -0.02 | 6493 |
| 8 | 1 | 8 | 10.01 | 10.2 | 10.13 | 10.13 | -0.07 | 24693 |
| 9 | 4 | 9 | 4.86 | 4.95 | 4.95 | 4.95 | 0.05 | 303240 |
| 10 | 1 | 10 | 144.5 | 146 | 145.04 | 145.04 | -0.76 | 292 |
| 11 | 1 | 11 | 38.12 | 44.49 | 41.92 | 41.92 | 3.23 | 1929 |
| 12 | 1 | 12 | 6.47 | 6.72 | 6.64 | 6.64 | 0.11 | 4637 |
| 13 | 1 | 13 | 185 | 190 | 185.2 | 185.2 | -1.01 | 75 |
| 14 | 4 | 14 | 10.8 | 11.3 | 11.02 | 11.02 | 0.1 | 229116 |
| 15 | 1 | 15 | 9.53 | 9.7 | 9.6 | 9.6 | -0.07 | 67937 |
| 16 | 1 | 16 | 24.5 | 25 | 24.77 | 24.77 | 0.19 | 700 |

Table 2. Clusters of stocks taken from table 1

We use the data mining tool named XLMiner™ for MS Excel. We select the agglomerative method of hierarchical clustering to find clusters of stocks. We experiment on all five variants of agglomerative method of hierarchical clustering and we have founded that the average linkage method will give the best results. We use as a stop rule for the process of clustering the number of clusters which is 4. [1]

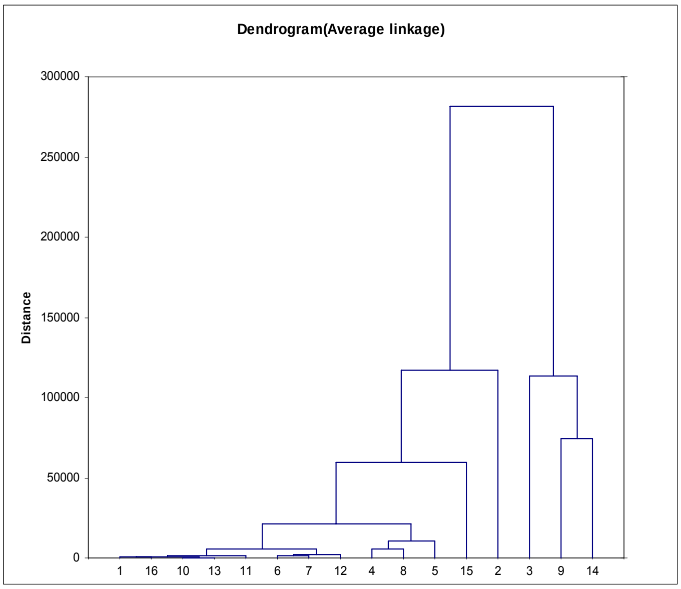


Figure 1. Dendrogram of the clusters from table 1

The dendrogram in Figure 1 shows how the numbered stocks are divided into the following four clusters: {1,4,5,6,7,8,10,11,12,13,15,16}, {2}, {3}, {9,14}. (Table 2) The last cluster is composed by two stocks that have the least prices, the greatest amounts traded and positive change. They are the most interesting for the investor. The second and the third cluster consist of only one stock. They have approximately equal prices and high amounts of them are traded but they differ from each other because stock 2 has positive change but stock 3 has negative change. The rest of stocks are grouped in another cluster. So this method is a good way to combine stocks that are preferred by the investors.

## Conclusion

Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning. Like statistics, data mining is not a business solution, it is just a technology. In this article it has been shown how a hierarchical clustering method can support an investor decision to choose stocks which can pretend to be participants in an investment portfolio by using a data mining tool. So the identification of clusters of companies of a given stock market can be exploited in the portfolio optimization strategies.

## Bibliography

1. G, Nitin, R. Patel, P. C. Bruce. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Hardcover, 2007.
2. Chris Westphal, Teresa Blaxton, Data Mining Solutions, John Wiley, 1998.
3. A.K.Jain, R.C. Dubes. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall, 1988.
4. L. Kaufman, P.J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley&Sons, 1990.
5. J.A. Harigan. Clustering Algorithms. New York: John Wiley&Sons,1975.
6. A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A survey. ACM Comput. Surv., 31:264-323, 1999.
7. J. Han, M. Kamber. Data mining: Concepts and Techniques, Morgan Kaufmann, 2000.

## Authors' Information

***Vera Marinova-Boncheva -*** *Institute of Information Technologies, Bulgarian Academy of Science, Sofia-1113, Bulgaria; e-mail: vboncheva@iit.bas.bg*

Transition P Systems

# A CIRCUIT IMPLEMENTING MASSIVE PARALLELISM IN TRANSITION P SYSTEMS

**Santiago Alonso, Luis Fernández, Fernando Arroyo, Javier Gil**

***Abstract**: Transition P-systems are based on biological membranes and try to emulate cell behavior and its evolution due to the presence of chemical elements. These systems perform computation through transition between two consecutive configurations, which consist in a m-tuple of multisets present at any moment in the existing m regions of the system. Transition between two configurations is performed by using evolution rules also present in each region.*

*Among main Transition P-systems characteristics are massive parallelism and non determinism. This work is part of a very large project and tries to determine one way to design a hardware circuit that can improve remarkably the process involved in the evolution of a membrane. Process in biological cells has two different levels of parallelism: the first one, obviously, is the evolution of each cell inside the whole set, and the second one is the application of the rules inside one membrane. This paper presents an evolution of the work done previously and includes an improvement that causes that the transition between two states is reached using massive parallelism. To achieve this, the initial set of rules is transformed into a new set that consists in all their possible combinations, and each of them is treated like a new rule (participant antecedents are added to generate a new multiset), converting an unique rule application in a way of parallelism in the means that several rules are applied at the same time. In this paper, we present a circuit that is able to process this kind of rules and to decode the result, taking advantage of all the potential that hardware has to implement P Systems versus previously proposed sequential solutions.*

***Keywords**: Transition P System, membrane computing, circuit design.*

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

## Introduction

Transition P-systems or Membrane Computing (designed by [Păun, 1998]) are based on the processes that occur among living cells. The idea behind it is the fact that a living cell may change its state depending on the set of elements that are present in it and, of course, depending on the chemical rules that can transform them. So we can create a computational model based on that behavior. So, there is a definition of a cellular structure that contains elements that can be repeated, conforming multisets, and rules that define how multisets are combined to do the cell evolution. One of these structures (membranes) may contain another ones, conforming a hierarchical relation whose components may communicate among them, always based on what the rules allow. Evolution due to a rule application may cause that a membrane passes information to the one immediately superior in the hierarchy or to any of the ones that are immediately inferior. All this, besides the fact that eventually, a membrane may be inhibited or dissolved by means of some rule application, and that they may have different priorities, does the P-systems very interesting in order to define their hardware implementation.

All these processes can be viewed as computational ones and P systems have been sufficiently characterized from a theoretical point of view and their computational power has been settled. However, nowadays, the way in which these models have to be implemented is still a problem not solved. This problem is having two different approaches: software and hardware models. There are many papers about software tools implementing different P system variants [Gutierrez-Naranjo, 2006], but in the case of P-systems hardware implementation, only a few references can be found: connectivity arrays for membrane processors [Arroyo, 2004], multisets and evolution rules representation in membrane processors [Arroyo, 2004b] or a hardware membrane system description using VHDL [Petreska, 2003]. However, in [Martinez, 2006a] and [Martinez, 2006b] there is a hardware approach that implements a circuit that covers the whole process that takes place inside a membrane. Authors describe the way a sequential circuit may control the application of active rules in a Transition P –system and its internal structure.

Being aware that P-systems are defined as "distributed, massively parallel and non deterministic", we think these characteristics should be strengthen. Parallelism takes place in this model in two different levels: the first one is due to the fact that every cell or membrane evolutions at the same time than the others, and the second one is due to the fact that rules inside each membrane may be applied at the same time.

It is at this point where this work pretends to be positioned: parallelism by means of application of multiple rule at the same time.

The structure of this paper presents, first, the problem and its methodological solution and afterwards, shows its data model and a general representation of the circuit, as well as each part in detail.

## The algorithm

As we may read in [Martinez, 2006a] and [Martinez, 2006b], a hardware approach to P-system is possible. These papers show how the general algorithm of an evolution system may be developed with a circuit. Authors clearly improved the basic algorithm by the way of the proposal of obtaining the number that represents the maximum times each rule could be applied to the current multiset. This number, called *applicability MAX*, is the higher limit for a random number that indicates how many times the rule will be applied, modifying the basic algorithm as:

Let R be the initial set of active rules, R = {$R_1$, $R_2$, …, $R_n$ } and W the initial multiset, being *input($R_i$)* the antecedents for rule $R_i$

```
1.   R ← InitialActiveRules
2.   REPEAT
3.   $R_i$ ← Aleatory (R)
4.   MAX ← Applicability ($R_i$, W)
5.   IF MAX = 0
6.   THEN  R ← R - {$R_i$}
7.   ELSE
8.   K ← Aleatory(1, MAX)
9.   W ← W – K * input($R_i$)
10.  count(K, $R_i$)
11.  UNTIL |R| = 0
```

As we can see, the algorithm works by selecting randomly one rule until there are no rules to apply (|R| = 0). Once the rule is selected, it calculates its *MAX* value; if this value is zero, it means that the rule is no more applicable and it has to be removed from the set of rules.

Afterwards, it generates a random number *K*, equal or less than *MAX* and the application of the rule consists in subtracting *K* times the antecedents input($R_i$) from W. This means that such rule is being *K* times used. Of course we have to store this value so we can check how many times a rule has been applied (step 10).

So, this algorithm is implementing some way of parallelism (in each iteration, a rule is applied K times). However, the importance of parallelism in this kind of model, as well as its possible importance in the field of NP problem solving, urged us to find a way to be able to apply several rules at the same time, improving its throughput (after all, the exposed algorithm just calculated *MAX* for one rule). Thus, the idea is to find a way to select several rules and apply them over the multiset in each evolution step. We could see that this could be achieved in a better way, improving its computational throughput just by considering the initial set of rules as a new set composed by the rules that result form calculating the power set P(R) from the original set of rules. So if we have:

$R = \{R_1, R_2, \dots, R_n\}$

its power set is:

$P(R) = \{\varnothing, R_1, R_2, \dots, R_n, R_1 R_2, \dots, R_1 R_n,, \dots, R_{n-1} R_n, \dots, R_1 R_2 \dots R_{n-1} R_n\}$

*As* $\{\varnothing\}$ is an element with no rules and it has no meaning for this work, the power set minus the empty set will be considered:

$P'(R) = P(R) - \{\varnothing\} = \{ R_1, R_2, \dots, R_n, R_1 R_2, \dots, R_1 R_n,, \dots, R_{n-1} R_n, \dots, R_1 R_2 \dots R_{n-1} R_n\}$

If we consider now this set P'(R) as the initial active rules set, what we are doing is to be able to apply several rules at the same time, by the meaning that if a rule $R' \in P'(R)$ / $R' = R_{x\dots}R_yR_z$ is chosen, a possible evolution may process the antecedents of several rules ($R_{x\dots}R_yR_z$ ) at the same time (as many as conform the chosen element). The algorithm, right now would be:

Let R be the initial set of active rules, $R = \{R_1, R_2, \dots, R_n \}$ and W the initial multiset, being *input*($R_i$) the antecedents for rule $R_i$

*Let P(R) be the power set of R and P'(R) = P(R) - {∅} with card( P(R)) =* $2^n$ *and card(P'(R) =* $2^n$ *-1) )*

```
1.  REPEAT
2.     $\forall R_i \in P'(R)$, ||      $MAX_i \leftarrow$ Applicability ($R_i$, W)
3.     $\forall R_i \in P'(R)$, ||      $K_i \leftarrow$ Aleatory(1, $MAX_i$)
4.     COBEGIN
5.        $\forall R_i \in P'(R)$, ||   $W_T \leftarrow K_i$ * input($R_i$)
6.        END $\leftarrow \neg \exists K_i$ <>0; IF NOT END
7.                     THEN BEGIN
8.                        $R_j \leftarrow$ Aleatory (P'(R)) / $K_i$ <>0
9.                        COBEGIN
10.                          W $\leftarrow$ W – $W_T$
11.                          count ($K_i$, $R_j$, R)
12.                       COEND
13.                    END
14.    COEND
15. UNTIL END
```

As we may see, this algorithm underlines the importance of parallelism, taking advantage in the processes that can be done simultaneously. As we will see ahead, there are two types of parallelism: first, some processes are applied to all the rules at the same time (indicated by the sign "||" in steps 2, 3 and 5) and second, some control processes may be done simultaneously (indicated by the clauses "COBEGIN … COEND").

Moreover, differences with the previous algorithm include (steps 2 and 3) calculating applicability *MAX* and a random number ($K_i$, between 1 and its *MAX* value) for each of the rules that are included in P'(R). As they should be calculated simultaneously, process time is not incremented. Once this is done, it calculates the product of each $K_i$ by the antecedents of each rule, but, at the same time this is happening, there is a special process (steps 6 through 8) that selects a random rule but just for the rules whose *MAX* value is different than zero (this means that $K_i$ is also different than zero). This causes that any selected rule is applicable and only in the case that no rule has *MAX* value greater than zero, the *END* condition is reached.

Once the rule is selected, of course the system has to subtract the antecedents ($W_T$) from the set of elements (W) but, again at the same time, it has to decode the participant rules, because not all the rules that are in P'(R) appear also in R. A rule could be the result from the composition of several rules from R and so, the process has to increase the counter for each of the rules from R.

## The model and data representation

Before we can start with the circuit design, there is the need for a definition of a data structure that contains information about the initial membrane state, the initial multiset of objects and the set of evolution rules. Continuing with the work done in precedent papers, and knowing that we have to establish some limits for a suitable circuit, the model should:

a. Limit the cardinality O = {a, b, c, d, e, f, g, h, i, j} of the alphabet to 10.

b. Define the initial multiset involved in a specific membrane *i*, $W_i$, is represented in a 4-bits register. The length of this register will be 10. The value in each register position will represent the number of occurrences for the object represented by the alphabet letter in that position.

c. The finite set of evolution rules R associated to the membrane *i* is represented by a set of registers, each of ones represents the antecedents of rule *i*, and the value in each position represents the element occurrences needed for the current rule to be applied.

d. The Application Rules Register is represented by a register which length is, at least, $\log_2 n$, being n = card(P(R))

In this work we have to consider two main aspects:

a. First, the initial set of rules is considered to be the power set of active rules at the beginning of the process. The circuit to obtain active rules from the initial multiset may be obtained from [Martinez, 2006a].

b. Shown solution will be scalable, so, increasing number of initial rules will not have a negative influence in the design (if card(R)=n, then card( P(R) ) = $2^n$ and card(P'(R) = $2^n$ -1) ). In this paper we will work with examples with a set of three initial rules, that makes card (P'(R)) = 7.

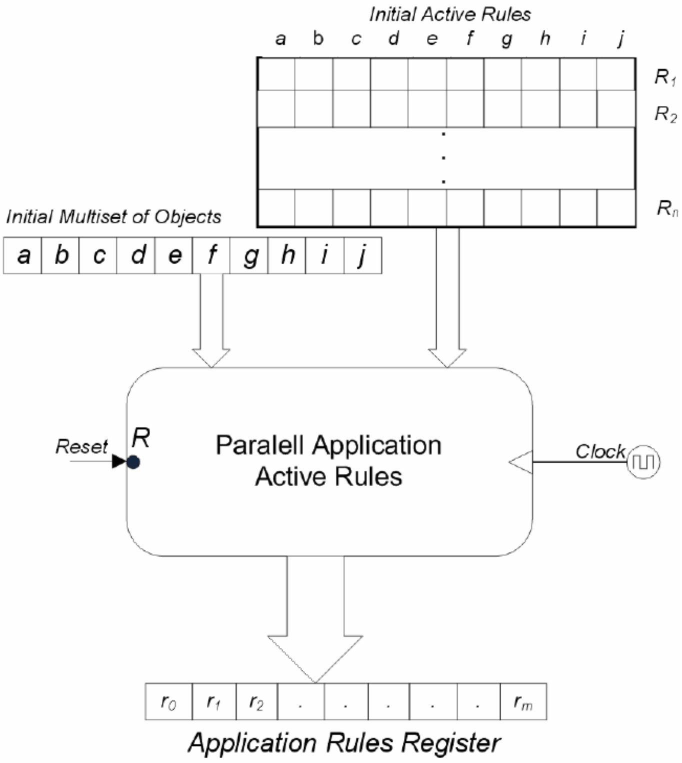

Initial Active Rules
a b c d e f g h i j
$R_1$
$R_2$
$R_n$
Initial Multiset of Objects
a b c d e f g h i j
Reset
R
Paralell Application Active Rules
Clock
$r_0$ $r_1$ $r_2$ . . . . . $r_m$
Application Rules Register

Figure 1. Circuit inputs and outputs

The circuit shown in figure 1 takes the set of rules, already P'(R) members (*Initial Active Rules*), and the initial multiset of objects and brings out a complete register (*Application Rules Register*) with the occurrences each rule should be applied to obtain a step of evolution.

## The circuit

The circuit is the result for assembling different functional units created each one to do a specific job. All of them should be coordinated by a "Logic Control Unit" not represented in figure 2, that takes the control and repeats the whole cycle until the signal provided by the Application Selector F.U. indicates that are no more active rules.

The different units are:

**Applicability MAX F.U.:** This functional unit is the one that receives an active rule and determines its *Aplicability Max* value, as explained before. This value is calculated as the largest number of times current rule can be applied without having in mind the other rules. So, this functional unit needs, as input, the antecedents of current rule and the multiset of objects. The output will be the MAX value for current rule.

The Max value may be obtained [Martinez, 2006a] by dividing each position value from the register for antecedents by its corresponding position value in the multiset register. Once obtained all this results, the smallest one will be the maximum value the rule may be applied.

**Random generator 1..MAX:** once the Applicability Max is obtained, the circuit should generate randomly a value for each of the available rules in the active rules register. This value represents the number that each rule should be applied in case that specific rule is chosen to be the one that consumes the elements and its lowest value will be 1 and the highest will be $Max_{i.}$

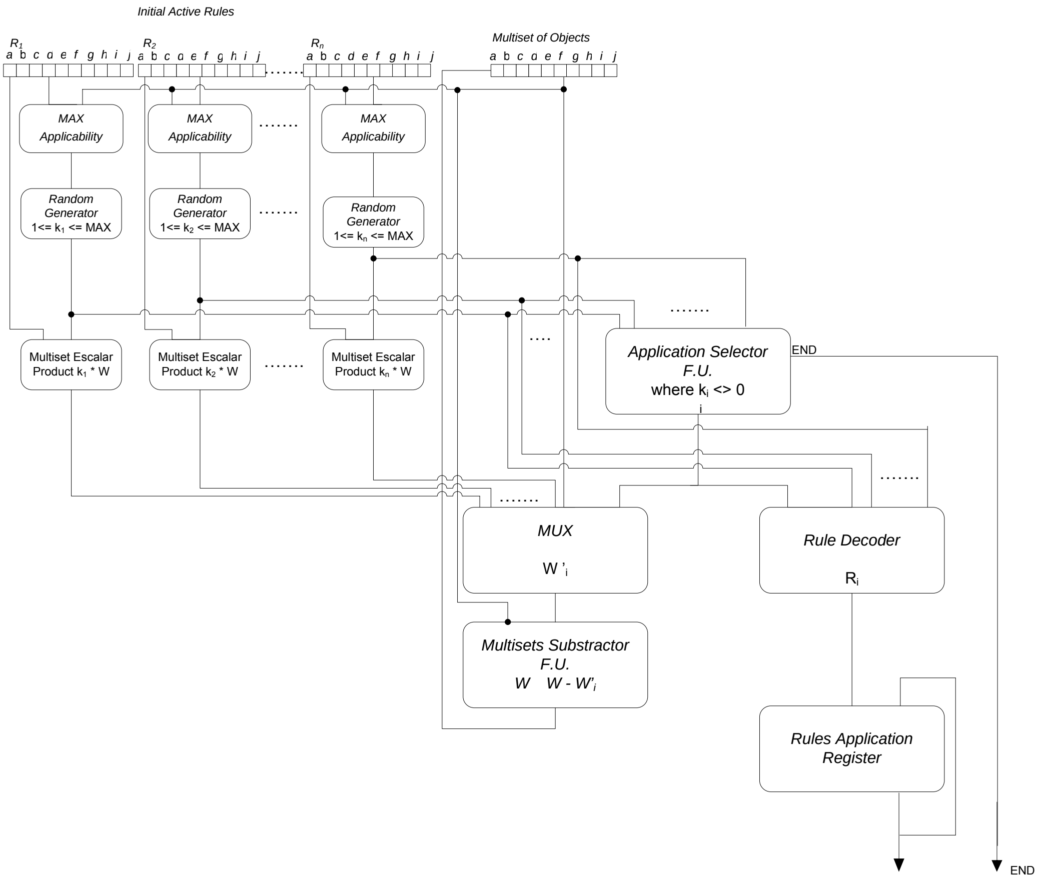

Initial Active Rules
R1
a b c d e f g h i j
R2
a b c d e f g h i j
Rn
a b c d e f g h i j
Multiset of Objects
a b c d e f g h i j
MAX Applicability
MAX Applicability
MAX Applicability
Random Generator 1<= k1 <= MAX
Random Generator 1<= k2 <= MAX
Random Generator 1<= kn <= MAX
Multiset Escalar Product k1 * W
Multiset Escalar Product k2 * W
Multiset Escalar Product kn * W
Application Selector F.U. where ki <> 0
i
END
MUX
W 'i
Rule Decoder
Ri
Multisets Substractor F.U.
W W - W'i
Rules Application Register
END

Figure 2: Circuit functional units

It is very important to realize that Max value could be zero, due to the fact that a rule could not be applied because there are no enough elements in the multiset. In another type of circuit, this would cause the rule to be invalid for the process and that could be a problem. In this case, this is solved by the Application Selector F.U. where a k not equal to zero is calculated.

If n is the cardinality of P'(R), all this calculation (n times a random number between 1 and $Max_i$) may be done at the same time, forcing the higher parallelism.

**Application Selector F.U.:** Once the previous random generator has calculated $k_i$ for each rule, we can find that any of these numbers may be zero (due to the fact that its *Application Max* value may be equal to zero). We have to implement a way of avoiding to choose a rule with $k_i$ equal to zero because it would cause a delay time in process dedicated, probably, to recalculate a new $k_i$. To avoid this kind of problems, we developed a functional unit that can generate a random number but just for those rules which k is greater than zero.

Achieving the developing of this functional unit included developing of one special cell that obtains the position of the first "1" appearing in the register, and another cell to get the position of the second "1", and another for the third, and so on. We will have as many cells as number of rules in P'(R). As result of this, we will get together all the positions that have a value for $k_i$ different from zero.

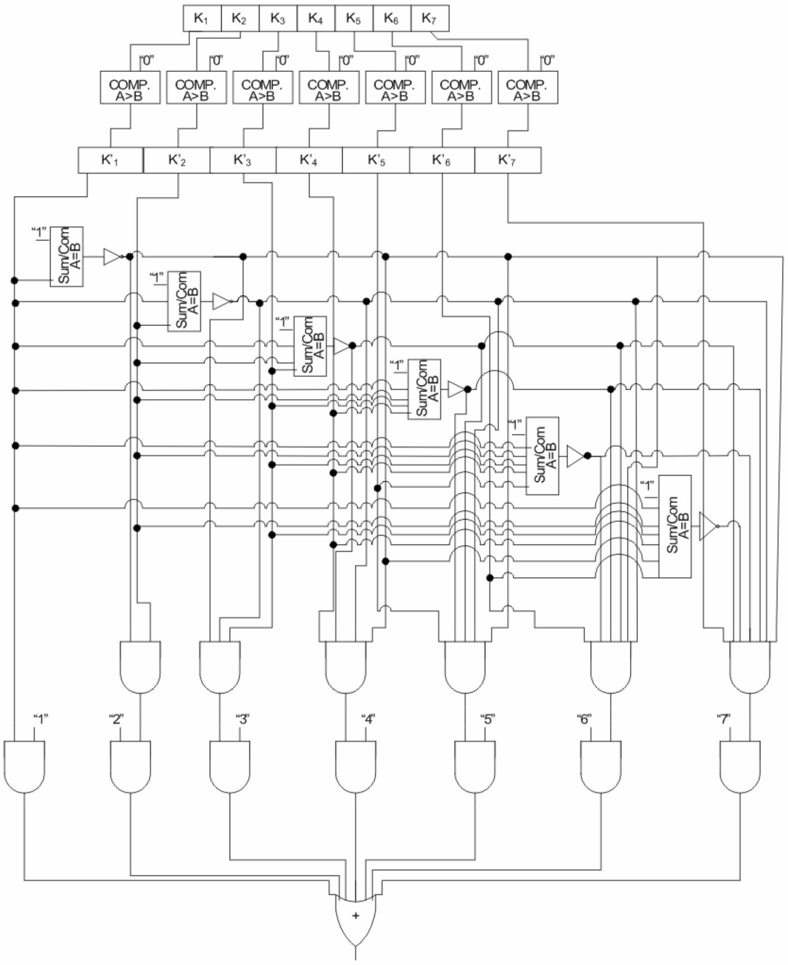


Figure 3: Detecting first rule with $k_i > 0$

As we can see in figure 3, to do this, first we need to transform k values, that can be greater than one, to another values (1 or 0) representing that $k_i$ has a value greater than zero or not. This can be done with a comparator.

Thus, there is a need to have a specific circuit to detect the first “1” in the register, that would be the position of the first rule that has a non zero value. In figure 3 we can see that there is a comparator that sets the position of the value “1” by deactivating the logical gates after it finds the value. Comparison with values 1 to 7 brings us the value of the position for the first rule that has a non zero value for the random number k. There has to be another specific circuit to detect the value for the second rule, the third, etc. Of course, these circuits are similar to the one shown but they ignore the registers behind the position they are looking for.

If we call each of these circuits A, B, C, D…, all of them should be added as we can see in figure 4, in such a way that the first values are all different from zero. Now, all we have to do is to generate a random value no greater than the position of the last number greater than zero. To achieve this, we just have to add the number of values different from zero that are stored in the register and use it as the input for the random generator. The output will be a number between 1 and the number of values different from zero. If we use it as the index for the multiplexer, we will obtain always a value indicating the position of a rule which random k is different from zero and so, we are sure the rule is applicable and active.

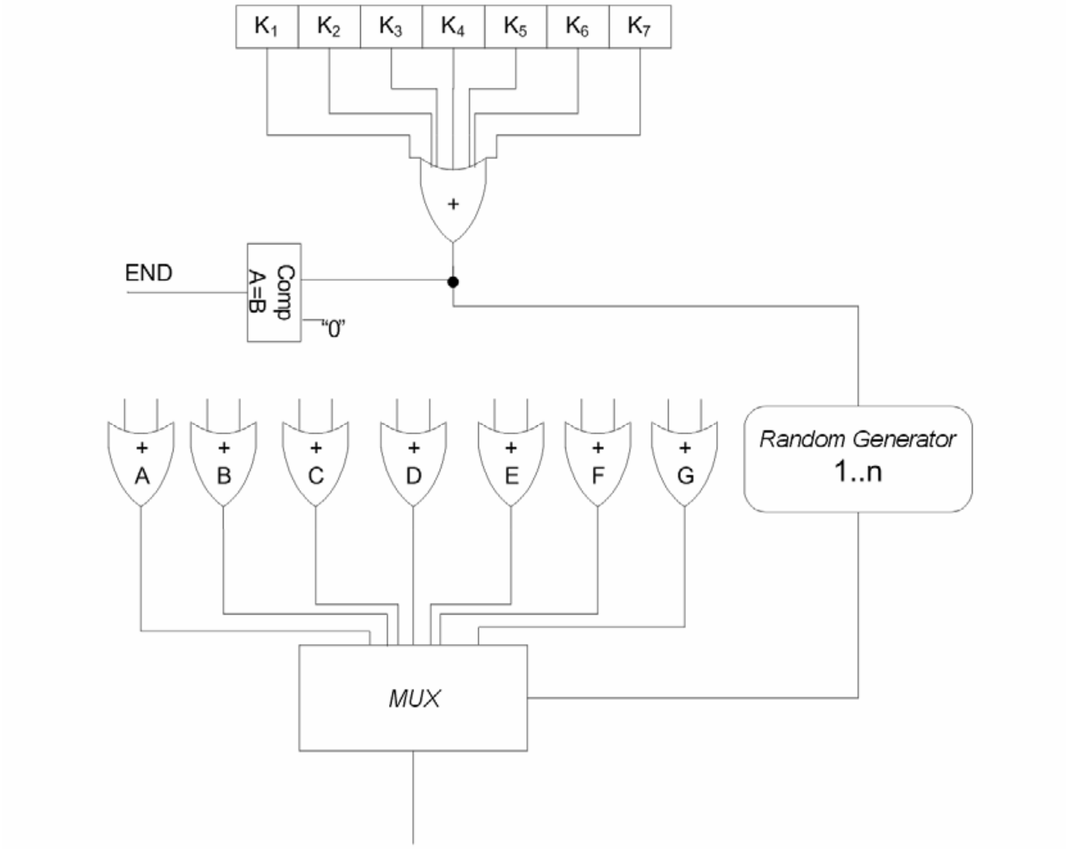


Figure 4: End signal and output for random generator $k_i$ <> 0

Of course, if no k value is different from zero, the addition would result in a zero value, which, once compared with “0”, results in the "*END*" signal for all the circuit because it means that no more rules are applicable.

**Rule decoder:** As we can see in figure 1, once a rule is chosen by the *Application Selector F.U.*, we need to perform two different processes: the first one is to calculate the occurrences of elements used to be able to decrement them from the multiset of objects. But there is still another problem: we should be able to register in the *Application Rules Register* the number of times each rule was applied. This means that if the rule applied i was one that belonged to P’(R) but was not in R (possible due to the way we conformed P’(R)), we have to “decode” that rule to the set of rules that conformed R. Circuit in figure 5 shows how it can be done.

The first set of comparators select the rule indicated by the functional unit “*Application Selector*”. As we just have seen, the value for the selected rule, $k_i$, can not be zero. Once we this value, we have to separate the components that conform it.

So, in the example with 3 rules for R ({ R1, R2, R3}) and 7 rules in P'(R):

P'(R) = { $R_1$, $R_2$, $R_3$, $R_1$ $R_2$, $R_1R_3$, $R_2R_3$, $R_1R_2R_3$}

If rule 1 is selected, the circuit will add only the k value for this rule (gate at the left), but if rule 7 is selected, then it will add the k value to rules $R_{1,}$ $R_2$ and $R_3$ because rule 7 is $R_1R_2R_3$ and all of them were applied k times.

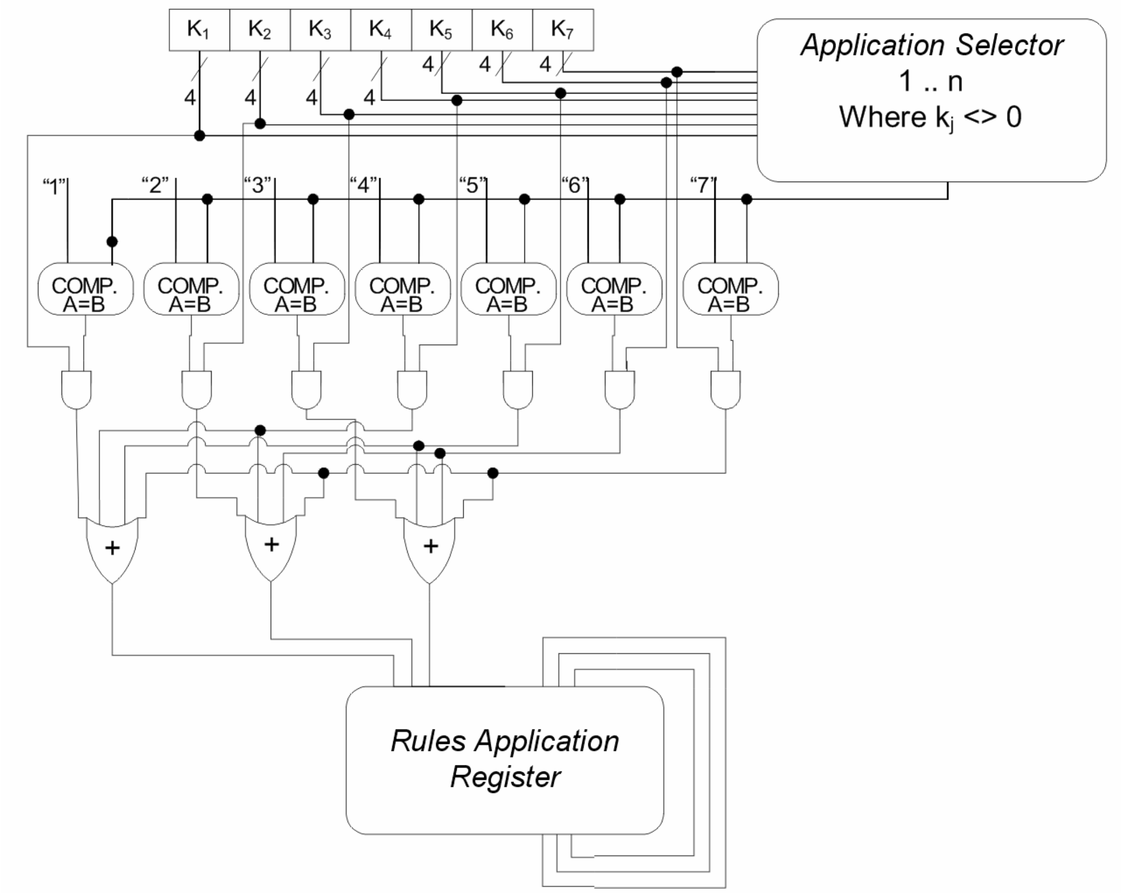


Figure 5: Rule decoder

Whenever the *Application Selector F.U.* enables de *END* signal the "Rules Application Register" will contain the final result, that is, the number of times each rule has to be applied to go forward with a transition. This number is referring the initial set of rules, R, and thus, should be $\log_2 n$ long.

**Other functional units:** Of course, there are more functional units that are in charge of calculating the final amount of elements that the circuit used during each step of evolution. There is a unit that is dedicated to calculate, for each rule, the result of multiplying k (random number generated by the first generator) by each *input($R_i$)* (elements in the antecedent of each rule). Of course this can be done for all the rules at the same time and it just needs a multiplier per rule.

The second one is just a multiplexer in charge of receiving the rule number (j) selected by the *Application Selector F.*U. and to select, according with it, the product $k_j$ * *input($R_j$)*.

Once this is done, we need just to decrement the product selected before from the global multiset, and this is the job for the last functional unit, storing its result in the Multiset Register of Objects to allow a new selection of a rule and let whole process go again.

## Conclusion

Nowadays there are several projects trying to conform different types of circuits to implement membrane computational model with hardware, obtaining active rules and forcing the system to evolution and obtain the number of rules applied. This paper presents how to improve this kind of circuits by emphasizing the massive parallel character P-systems have.

The circuit provides the number of times each rule should be applied to do a complete transition between two configurations, according to its initial set of rules and initial multiset of objects. Of course, different applications over the same sets, do not have to produce the same result.

Hardware implementation is based on basic components like registers, counters, multiplexers, logical gates and so on. The development of the system can be done using hardware-software architectures like VHDL and physical implementation can be accomplished on hardware programmable devices like FPGA's.

## Bibliography

[Arroyo, 2004a] F. Arroyo, C. Luengo, Castellanos, L.F. de Mingo. A binary data structure for membrane processors: Connectivity Arrays. A. Alhazov, C. Martin-Vide, G. Mauri, G. Paun, G.Rozenberg, A. Saloma (eds.): Lecture Notes in Computer Science, 2933, Springer Verlag, 2004, 19-30.

[Arroyo, 2004b] F. Arroyo, C. Luengo, Castellanos, L.F. de Mingo. Representing Multisets and Evolution Rules in Membrane Processors. Pre-proceedings of the Fifth Workshop on Membrana Computing (WMC5). Milano, Italy. June 2004, 126-137.

[Gutiérrez-Naranjo, 2006] M.A. Gutiérrez-Naranjo, M.J. Pérez-Jiménez, A. Riscos-Nez. Available membrane computing software. In G. Ciobanu, Gh. Paun, M.J. Pérez (eds.) Applications of Membrane Computing.Berlin, Germany. Springer Verlag, 2006. pp.411-436. ISBN: 3-540-25017-4.

[Martínez, 2006a] V. Martinez, L.Fernández, F.Arroyo, I.García, A. Gutierrez. A HW circuit for the application of Active Rules in a Transition P System Region. Proceedings on Fourth International Conference Information Research and Applications (i.TECH-2006). Varna (Bulgary) June, 2006. pp. 147-154. ISBN-10: 954-16-0036-0.

[Martínez, 2006b] V. Martínez, L. Fernández, F. Arroyo, A. Gutierrez. HW Implementation of a Bounded Algorithm for Application of Rules in a Transition P-System. Proceedings on 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC-2006). Timisoara (Romania) septiembre, 2006. pp. 32-38.

[Păun, 1998] Gh.Păun. Computing with membranes. Journal of Computer and System Sciences, 61 (2000), and Turku Center for Computer Science-TUCS Report No 208, 1998.

[Păun, 1999] Gh.Păun. Computing with membranes. An introduction. Bulletin of the EATCS, 67, 139-152, 1999.

[Petreska, 2003] B. Petreska and C. Teuscher. A hardware membrane system. A. Alhazov, C. Martin-Vide, Gh. Paun (eds.):Pre-proceedings of the workshop on Membrane Computing Tarragona, July 17-22 2003, 343-355.

## Authors' Information

***Santiago Alonso Villaverde*** *– Natural Computing Group of Universidad Politécnica de Madrid. - Dpto. Organización y Estructura de la Información de la Escuela Universitaria de Informática, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: salonso@eui.upm.es*

***Luis Fernández Muñoz*** *– Natural Computing Group of Universidad Politécnica de Madrid. - Dpto. Lenguajes, Proyectos y Sistemas Informáticos de la Escuela Universitaria de Informática, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: setillo@eui.upm.es*

***Fernando Arroyo Montoro*** *– Natural Computing Group of Universidad Politécnica de Madrid. - Dpto. Lenguajes, Proyectos y Sistemas Informáticos de la Escuela Universitaria de Informática, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: farroyo@eui.upm.es*

***Javier Gil Rubio*** *– Natural Computing Group of Universidad Politécnica de Madrid. - Dpto. Organización y Estructura de la Información de la Escuela Universitaria de Informática, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: jgil@eui.upm.es*

# A HIERARCHICAL ARCHITECTURE WITH PARALLEL COMUNICATION FOR IMPLEMENTING P SYSTEMS

**Ginés Bravo, Luis Fernández , Fernando Arroyo, Juan A. Frutos**

***Abstract***: *Membrane systems are computational equivalent to Turing machines. However, its distributed and massively parallel nature obtain polynomial solutions opposite to traditional non-polynomial ones.*

*Nowadays, developed investigation for implementing membrane systems has not yet reached the massively parallel character of this computational model. Better published approaches have achieved a distributed architecture denominated "partially parallel evolution with partially parallel communication" where several membranes are allocated at each processor, proxys are used to communicate with membranes allocated at different processors and a policy of access control to the communications is mandatory. With these approaches, it is obtained processors parallelism in the application of evolution rules and in the internal communication among membranes allocated inside each processor. Even though, external communications share a common communication line, needed for the communication among membranes arranged in different processors, are sequential.*

*In this work, we present a new hierarchical architecture that reaches external communication parallelism among processors and substantially increases parallelization in the application of evolution rules and internal communications. Consequently, necessary time for each evolution step is reduced. With all of that, this new distributed hierarchical architecture is near to the massively parallel character required by the model.*

***Keywords:*** *Architecture, hierarchy, P systems*

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

## Introduction

Possibilities offered by Natural Computation and, specifically P-Systems, for solving NP-problems, have made researchers concentrate their work towards HW and SW implementations of this new computational model. Transition P systems were introduced by [Păun, 1998]. They were inspired by "basic features of biological membranes". One membrane defines a region where there are a series of chemical components (multisets) that are able to go through chemical reactions (evolution rules) to produce other elements. Inside the region delimited by a membrane can be placed other membranes defiining a complex hierarchical structure that can be represented as a tree. Generated products by Chemical reactions can remain in the same region or can go to another region crossing a membrane. As a result of a reaction, one membrane can be dissolved (its chemical elements are transferred to the container membrane) or can be inhibited (the membrane becomes impermeable and not let objects to pass through).

Membrane systems are dynamics because chemical reactions produce elements that go through membranes to travel to other regions and produce new reactions. This dynamic behaviour is possible to be sequenced in a series of evolution steps between one system configuration to another. These system configurations are determined by the membrane structure and multisets present inside membranes. In the formal Transition P systems model can be distinguished two phases in each evolution step: rules application and communication. In application rules phase, rules of a membrane are applied in parallel to the membrane multiset inside of it. Once application rules phase is finished, then it begins communication phase, where those generated multisets travel through membranes towards their destination in case it is another region. These systems carry out computations through transitions between two consecutive configurations, what turn them into a computational model with the same capabilities as Turing machines.

Power of this model lies in the fact that the evolution process is massively parallel in application rules phases as well as in communication phase. The challenge for researchers is to achieve hardware and/or software implementations of P systems respecting the massively parallelism in both phases. The goal of this work is to design a new hierarchical communication architecture that approaches the best possible way to the inherent characteristics of P systems: application and communication massively parallel.

This paper is structured in the following way: in the first place, the related works are enumerated analyzing the proposed architectures, next a communication hierarchical architecture model is introduced stating detailed analysis of the model. Afterward a comparative analysis with other architectures is presented and finally the conclusions obtained are presented.

## Related Works

In [Syropoulos, 2003] and [Ciobanu, 2004] distributed P systems implementations are presented. They use respectively, the Java Remote Method Invocation (RMI) and the Message Passing Interface (MPI) over a PC cluster's Ethernet network. These authors don't make a detailed analysis about importance of time spent in communication phase respect total time of P system evolution, although Ciobanu declares that "the response time of the program has been acceptable. There are however executions that could take a rather long time due to unexpected network congestion" [Ciobanu, 2004].

In reply to this problem, [Tejedor, 2007] presents an analysis of an architecture named "partially parallel evolution with partially parallel communication". This architecture is based on the following pillars:

a. Membranes distribution. At each processor, *K* membranes are allocated that will evolve, at worst, sequentially. Where,

$$K = \frac{M}{P} \quad ,K \geq 1 \tag{1}$$

and *M* is the total number of membranes of the P system and *P* is the number of processors of the distributed architecture. The physical interconnection of processors is made through a shared communication line. In this scenario, there are two sorts of communications,

- internal communications that are the ones that occur between membranes allocated at the same processor, and whose communication times is negligible because they are carried out using shared memory techniques.
- external communications that are those that occur between different processors because the membranes that needs to communicate are in different processors.

The benefit obtained is that the number of the external communications decreases.

b. Proxy for processor. Membranes that are in different processors do not communicate directly. They do by the means of proxys hosted at their respective processor. Proxys are used to communicate among processors. A proxy assumes communications among membranes of one processor towards the proxy of another one. In the same way, when information from other proxys is receive, it is redistributed to the membranes of the processor.

The benefit of using proxys in the communication among membranes instead of direct communication occurs because the communication protocols penalize the transmission of small packets due to protocol overhead. So, communicate N messages of L length is slower than one message of (S * L) length.

c. Tree topology of processors. The benefit obtained by using a tree topology in the processors interconnection is that the total number of external communications is minimized due to proxys only communicate with their direct ancestor and direct descendants. This way, total number of external communications is 2(P-1).

d. Token passing in the communications. In order to avoid collision and network congestion, it has been established and order in the communication. The idea is not to have more than one proxy trying to transmit at the same time.

The analysis of this distributed architecture leads to the following conclusions:

- This solution avoids communication collisions and reduces the number and length of the external communications.
- In this model, minimum time for an evolution step ($T_{min}$) is determined by the formula:

$$T_{\min} = 2\sqrt{2\ M\ T_{apl}\, T_{com}} - 2T_{com} \tag{2}$$

where, $\boldsymbol{T_{apl}}$ is the maximum time used by the slowest membrane in applying its rules**,** and $\boldsymbol{T_{com}}$ is the maximum time used by the slowest membrane for communication

- The number of processors ($P_{opt}$) that leads to the minimum time is:

$$P_{opt} = \sqrt{\frac{T_{apl}\ M}{2T_{com}}} \tag{3}$$

## Hierarchical Architecture

Previous model parallelize over $P_{opt}$ processors the application of rules and the internal communications among membranes in the same processor. On the other hand, external communications, necessaries for the communication among membranes allocated at the same processor, are sequential. For that reason, we propose a variation that permits to parallelize, up to a certain degree, external communications among nodes. This way, time of an evolution step is reduced drastically and it will tend towards the massively parallel character of a P system.

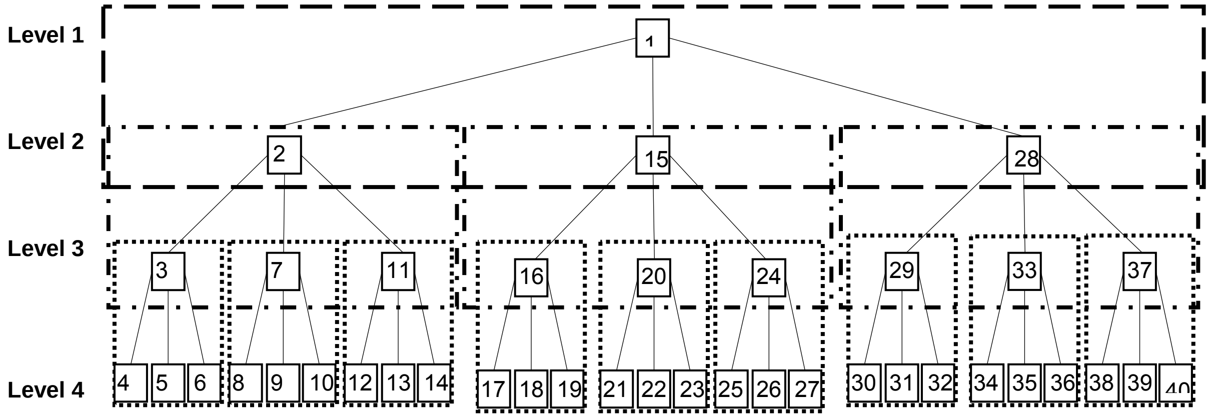


Figure 1. Hierarchical Architecture of 4 levels and amplitude equal to 3.

The new architecture consists of having at its distribute the processors in a hierarchical way, specifically, in a balanced tree of *N* levels depth and *A* processors in amplitude. For instance, figure 1 shows a balanced tree of *N* = 4 and *A* = 3.

For example of figure 1, when every node have applied its rules in parallel, external communications are carried out sequentially in each one of the 9 subtrees arranged between levels 3 and 4; hence, at every instant, as many external communications are carried out as subtrees exist between levels 3 and 4. Subsequently, external communications in each one of the three subtrees arranged between levels 2 and 3 are carried out sequentially; hence, at every instant, as many external communications are carried out as subtrees exist between levels 2 and

3. And finally, external communications in the subtree arranged between levels 1 and 2 are carried out sequentially.

From a logical point of view, each subtree requires a particular physical network to reach the parallelism of its external communications. This way, the processors of intermediate subtrees need 2 communication interfaces, one for the network of the subtree which is root, and another one for network of the subtree which is a leaf. On the other hand, only one interface is required for the processors in the extreme levels 1 and N because they are part of just one subtree. On the other hand, from a physical point of view, the number of logical networks can be reduced to one using Ethernet switches because they permit the separation of collision domains.

Figure 2 chronogram shows the parallelism in application times and in the external communications of the previous example.

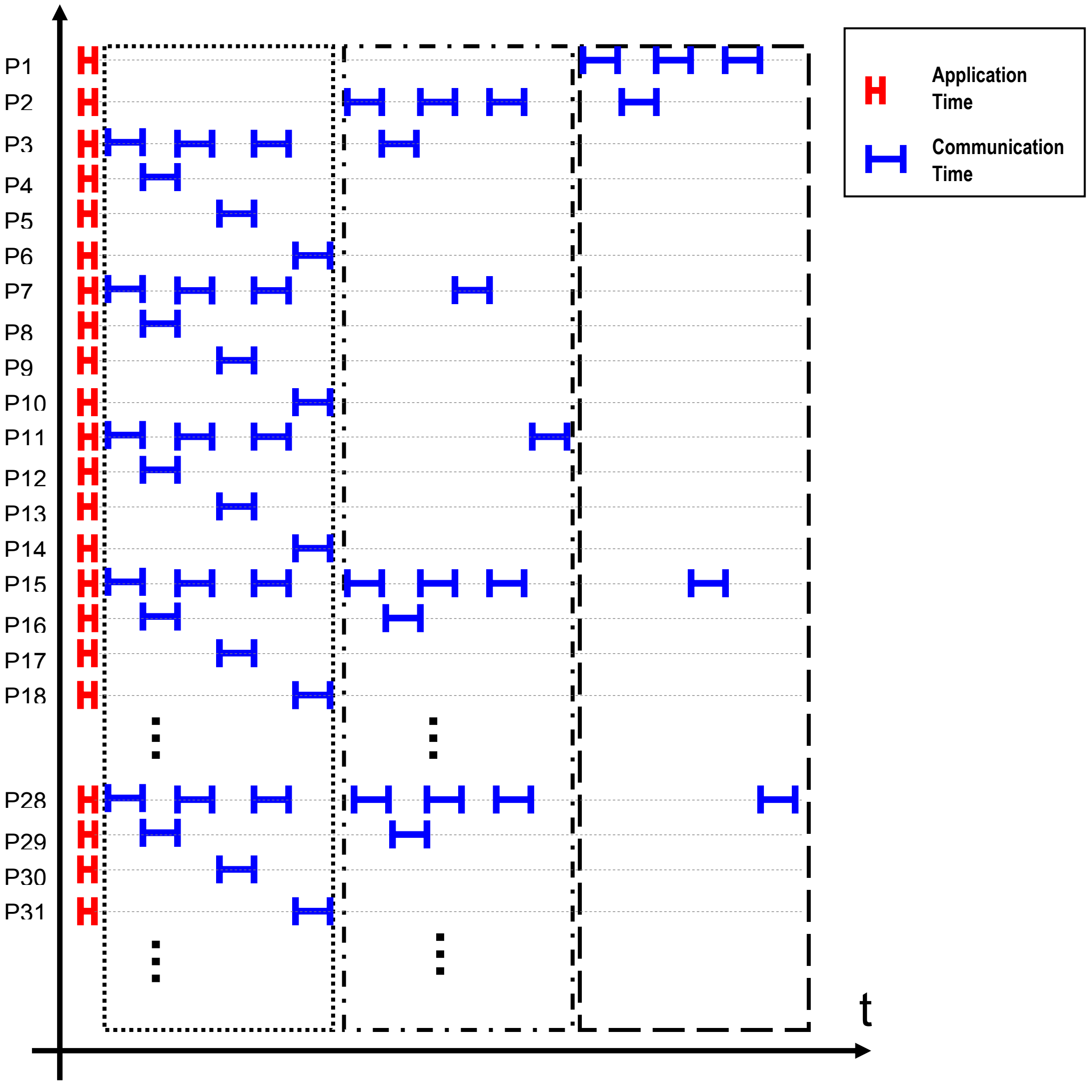


Figure 2. Chronogram of the Hierarchical Architecture of 4 levels and amplitude equal to 3.

Considering the hierarchical distribution of processors, the pillars of this model are:

a. Membranes distribution as in [Tejedor, 2007].

b. Proxy for processor as in [Tejedor, 2007].

c. Balanced tree topology of processors. Benefit obtained from this interconnection topology among processors is that the number of total external communications is minimized because proxys only exchange information with their direct descendants so, total number of external communications is *2(P-1)*, where

$$P = \frac{A^N - 1}{A - 1} \tag{4}$$

d. Token passing in the communication. A sequential order of communication is established for each processor in the same subtree; this way, there can not be more than proxy trying to transmit at the same time in the same subtree which is in. But, sequential external communications of a subtree are carried out in parallel with the ones of any other subtree of the same level. Last, established order for different levels is bottom-upm, i. e., no subtree of a given level begins its communications until every subtree of lower levels have finished.

   This communication policy avoids collisions and network congestion, but additionally permits to be parallelized the 2(P-1) external communications so, the longest external communication sequence in each evolution step will be:

$$2(A-1)(N-1) \tag{5}$$

Hence, in this hierarchical architecture *K* membranes have been located in each processor. At the worst, the application of the rules in each one of these membranes will be made sequentially in each processor. Therefore, the required time to carry out the application of the rules of *M* membranes will be:

$$K\,T_{apl} \tag{6}$$

From (1), (4) and (6) the required time to carry out the application of the rules of *M* membranes will be:

$$M\,\frac{(A-1)}{A^N - 1} T_{apl} \tag{7}$$

On the other hand, from (5) it is obtained the required time to carry out the communication among processors of the architecture:

$$2(A-1)(N-1)T_{com} \tag{8}$$

Therefore, from (7) and (8) the required time to perform a complete evolution step will be:

$$T = M\,\frac{(A-1)}{A^N - 1} T_{apl} + 2T_{com}(N-1)(A-1) \tag{9}$$

Once the required time to perform an evolution step is known, we can determine the number of levels (*L<sub>opt</sub>*) and the amplitude (*A<sub>opt</sub>*) of the architecture in order to minimize this time:

$$A_{opt} = 2 \tag{10}$$

$$L_{opt} = \frac{\ln\left(\sqrt{T_{apl}\frac{M}{T_{com}}\ln(2)}\cdot\sqrt{T_{apl}\frac{M}{T_{com}}\ln(2)+8}+T_{apl}\frac{M}{T_{com}}\ln(2)+4\right)}{\ln(2)} - 2 \tag{11}$$

From (9) and (10) the minimum time required to perform an evolution step is:

$$T_{\min} = \frac{M}{2^{L_{opt}} - 1} T_{apl} + 2T_{com}(L_{opt} - 1) \tag{12}$$

And, from (4) and (10) the number of processors necessary to run the P system minimizing the necessary time to carry out an evolution step will be:

$$P_{opt} = 2^{L_{opt}} - 1 \tag{13}$$

## Comparative Analysis

In this section, we present an empirical analysis comparing proposed architectures in [Tejedor, 2007] with the hierarchical architecture proposed here.

Figure 3 shows the number of processors of both architectures to reach their respective optimum times for an evolution step. As it can be seen, hierarchical architecture have a bigger number of processors than previous work. Also, the growing slope becomes steeper as the number of membranes of the P system is growing. This way, hierarchical architecture reaches a better parallelism degree in proportion to a bigger number of processors in the architecture. This fact increases the parallel application of evolution rules and the internal communication among membranes allocated at the same processor.

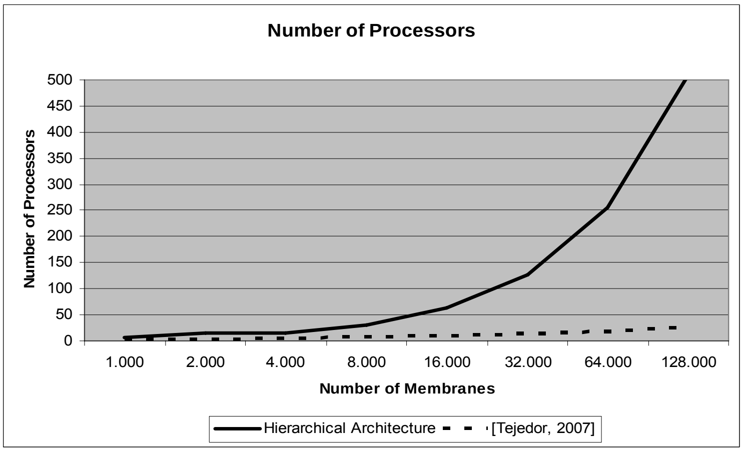


Figure 3. Number of processors to reach optimum times per evolution step among membranes in both architectures.

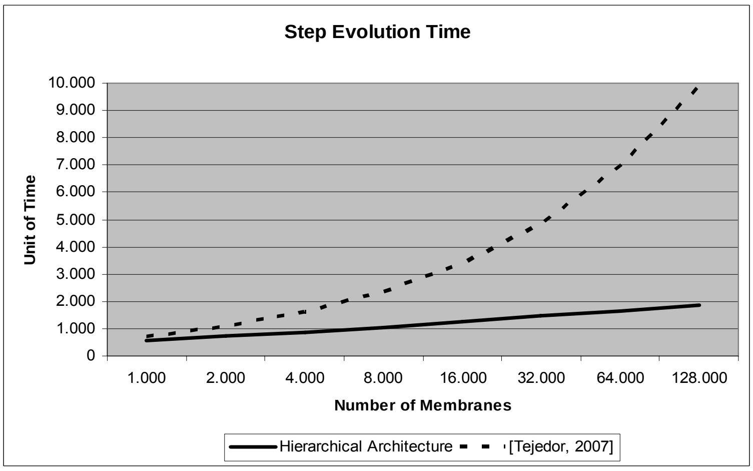


Figure 4. Optimum times per evolution step in both architectures.

Consequently, the bigger parallelization degree of our architecture and external communications parallelization between subtrees of same level obtains smaller minimum times per evolution step. Figure 4 shows resulting times for both architectures as the number of membranes of the P system grow up.

## Conclusions

In this paper a hierarchical architecture of communications to implement P system has been introduced. This architecture is based on the location of several membranes at the same processor, the use of proxys for communicating processors placed in a balanced tree topology and token passing in the communication.

This solution, just like previous architectures, avoids communication collisions, reduces the number and length of the external communications, but permits for the first time the parallelization of external communications and increases drastically the application rules and internal communications parallelization degree. All this, allows us to obtain a better step evolution time than any other suggested architectures and is closer to the massively parallelism character inherent to the membranes computer model.

## Bibliography

[Păun, 1998] Gh.Păun. Computing with membranes. Journal of Computer and System Sciences, 61 (2000), and Turku Center for Computer Science-TUCS Report No 208, 1998.

[Tejedor, 2007] A. Tejedor, L. Fernandez, F. Arroyo, G. Bravo, An architecture for attacking the bottleneck communication in P Systems. In: M. Sugisaka, H. Tanaka (eds.), Proceedings of the 12th Int. Symposium on Artificial Life and Robotics, Jan 25-27, 2007, Beppu, Oita, Japan, 500-505.

[Ciobanu, 2004] G.Ciobanu, W.Guo. P Systems Running on a Cluster of Computers. Workshop on Membrane Computing (Gh. Păun, G. Rozenberg, A. Salomaa Eds.), LNCS 2933, Springer, 123-139, 2004.

[Syropoulos, 2003] A. Syropoulos, E.G. Mamatas, P.C. Allilomes, K.T. Sotiriades, A distributed simulation of P systems, A. Alhazov, C. Martin-Vide and Gh. Păun (Editors): Preproceedings of the Workshop on Membrane Computing; Tarragona, July 17-22 2003, 455-460.

## Authors' Information

***Ginés Bravo García –*** *e-mail: gines@eui.upm.es*

***Luis Fernández Muñoz –*** *e-mail: setillo@eui.upm.es*

***Fernando Arroyo Montoro –*** *e-mail: farroyo@eui.upm.es*

***Juan Alberto Frutos Velasco –*** *e-mail: jafrutos@eui.upm.es*

*Natural Computing Group of Universidad Politécnica de Madrid, Ctra. de Valencia, km. 7, 28031 Madrid (Spain)*

# STATIC ANALYSIS OF USEFULNESS STATES IN TRANSITION P SYSTEMS

**Juan Alberto Frutos, Luis Fernandez, Fernando Arroyo, Gines Bravo**

***Abstract**: Transition P Systems are a parallel and distributed computational model based on the notion of the cellular membrane structure. Each membrane determines a region that encloses a multiset of objects and evolution rules. Transition P Systems evolve through transitions between two consecutive configurations that are determined by the membrane structure and multisets present inside membranes. Moreover, transitions between two consecutive configurations are provided by an exhaustive non-deterministic and parallel application of evolution rules. But, to establish the rules to be applied, it is required the previous calculation of useful, applicable and active rules. Hence, computation of useful evolution rules is critical for the whole evolution process efficiency, because it is performed in parallel inside each membrane in every evolution step. This work defines usefulness states through an exhaustive analysis of the P system for every membrane and for every possible configuration of the membrane structure during the computation. Moreover, this analysis can be done in a static way, therefore membranes only have to check their usefulness states to obtain their set of useful rules during execution.*

***Keywords**: Evolution Rules, Usefulness States, Transition P System, Sequential Machines, Static Analysis*

***ACM Classification Keywords**: F.1.1 Computation by abstract devices – Models of computation. D.1.m Miscellaneous – Natural Computing*

## Introduction

Membrane Computing was introduced by Gh. Păun in [Păun, 1998], as a new branch of natural computing, inspired on living cells. Membrane systems establish a formal framework in which a simplified model of cells is considered a computational device. Starting from a basic model, Transition P systems, many different variant have been considered; and many of them have been demonstrated to be, in power, equivalent to the Turing Machine. An overview of this model is described in the next section.

Nowadays, a challenge for researchers of these kinds of devices is to get real implementations of membrane systems with a high degree of parallelism. Accordingly with this fact, there are some published works related to parallel implementation of membrane systems [Ciobanu, 2004 ], [Syropoulos, 2003] and [Tejedor, 2007].

In [Tejedor, 2007] set up two different phases in the inner dynamic of the evolution step: first phase is related to inner application of evolution rules inside membranes; second phase is related to communication among membranes in the systems. Then it is computed the total time the system spend during the evolution step, and what is important to note is the fact that reducing the time membranes spend in the application phase, the system gets an important gain in the total time it needs for the evolution step. The work presents in this paper is to improve the first phase –application of evolution rules inside membranes- getting useful rules in a faster way. In order to do it, it is introduced the concept of *usefulness states* of membranes in Transition P systems. The main idea is to carry out a static analysis of the P system in order to obtain all usefulness states and transitions between states in each membrane. During execution, membranes will obtain the set of useful evolution rules directly from their usefulness states.

This paper is structures as follows: first Transition P systems are formally defined. Second, usefulness states associated to membranes of Transition P systems with rules able to dissolve membranes are established. Third, the inhibition capability in P systems is incorporated. Fourth, a way for encoding usefulness states is introduced in order to reduce the needed space for implementing. Finally, conclusions are presented.

## Transition P Systems

Formally, a transition P system of degree *m* is a construct of the form

$$\Pi = (O, \mu, \omega_1, \ldots, \omega_m, (R_1, \rho_1), \ldots, (R_m, \rho_m), i_0), \text{ where:}$$

- *O* is the alphabet of objects
- $\mu$ is a membrane structure, consisting of *m* membranes, labelled with *1,2,...., m*. It is a hierarchically arranged set of membranes, contained in a distinguished external membrane, called skin membrane. Several membranes can be placed inside a parent membrane; and finally, a membrane without any other membrane inside is said to be elementary.
- $\omega_i \mid 1 <= i <= m$ are strings over O, representing multisets of objects placed inside the membrane with label *i*.
- $R_i \mid 1 <= i <= m$ are finite sets of evolution rules associated to the membrane with label *i*. Rules have the form $u \rightarrow v$, $u \rightarrow v\ \delta$ or $u \rightarrow v\ \tau$, with $u \in O^+$ and $v \in (O^+ \times TAR)^*$, where $TAR=\{here, out\} \cup \{in_j \mid 1 <= i <= m\}$. Symbol $\delta$ represents membrane dissolution, while symbol $\tau$ represents membrane inhibition. $\rho_i$, $1 <= i <= m$, are priority relations defined over $R_i$, the set of rules of membrane *i*.
- $i_0$ represents the label of the membrane considered as output membrane.

The initial configuration of a P system is given by specifying the membrane structure and the multisets of objects placed inside membranes. $C = (\mu, \omega_1, \ldots, \omega_m)$. A transition takes place by application of evolution rules inside each membrane in the system, in a non-deterministic and maximally parallel manner. This implies that every object in the system able to evolve by the application of one evolution rule must evolve and rules are applied in a non-deterministic way. A computation is defined as a sequence of transitions between system configurations in which the final configuration has no objects able to evolve at any membrane of the system.

Figure 1 shows an example of transition P system, although only multiset and rules associated to membrane 1 are represented.

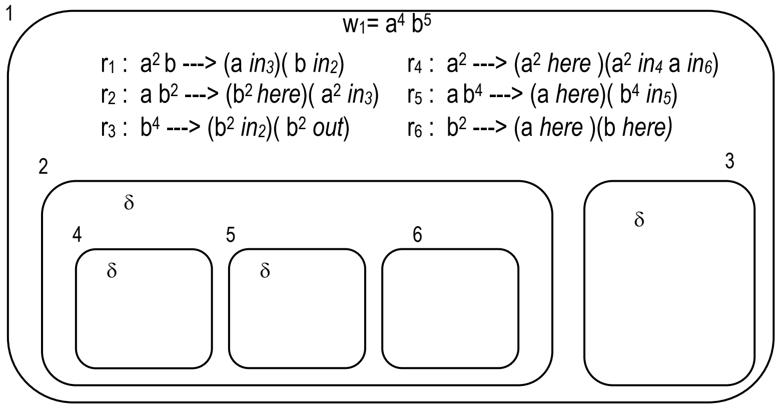

1
$w_1 = a^4 b^5$
$r_1$ : $a^2 b$ ---> (a *in*$_3$)( b *in*$_2$)
$r_2$ : a $b^2$ ---> ($b^2$ *here*)( $a^2$ *in*$_3$)
$r_3$ : $b^4$ ---> ($b^2$ *in*$_2$)( $b^2$ *out*)
$r_4$ : $a^2$ ---> ($a^2$ *here* )($a^2$ *in*$_4$ a *in*$_6$)
$r_5$ : a $b^4$ ---> (a *here*)( $b^4$ *in*$_5$)
$r_6$ : $b^2$ ---> (a *here* )(b *here)*
2
3
δ
4
5
6
δ
δ
δ

Figure 1: Transition P System

## Usefulness states in transition P systems with membrane dissolving capability

Transition P systems with membrane dissolving capability are characterized by *OP*$_m$*(*$\alpha$*,tar,*$\delta$*,pri)*. This notation denotes the class of P systems with simple objects, priorities and dissolving capability. In this class of P systems, rules have the capability for dissolving membranes in the systems and, hence, they can modify the membrane structure of the P system during execution. Evolution rules in these systems are of the form $u \rightarrow v \quad or \quad u \rightarrow v\delta$ . In a different way, $r = (u,v,\xi), where\ \xi \in \{\delta,\lambda\}$ .

Evolution rules able to be applied at any evolution step of the P system must accomplish three requisites: useful, applicable and active. A rule is *useful* in a evolution step if all targets are adjacent and not dissolved. In membrane 1 of the Figure 1, evolution rule $r_4$ is not useful in the initial configuration, but if membrane 2 is dissolved then membrane 4 and 6 become adjacent, and rule $r_4$ useful. On the other hand, a rule is *applicable* if its antecedent is included in the multiset of the membrane. Finally, a rule is *active* if there is no other applicable rule with higher priority.

The main goal of this work is to reduce the time of getting useful rules, avoiding communication as much as possible. The proposed solution is to define the membrane context associated to membranes and configurations in the P system.

**Definition**: The *membrane context* in a time is the set of children membranes to which rules can send objects in the current membrane structure of the system. These membranes are adjacent to the current one.

The basic idea is the following: every membrane in the P system has to know its context at every time. When a membrane is dissolved, then it has to report the dissolution to its father, and the latter will update its context.

**Definition**: a *usefulness state* $q_i^j$ in a membrane *j* is a valid context in that membrane, $C(q_i^j)$ . A context in a membrane is valid if it could be reached in any configuration of the system.

The target of this work is to find out statically all valid usefulness states at any membrane of a P System, the useful rules associated to each usefulness state, and transitions between states when membranes are dissolved.

***Definition***: Let $Child\_Of(j) = \{i \mid 1 <= i <= m;\ i\ is\ a\ child\ of\ \ j\ en\ \mu\}$ , that is, all membrane *j* children in μ.

***Definition:*** Let $Q^j$ the set of *Usefulness states* associated to the membrane labelled with *j* in the P system Π, defined as follows:

1. if the membrane *j* is an elementary membrane: $Q^j = \{q_0^j\}$ , *where* $C(q_0^j) = \{\varnothing\}$
2. if the membrane *j* is not an elementary membrane:

$$Q^j = \underset{i \in Child\_Of(j)}{X} Q'^{\,i} \text{ , where } Q'^{\,i} = \begin{cases} \{q_N^i\} \textit{ if membrane i cannot be dissolved} \\ \{q_N^i\} \cup Q^i \textit{ if membrane i can be dissolved} \end{cases}$$

$q_N^i$ is a state representing that membrane *i* is not dissolved, therefore the context in $q_N^i$ is $C(q_N^i) = \{i\}$.

Context for each one of the states belonging to the *Cartesian product* is obtained by the union of contexts which configure the corresponding state. $C((q_{s_1}^{i_1},..,q_{s_n}^{i_n})) = \bigcup_{i_k \in Child\_Of(j)} C(q_{s_k}^{i_k})$.

Considering the P system depicted in Figure 1, only evolution rules associated to membrane 1 are shown. Other membranes only show if there is any rule which can dissolve them; hence membranes with labels 2, 3, 4 and 5 can be dissolved during execution of the system. In order to determine *Usefulness states* per membrane, we shall start from inside to outside of the membrane system; that is, from elementary membranes to the skin membrane.

It seems to be clear that elementary membranes cannot have more than one state, with null context. Therefore, $Q^3 = \{q_0^3\}$, $Q^4 = \{q_0^4\}$, $Q^5 = \{q_0^5\}$ and $Q^6 = \{q_0^6\}$. Each one of these states has context {∅}.

For membrane 2:

$$Q^2 = Q'^{\,4} \times Q'^{\,5} \times Q'^{\,6}$$

$$Q'^{\,4} = \{q_N^4\} \cup \{q_0^4\} \quad Contexts = \{\{4\},\{\varnothing\}\}$$

$$Q'^{\,5} = \{q_N^5\} \cup \{q_0^5\} \quad Contexts = \{\{5\},\{\varnothing\}\}$$

$$Q'^{\,6} = \{q_N^6\} \quad Contexts = \{\{6\}\}$$

$$Q^2 = \{\overbrace{(q_N^4,q_N^5,q_N^6)}^{q_0^2},\overbrace{(q_N^4,q_0^5,q_N^6)}^{q_1^2},\overbrace{(q_0^4,q_N^5,q_N^6)}^{q_2^2},\overbrace{(q_0^4,q_0^5,q_N^6)}^{q_3^2}\} \quad Contexts = \{\{4,5,6\},\{4,6\},\{5,6\},\{6\}\}$$

And finally, for membrane 1:

$$Q^1 = Q'^{\,2} \times Q'^{\,3}$$

$$Q'^{\,2} = \{q_N^2\} \cup \{q_0^2,q_1^2,q_2^2,q_3^2\} \quad Contexts = \{\{2\},\{4,5,6\},\{4,6\},\{5,6\},\{6\}\}$$

$$Q'^{\,3} = \{q_N^3\} \cup \{q_0^3\} \quad Contexts = \{\{3\},\{\varnothing\}\}$$

$$Q^1 = \{\overbrace{(q_N^2,q_N^3)}^{q_0^1},\overbrace{(q_N^2,q_0^3)}^{q_1^1},\overbrace{(q_0^2,q_N^3)}^{q_2^1},\overbrace{(q_0^2,q_0^3)}^{q_3^1},\overbrace{(q_1^2,q_N^3)}^{q_4^1},\overbrace{(q_1^2,q_0^3)}^{q_5^1},\overbrace{(q_2^2,q_N^3)}^{q_6^1},\overbrace{(q_2^2,q_0^3)}^{q_7^1},\overbrace{(q_3^2,q_N^3)}^{q_8^1},\overbrace{(q_3^2,q_0^3)}^{q_9^1}\}$$

$$Contexts = \{\{2,3\},\{2\},\{4,5,6,3\},\{4,5,6\},\{4,6,3\},\{4,6\},\{5,6,3\},\{5,6\},\{6,3\},\{6\}\}$$

**Useful rules associated to usefulness states**

Every Usefulness state is characterized by its context, that is, the set of children membranes directly enclosed in the original membrane. Hence, the context or state determines the set of useful rules in the membrane. Moreover, what is important to note is that the set of usefulness states, contexts and, hence, the set of evolution rules for each one of the membranes and possible configuration of the system can be established in a static analysis.

***Lemma:*** An evolution rule $r = (u,v\xi), where\, \xi \in \{\delta,\lambda\}$ is useful in $q_i^j$ if and only if $\forall\, TAR\; in_k \in v, k \in C(q_i^j)$.

Considering the previous P system Π for membrane 1, the table 1 shows the whole set of usefulness states – contexts and their corresponding sets of useful evolution rules accordingly to the states.

| Usefulness states | Useful Rules |
|---|---|
| $q_0^1$ {2, 3} | r$_1$, r$_2$, r$_3$, r$_6$ |
| $q_1^1$ {2} | r$_3$, r$_6$ |
| $q_2^1$ {4, 5, 6, 3} | r$_2$, r$_4$, r$_5$, r$_6$ |
| $q_3^1$ {4, 5, 6} | r$_4$, r$_5$, r$_6$ |
| $q_4^1$ {4, 6, 3} | r$_2$, r$_4$, r$_6$ |
| $q_5^1$ {4, 6} | r$_4$, r$_6$ |
| $q_6^1$ {5, 6, 3} | r$_2$, r$_5$, r$_6$ |
| $q_7^1$ {5, 6} | r$_5$, r$_6$ |
| $q_8^1$ {6, 3} | r$_2$,, r$_6$ |
| $q_9^1$ {6} | r$_6$ |

Table 1: Useful Evolution Rules associated to Usefulness States for Membrane 1

**Transitions between usefulness states**

***Definition***: Let $Child\_D(j)=\{i\in Child(j)\wedge\exists r=(u,v,\delta)\in R_j\}$, be the set of child membranes to membrane *j* that can be dissolved.

***Definition***: Let $TC\_D(j)=Child\_D(j)\bigcup_{i\in Child\_D(j)}TC\_D(i)$ , be the total context for membrane *j*, including only those membranes that can be dissolved. By total context is understood those membrane that eventually can become children of membrane *j*.

A transition between two *usefulness states* in a membrane is produced when a child membrane is dissolved. In this case, father membrane is affected and its usefulness state must change. The way for representing this behaviour is through a Moore's Sequential Machine in every membrane labelled with *j*.

$$MS^j=\left(\sum{}_I^j,\sum{}_O^j,Q^j,q_0^j,g^j,f^j\right)\text{, where:}$$

- Input alphabet: $\sum{}_I^j=\{\delta(i,q_s^i)\mid i\in TC\_D(j),\, q_s^i\in Q^i\}$, the sequential machine will transits when a child membrane is dissolved. Child membrane must send to membrane *j* that is dissolved and its usefulness state because the context of the membrane child will pass to be part of the parent context.
- Output alphabet: $\sum{}_O^j=\{r_k\mid r_k\in R_j\}$, the set of useful rules in membrane *j*.
- Set of states: $Q^j=\{(q_{s_1}^{i_1},..,q_{s_n}^{i_n})\mid i_k\in Child\_Of(j)\,,q_{s_k}^{i_k}\in Q'^{\,i_k}\}$ , the set of usefulness states of membrane *j*.
- Initial state: $(q_N^{i_1},..,q_N^{i_n})\mid i_k\in Child\_Of(j)$, that is, the state in which every child membrane is not dissolved.
- Output function: $g^j:Q\to\mathcal{P}(R_j)$. the function that assigns a set of useful rules to each one of the *usefulness state* of the membrane *j*; as it was shown in table 1.
- Transition function: $f^j:Q\times\Sigma_I^j\to Q$. the function provides the new usefulness state to transit given the current one and the dissolution of a child membrane. This function is defined as follows:

  $\forall i_k\in Child\_D(j)$

  1. If $i_k$ is dissolved, $i_k: f^j((q_{s_1}^{i_1},...,q_N^{i_k},...,q_{s_n}^{i_n}),\delta(i_k,q_s^{i_k}))=(q_{s_1}^{i_1},...,q_s^{i_k},...,q_{s_n}^{i_n})$.
  2. If membrane m is dissolved being child of *j* and $m\in TC\_D(i_k)$:

$$f^j((q_{s_1}^{i_1},...,q_{s_k}^{i_k},...,q_{s_n}^{i_n}),\delta(m,q_s^m))=(q_{s_1}^{i_1},...,q_p^{i_k},...,q_{s_n}^{i_n})\quad where\ f^{i_k}(q_{s_k}^{i_k},\delta(m,q_s^m))=q_p^{i_k}$$

Hence, starting from states transition tables of children membranes, it will be obtained the transition table for membrane *j*. Of course, elementary membranes have not transition tables because of they have only one state. As an example, the transition function $f^1$ for membrane 1 of the P system of Figure 1 is depicted in table 2.

In Table 2 transitions for dissolutions of membranes 4 and 5 have been obtained from transition function of membrane 2 –shows in Table 3-, because they belong to the total context of membrane 2.

| | $\delta(2,q_0^2)$ | $\delta(2,q_1^2)$ | $\delta(2,q_2^2)$ | $\delta(2,q_3^2)$ | $\delta(4,q_0^4)$ | $\delta(5,q_0^5)$ | $\delta(3,q_0^3)$ |
|---|---|---|---|---|---|---|---|
| $(q_N^2,q_N^3)$ | $(q_0^2,q_N^3)$ | $(q_1^2,q_N^3)$ | $(q_2^2,q_N^3)$ | $(q_3^2,q_N^3)$ | --- | --- | $(q_N^2,q_0^3)$ |
| $(q_N^2,q_0^3)$ | $(q_0^2,q_0^3)$ | $(q_1^2,q_0^3)$ | $(q_2^2,q_0^3)$ | $(q_3^2,q_0^3)$ | --- | --- | --- |
| $(q_0^2,q_N^3)$ | --- | --- | --- | --- | $(q_2^2,q_N^3)$ | $(q_1^2,q_N^3)$ | $(q_0^2,q_0^3)$ |
| $(q_0^2,q_0^3)$ | --- | --- | --- | --- | $(q_2^2,q_0^3)$ | $(q_1^2,q_0^3)$ | --- |
| $(q_1^2,q_N^3)$ | --- | --- | --- | --- | --- | $(q_3^2,q_N^3)$ | $(q_1^2,q_0^3)$ |
| $(q_1^2,q_0^3)$ | --- | --- | --- | --- | --- | $(q_3^2,q_0^3)$ | --- |
| $(q_2^2,q_N^3)$ | --- | --- | --- | --- | $(q_3^2,q_N^3)$ | --- | $(q_2^2,q_0^3)$ |
| $(q_2^2,q_0^3)$ | --- | --- | --- | --- | $(q_3^2,q_0^3)$ | --- | --- |
| $(q_3^2,q_N^3)$ | --- | --- | --- | --- | --- | --- | $(q_3^2,q_0^3)$ |
| $(q_3^2,q_0^3)$ | --- | --- | --- | --- | --- | --- | --- |

Table 2: Usefulness states transition function for membrane 1.

As an example, if from the state $(q_0^2, q_N^3)$ with context {4,5,6,3}, it is produced $\delta(4,q_0^4)$, then looking at transition table for membrane 2 from $q_0^2$ with $\delta(4,q_0^4)$, the result is $q_2^2$, and then the corresponding transition is to $(q_2^2, q_N^3)$ with context {5,6,3}.

Finally, it can be changed the notation for representing usefulness states, in this case, they are numbering in a correlative manner starting from 0. That is, $\{q_0^1,q_1^1,q_2^1,q_3^1,q_4^1,q_5^1,q_6^1,q_7^1,q_8^1,q_9^1\}$, like in table 3 for membrane 2.

| | $\delta(4,q_0^4)$ | $\delta(5,q_0^5)$ |
|---|---|---|
| $q_0^2$ | $q_2^2$ | $q_1^2$ |
| $q_1^2$ | --- | $q_3^2$ |
| $q_2^2$ | $q_3^2$ | --- |
| $q_3^2$ | --- | --- |

Table 3: Usefulness states transition function for membrane 2.

## Usefulness states in transition P systems with Dissolution and Inhibition Capability.

Evolution rules in these systems are of the form, $u \to v\ \xi$, *where* $\xi \in \{\delta,\tau,\lambda\}$. Symbol τ indicates that after rule application membrane containing the rule will be not permeable to objects communication. This membrane will come back to be permeable to objects communication by the application of one evolution rule having the symbol *δ*. If during the application phase of evolution rules different rules having symbols δ and τ are applied, then membrane will not change its communication state.

Hence, it would be considered three different membrane states concerning to objects communication: Dissolved, Permeable and inhibited or impermeable. These three states and their transition graph are depicted in Figure 2

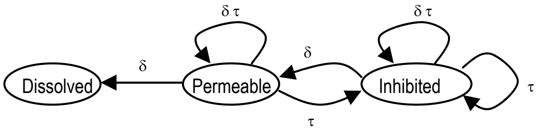


Figure 2

Inhibition capability modifies the previous study for *usefulness states* where only dissolution was allowed. Application of rules having the $\tau$ symbol in a membrane could modify its capability for accepting objects coming from outside. Hence, this fact modifies the context of the parent membrane, because rules sending objects to a membrane in the inhibited state are not useful.

***Definition***: Let $Q^j$ the set of *Usefulness states* associated to the membrane labelled with *j*, defined as follows:

1. if the membrane *j* is an elementary membrane: $Q^j = \{q_0^j\}$, *where* $C(q_0^j) = \{\varnothing\}$
2. if the membrane *j* is not an elementary membrane: $Q^j = \underset{i \in Child\_Of(j)}{X} Q'^{\,i}$, where

$$Q'^{\,i} = \begin{cases} \{q_N^i\} \textit{ if membrane i can be neither dissolved nor inhibited} \\ \{q_N^i, q_I^i\} \textit{ if membrane i can be inhibited, but not dissolved} \\ \{q_N^i\} \cup Q^i \textit{ if membrane i can be dissolved, but not inhibited} \\ \{q_N^i, q_I^i\} \cup Q^i \textit{ if membrane i can be inhibited and dissolved} \end{cases}$$

$q_N^i$ represents the permeable state fo the membrane i, therefore $C(q_N^i) = \{i\}$.

$q_I^i$ represents the inhibited state of the membrane *i*, therefore $C(q_I^i) = \varnothing$.

### Useful rules associated to usefulness states

In order to determine which evolution rules are useful in a determined membrane and evolution step, now it is needed to assure not only that evolution rules targets of type *in<sub>k</sub>* are all of then in the membrane context, but also current membrane must be permeable if target of type *out* is included. Hence, it is needed to consider the *usefulness state* and permeability state of the membrane; and then, it could be possible to abroad the static analysis of usefulness states for P systems with membrane dissolution and permeability control.

***Lemma:*** An evolution rule $u \rightarrow v\ \xi$, *where* $\xi \in \{\delta, \tau, \lambda\}$ is useful in $q_i^j$ *and* $q_{perm}^j$ if and only if $\forall\ TAR\ in_k \in v, k \in C(q_i) \wedge Si\ \exists TAR\ out \in v, q_{perm}^j = Permeable$

**Transitions between usefulness states**

***Definition***: Let $Child\_I(j) = \{i \in Child\_Of(j) \wedge \exists r = (u,v,\tau) \in R_j\}$ be the set of children membranes of membrane *j* that can be inhibited.

***Definition***: Let $TC\_I(j) = Child\_I(j) \bigcup_{i \in Child\_D(j)} TC\_I(i)$ , be the membrane *j* total context considering only those children membranes that can be inhibited.

In these systems, transitions are not only produced by membranes dissolution (δ), but also with membranes inhibition (τ) and come back permeable (¬τ). Therefore, the alphabet for the sequential states machines is:

$$\sum_l^j = \{\delta(i, q_s^i) \mid i \in TC\_D(j), q_s^i \in Q^i\} \cup \{\tau\, i \mid i \in TC\_I(j)\} \cup \{\neg\tau\, i \mid i \in TC\_D(j) \cap TC\_D(j)\}$$

And the transition function is:

$\forall i_k \in Child\_Of(j)$

If $i_k$ is dissolved: $f^j((q_{s_1}^{i_1},...,q_N^{i_k},...,q_{s_n}^{i_n}), \delta(i_k, q_s^{i_k})) = (q_{s_1}^{i_1},...,q_s^{i_k},...,q_{s_n}^{i_n})$

if *m* is dissolved being child of *j* and $m \in TC\_D(i_k)$: $f^j((q_{s_1}^{i_1},...,q_{s_k}^{i_k},...,q_{s_n}^{i_n}), \delta(m, q_s^m)) = (q_{s_1}^{i_1},...,q_p^{i_k},...,q_{s_n}^{i_n})$

$$where\ f^{i_k}(q_{s_k}^{i_k}, \delta(m, q_s^m)) = q_p^{i_k}$$

If $i_k$ is inhibited: $f^j((q_{s_1}^{i_1},...,q_N^{i_k},...,q_{s_n}^{i_n}), \tau\, i_k) = (q_{s_1}^{i_1},...,q_I^{i_k},...,q_{s_n}^{i_n})$

If *m* is inhibited being child of *j* and $m \in TC\_I(i_k)$: $f^j((q_{s_1}^{i_1},...,q_{s_k}^{i_k},...,q_{s_n}^{i_n}), \tau\, m) = (q_{s_1}^{i_1},...,q_p^{i_k},...,q_{s_n}^{i_n})$

$$where\ f^{i_k}(q_{s_k}^{i_k}, \tau\, m) = q_p^{i_k}$$

If $i_k$ comes back to be permeable: $f^j((q_{s_1}^{i_1},...,q_I^{i_k},...,q_{s_n}^{i_n}), \neg\tau\, i_k) = (q_{s_1}^{i_1},...,q_N^{i_k},...,q_{s_n}^{i_n})$

if *m* comes back to be permeable being child of *j* and $m \in TC\_D(i_k) \cap TC\_I(i_k)$:

$$f^j((q_{s_1}^{i_1},...,q_{s_k}^{i_k},...,q_{s_n}^{i_n}), \neg\tau\, m) = (q_{s_1}^{i_1},...,q_p^{i_k},...,q_{s_n}^{i_n})\ where\ f^{i_k}(q_{s_k}^{i_k}, \neg\tau\, m) = q_p^{i_k}$$

## Encoding usefulness states

The main problem when usefulness states are encoded in a determined Hardware/Software architecture could be the size of transition states tables used for representing usefulness states transition functions in membranes. This is the reason why in this paper is proposed a way for encoding usefulness states with the purpose of making transition without using usefulness states transition tables.

***Definition***: Let $TC(j) = Child\_Of(j) \bigcup_{i \in Child\_D(j)} TC(i)$ , the total context for membrane $j$ . Independely of dissolving or inhibition.

The appearing membranes order in $TC(j)$, is normalized going down into the sub-tree of *µ* starting in membrane *j* in depth and in pre-order. And they are represented in this order in the *Normalized Total Context* of membrane *j*.

***Definition***: Let $TC_{Normal}(j) = (i_1, TC_{Normal}(i_1),.....,i_n, TC_{Normal}(i_n))$

$where\ i_k \in Child\_Of(j)\ from\ left\ to\ right\ in\ \mu$

Each one of the usefulness states of membrane *j*, $q_i^j$ is enconded on $TC_{Normal}(j)$ depending on its context, $C(q_i^j)$, with binary logic. The value 1 set out that membrane *k* is present in $C(q_i^j)$, while value 0 will represents that membrane *k* is not in $C(q_i^j)$. As an example, for membrane 1 of the P system depicted in Figure 1, it is obtained the total context $TC_{Normal}(1) = (2,4,5,6,3)$, and the usefulness states enconded are represented in table 4.

If $q^j(t) = (i_1,....,i_k,.....,i_n)$ *encoded by* $TC_{Normal}(j)$ is the usefulness state of membrane *j* at time *t*, the transitional logic will be the following:

1. If the child membrane of *j*, $i_k$ , at time *t* is inhibited: $q^j(t+1) = (i_1,....,0,.....,i_n)$
2. If the child membrane of *j*, $i_k$ , at time *t* comes back to be permeable: $q^j(t+1) = (i_1,....,1,.....,i_n)$
3. If the child membrane of *j*, $i_k$ , at time *t* is dissolved, it has to send its usefulness state $q^{i_k}(t+1)$, encoded by its normalized total context, $TC_{Normal}(i_k)$. It can be considered in a deeper sight the usefulness state for membrane *j* as $q^j(t) = (i_1,....,i_k,TC(i_k),.....,i_n)$ and the transition is $q^j(t+1) = (i_1,....,0,q^{i_k}(t+1),.....,i_n)$

In the proposed example, if membrane 1 is in usefulness state $q^1(t) = (10001)$ and membrane 2 is dissolved in $q^2(t) = (101)$ encoded by its normalized total context $TC_{Normal}(2) = (4,5,6)$, it is obtained the transition $q^1(t+1) = (01011)$. This is the transition of table 2 $f^1((q_N^2,q_N^3),\delta(2,q_1^2)) = (q_1^2,q_N^3)$ without making use of table.

| Usefulness states | Encoding |
|---|---|
| $q_0^1$ {2, 3} | 10001 |
| $q_1^1$ {2} | 10000 |
| $q_2^1$ {4, 5, 6, 3} | 01111 |
| $q_3^1$ {4, 5, 6} | 01110 |
| $q_4^1$ {4, 6, 3} | 01011 |
| $q_5^1$ {4, 6} | 01010 |
| $q_6^1$ {5, 6, 3} | 00111 |
| $q_7^1$ {5, 6} | 00110 |
| $q_8^1$ {6, 3} | 00011 |
| $q_9^1$ {6} | 00010 |

Table 4: Encoding of usefulnes states

## Conclusion

This paper presents the study of usefulness states associated to membranes of Transition P system. The aim of the work developed here is to reduce the evolution rules application time. In order to get the necessary efficiency in the application phase of rules, the analysis of usefulness states can be done in a static manner, and this implies an important reduction in time needed for evolution steps in the system. Moreover, not only usefulness states are defined here, but also the logic of transition between them. Each one of the usefulness states is associated to its own set of useful rules, and in this way there is no computation needed to obtain them because the computation of usefulness states and context is done before starting system execution or simulation.

## Bibliography

[Ciobanu 2004] G.Ciobanu, G.Wenyuan, "A P System runnning on a cluster of computers", Proceedings of Membrane Computing. International Workshop, Tarragona (Spain). Lecture Notes in Computer Science, vol 2933, 123-150.

[Păun, 1998] Gh.Păun, "Computing with Membranes", Journal of Computer and System Sciences, 61(2000), and Turku Center of Computer Science-TUCS Report nº 208, 1998.

[Syropoulos 2003] A. Syropoulos, E.G. Mamatas, P.C. Allilomes, K.T. Sotiriades, "A distributed simulation of P systems". Preproceedings of the Workshop on Membrane Computing (A. Alhazov, C.Martin-Vide and Gh.Păun, eds); Tarragona, vol July 17-22 (2003), 455-460.

[Tejedor, 2007] J.Tejedor, L.Fernández, F.Arroyo, G.Bravo, *An architecture for attacking the bottleneck communication in P systems*. In: M. Sugisaka, H. Tanaka (eds.), Proceedings of the 12th Int. Symposium on Artificial Life and Robotics, Jan 25-27, 2007, Beppu, Oita, Japan, 500-505.

## Authors' Information

***Juan Alberto de Frutos -*** *Dpto. Lenguajes, Proyectos y Sistemas Informáticos (LPSI) de la Escuela Universitaria de Informática (EUI) de la Universidad Politécnica de Madrid (UPM); ); Ctra. Valencia, km. 7, 28031 Madrid (Spain); e-mail: jafrutos@eui.upm.es*

***Luis Fernández -*** *Dpto. LPSI de la EUI de la UPM; e-mail: setillo@eui.upm.es*

***Fernando Arroyo -*** *Dpto. LPSI de la EUI de la UPM; e-mail: farroyo@eui.upm.es*

***Gines Bravo –*** *EUI de la UPM; e-mail: gines@eui.upm.es*

# DELIMITED MASSIVELY PARALLEL ALGORITHM BASED ON RULES ELIMINATION FOR APPLICATION OF ACTIVE RULES IN TRANSITION P SYSTEMS

## Francisco Javier Gil, Luis Fernández, Fernando Arroyo, Jorge Tejedor

***Abstract***: *In the field of Transition P systems implementation, it has been determined that it is very important to determine in advance how long takes evolution rules application in membranes. Moreover, to have time estimations of rules application in membranes makes possible to take important decisions related to hardware / software architectures design.*

*The work presented here introduces an algorithm for applying active evolution rules in Transition P systems, which is based on active rules elimination. The algorithm complies the requisites of being nondeterministic, massively parallel, and what is more important, it is time delimited because it is only dependant on the number of membrane evolution rules.*

***Keywords***: *Natural computing, Membrane computing, Transition P systems, rules application algorithms*

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

## Introduction

Transition P systems are a distributed parallel computational model introduced by Gheorghe Păun based on basic features of biological membranes and the observation of biochemical processes [Păun, 1998]. In this model, membrane contains objects multisets, which evolve according to given evolution rules. Applying the later ones in a nondeterministic maximally parallel way the system changes from a configuration to another one making a computation. This model has become, during last years, an influential framework for developing new ideas and investigations in theoretical computation. "P systems with simple ingredients (number of membranes, forms and sizes of rules, controls of using the rules) are Turing complete" [Păun, 2005]. Moreover, P systems are a class of distributed, massively parallel and non-deterministic systems. "As there do not exist, up to now, implementations in laboratories (neither in vitro or in vivo nor in any electronically medium), it seems natural to look for software tools that can be used as assistants that are able to simulate computations of P systems" [Ciobanu, 2006]. "An overview of membrane computing software can be found in literature, or tentative for hardware implementations, or even in local networks is enough to understand how difficult is to implement membrane systems on digital devices" [Păun, 2005].

In addition, Gheorghe Păun says that: "we avoid to plainly say that we have 'implementations' of P systems, because of the inherent non-determinism and the massive parallelism of the basic model, features which cannot be implemented, at least in principle, on the usual electronic computer -but which can be implemented on a dedicated, reconfigurable, hardware [...] or on a local network". Thereby, there exists many P systems simulators in bibliography but “the next generation of simulators may be oriented to solve (at least partially) the problems of storage of information and massive parallelism by using parallel language programming or by using multiprocessor computers” [Ciobanu, 2006].

This work presents a time delimited massively parallel algorithm based on rules elimination for application of active rules in transition P systems. After this introduction, other related works appear, where the problem that is tried to solve is exposed. Later the massively parallel algorithm of application of rules appears developed, including the synchronization between processes and the analysis of its efficiency.

## Related Work

J. Tejedor proposes a software architecture for attacking the bottleneck communication in P systems denominated “partially parallel evolution with partially parallel communications model” where several membranes are located in each processor, proxies are used to communicate with membranes located in different processors and a policy of access control to the network communications is mandatory [Tejedor, 2006]. This obtains a certain parallelism yet in the system and an acceptable operation in the communications. In addition, it establishes a set of equations that they allow to determine in the architecture the optimum number of processors needed, the required time to execute an evolution step, the number of membranes to be located in each processor and the conditions to determine when it is best to use the distributed solution or the sequential one. Additionally it concludes that if the maximum application time used by the slowest membrane in applying its rules improves $N$ times, the number of membranes that would be executed in a processor would be multiplied by $\sqrt{N}$ , the number of required processors would be divided by the same factor, and the time required to perform an evolution step would improve approximately with the same $\sqrt{N}$ factor.

Therefore, to design software architectures it is precise to know the necessary time to execute an evolution step. For that reason, algorithms for evolution rules application that they can be executed in a delimited time are required, independently of the object multiset cardinality inside the membranes. Nevertheless, this information cannot be obtained with the present algorithms since its execution time depends on the cardinality of the objects multiset on which the evolution rules are applied.

In addition, Ciobanu presents several related papers about parallel implementation of P systems [Ciobanu 2002, 2004, 2006], in which “the rules are implemented as threads. At the initialization phase, one thread is created for each rule. Rule applications are performed in term of rounds” [Ciobanu, 2006]. Again, the author recognizes that: “since many rules are executing concurrently and they are sharing resources, a mutual exclusion algorithm is necessary to ensure integrity” [Ciobanu, 2004]. So, “when more than one rule can be applied in the same conditions, the simulator randomly picks one among the candidates” [Ciobanu, 2006]. Hence, processes will have pre-protocols and post-protocols for accessing to critical sections included into their code in order to work under mutual exclusion. Then, each evolution rule set associated to a membrane must access to the shared multiset of objects under mutual exclusion; but different sets of evolution rules associated to different membranes there are no competition among them because they are disjoint processes. Hence, some degree of parallelism is achieved spite of having a thread for each evolution rule. The implementation of evolution rules application will be concurrent inside membranes but not massively parallel.

On the other hand, L. Fernández proposes a massively parallel algorithm for evolution rules application [Fernández, 2006]. In this solution a process by each rule is generated and exist one more controller process that simulates the membrane containing the objects multiset. In a loop, each rule process proposes simultaneously a object multiset to be consumed and the membrane process determines if it is possible to apply the proposed multiset, until the proposal is correct. The algorithm execution finishes when there is no active rule. This last solution contains a high degree of parallelism, but its execution time is not delimited. Therefore, this algorithm is not appropriate to be used in the previously commented [Tejedor, 2006] software architecture.

Finally, in [Tejedor, 2007] is exposed an algorithm for application of evolution rules based on active rules elimination. In this algorithm, in each loop iteration all the rules -except the last one- are sequentially applied a random number of times. Next, the last active rule is applied the greater possible number of times, reason why it became inactive. This algorithm reaches a certain degree of parallelism, since one rule can be simultaneously several times applied in a single step. In this algorithm, the execution time depends on the number of rules, not of the objects multiset cardinality. In the experimental tests, this algorithm has obtained better execution times than the previously published sequential algorithms. This sequential solution is, of course, a minimal parallelism solution.

## Delimited Massively Parallel Algorithm based on Rules Elimination for Application of Active Rules in Transition P Systems

Here we present a time delimited massively parallel algorithm for application of active rules. The initial input is a set of active evolution rules for the corresponding membrane -the rules are applicable and useful- and the initial membrane multiset of objects. The final results are the complete multiset of applied evolution rules and the obtained multiset of objects after rules application. In order to achieve this, we propose one process for each rule and one more controller process that simulate the membrane containing the multiset of objects.

The general idea is that each rule -except the last one- randomly proposes, in an independent manner, a multiset to be consumed from the membrane multiset (the obtained algorithm is nondeterministic due to this random proposal). If the addition of all the proposed multiset by rules is smaller than the membrane multiset, then the proposed multiset is subtracted from the membrane multiset. Next, the last active rule determines and applies its maximal applicability benchmark over the membrane multiset, subtracting the correspondent multiset from the membrane multiset. At this point, rules that are not applicable over the new membrane multiset finish their process execution –obviously, including the last active rule-. The resting active rules come back to the starting point, and again, propose a new multiset to be consumed. This process is repeated until none rule is applicable over the membrane multiset.

This idea can be divided into seven phases:

Phase 1 *Membrane initialization*. A global probability for proposing multiset to be consumed by rules is initialized. One rule is able to consume with a determined probability; but if a rule is not allowed to consume then it will not propose multiset to be consumed until the next loop iteration.

This phase is performed only by the controller process, while rules are waiting to second phase.

Phase 2 *Evolution rules initialization*. Each rule -except the last active rule- determines its applicability benchmark to its maximal applicability benchmark over the membrane multiset. On the other hand, every rule is settled to the state in which rules can propose.

This phase is performed in parallel by every rule. The controller process -membrane- waits until phase 5.

Phase 3 *Multiset propositions*. Considering the global probability established by the membrane, each rule proposes in a randomly manner one multiset of objects to be consumed from the membrane multiset. The proposed rule multiset can be the empty multiset or the scalar product of its antecedent by a natural number chosen in a random manner in between 1 and its applicability benchmark.

This phase is performed in parallel for every evolution rule, except the last one.

Phase 4 *Sum of Multiset Proposals*. The addition of the proposed multisets by rules is performed two by two by neighborhood with respect to their number. For example, rule number 1 with rule number 2, rule number 3 with rule number 4, and so on. After finishing this step, the resulting multisets are added two by two again. For example, rule number 1 with rule number 3. And so on until reaching one single multiset. This way for adding multiset develops a binary tree of additions performed in parallel at each level of the tree.

This phase is performed in parallel for every rule.

Phase 5 *Proposal management and last active rule maximal application*. Membrane analyzes the proposed multiset by the rules. If the proposed multiset is valid (not empty and included in the membrane multiset), then the membrane subtract from its own multiset the proposed multiset from phase 4. Next, the membrane process determines and applies the maximal applicability benchmark of the last active rule over the membrane multiset, subtracting the correspondent multiset from the membrane multiset. Moreover, the membrane process indicates to the rule processes the executed operation. Finally, it initializes the information about its active evolution rules for the next loop in the algorithm.

This phase is performed only by the membrane process while rules wait until the phase 6.

Phase 6 *Checking rules halt*. Each one of the evolution rules accumulates the number of proposed application over the membrane multiset. Moreover, it computes its maximal applicability benchmark over the new resting membrane multiset for the next iteration and, if it is bigger than 0, they pass to the state in which rules can propose and indicate it into the active evolution rules data structure. Otherwise, they finish their execution.

This phase is performed in parallel by every evolution rule except into the access to the active evolution rules data structure. Membrane is waiting until phase 7.

Phase 7 *Checking membrane halt*. Membrane checks if there exists some active rule for the next loop and, in this case, it returns to establish the global probability to propose multisets by the rules. If so, it come back to phase 5 waiting the proposal management, otherwise it finishes the execution.

This phase is performed only by the membrane and the rules wait for coming back to phase 3 -if they are active for next loop- of for finish their execution.

Next we will deal with two different aspects for the exposed general idea: the phases synchronization and finally efficiency analysis.

## Synchronization Design

Accordingly whit the previous explanation, these phases shared between the two different processes types, evolution rules and membrane, as it can be observed in tables 1 and 2.

```
Phase 1: Membrane initialization
REPEAT
  Phase 5: Proposal Management &
           Last active rule maximal application
  Phase 7: Checking membrane halt
UNTIL End
```

Table 1: Process Type Membrane

```
Phase 2: Rules initialization
REPEAT
  Phase 3: Multiset proposition
  Phase 4: Sum of Multiset Proposals
  Phase 6: Checking rules halt
UNTIL End
```

Table 2: Process Type Evolution Rule

Both processes types are not disjoint and they must preserve the following synchronizations (Fig. 1 presents the activity diagram showing the needed synchronization in the different phases for the process membrane and two evolution rules processes):

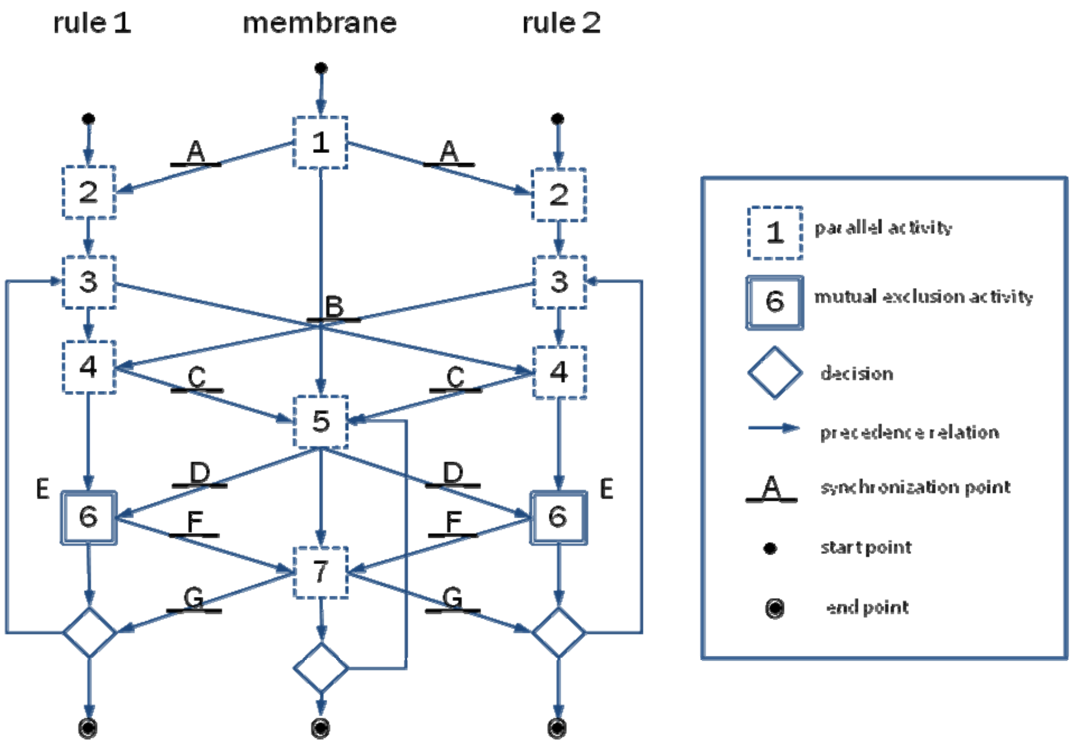


Figure 1: The activity diagram showing the needed synchronization in the different phases for the membrane and two evolution rules

A. Every evolution rule must wait for initialization until membrane initialization finishes.

B. Each evolution rule must wait for their neighbor evolution rules finish their respective additions of proposed multisets by their neighbor evolution rules.

C. Membrane must wait to start management collision until evolution rules finish accumulating the proposed multisets.

D. Every evolution rule must wait to start checking halting condition until membrane finishes multisets subtraction.

E. Every evolution rule must wait for the mutual exclusion to access into the active evolution rule data structure and it can perform its register for the next loop iteration.

F. Membrane must wait to checking halting condition until evolution rules finish their corresponding checking for halting conditions.

G. Every evolution rule must wait to start to determine, it they finish their execution or come back to propose a new multiset, until membrane halt checking finishes.

## Efficiency Analysis

Analyzing the membrane process it can be observed that the proposed algorithm executes, at the most, so many times as rules exist initially (we will denominate it by *R*), since in each iteration at least one rule is eliminated in the worse case. The rest of operations executed by the process membrane can be considered like basic operations. Moreover, the operations executed by the rules processes are simple, except the sum of proposals made in phase 4. This sum is performed two by two by neighborhood, reason why the obtained complexity order is *log*$_2$ $R_i$, being $R_i$ the number of active rules minus one in each loop iteration. Consequently, the complexity order of the proposed algorithm is:

$$\sum_{i=R-1}^{2} log_2 i = log_2(R-1) + log_2(R-2) + \ldots + log_2 2 = log_2(R-1)!$$

Consequently, we can conclude that the complexity order of the proposed algorithm -in the worse case- is *log*$_2$*(R - 1)*!, but better results can be expected experimentally than the ones obtained theoretically, because exists the possibility that in a same loop iteration disappears more than a rule.

## Future Work

In first phase of the presented algorithm -membrane initialization- a global probability for the rule processes is determined. This value determines the probability of proposing an objects multiset by a rule. Assigning a value of 1/*R* to this probability very good results in the made tests have been obtained. At the moment we are working in the process of determination of this value, trying of obtaining a better efficiency.

In addition, since one has studied in this work, the presented algorithm eliminates an active rule in each loop iteration. Evidently, the order in that the rules are applied influences in the final results obtained. Therefore, to improve the algorithm efficiency it seems interesting to study as it would have to be the last applied rule, studying the relations between the antecedents of the rules available.

## Conclusions

This paper introduces an algorithm of active rules application based on rules elimination in transition P systems. The two most important characteristics of this algorithm are:

- The presented algorithm is massively parallel
- The execution time of the algorithm is time delimited, because it only depends on the number of rules of the membrane. The number of rules of the membrane is a well-known static information studying the P system

We think that the presented algorithm can represent an important contribution in particular for the problem of the application of rules in membranes, because it presents high productivity and it allows estimate the necessary time to execute an evolution step. Additionally, this last one allows to make important decisions related to the implementation of P systems, like the related ones to the software architecture.

## Bibliography

[Ciobanu, 2002] G. Ciobanu, D. Paraschiv, “Membrane Software. A P System Simulator”. Pre-Proceedings of Workshop on Membrane Computing, Curtea de Arges, Romania, August 2001, Technical Report 17/01 of Research Group on Mathematical Linguistics, Rovira i Virgili University, Tarragona, Spain, 2001, 45-50 and Fundamenta Informaticae, vol 49, 1-3, 61-66, 2002.

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

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

[Fernández 2006] L. Fernández, F. Arroyo, J. Tejedor, J. Castellanos. “Massively Parallel Algorithm for Evolution Rules Application in Transition P System”. Seventh Workshop on Membrane Computing, WMC7, Leiden (The Netherlands). July, 2006

[Păun, 1998] G. Păun. “Computing with Membranes”. In: Journal of Computer and System Sciences, 61(2000), and Turku Center of Computer Science-TUCS Report nº 208, 1998.

[Păun, 2005] G. Păun. “Membrane computing. Basic ideas, results, applications”. In: Pre-Proceedings of First International Workshop on Theory and Application of P Systems, Timisoara (Romania), pp. 1-8, September, 2005.

[Tejedor, 2006] J. Tejedor, L. Fernández, F. Arroyo, G. Bravo. “An Architecture for Attacking the Bottleneck Communications in P systems”. In: Artificial Life and Robotics (AROB 07). Beppu (Japan), January 2007.

[Tejedor, 2007] J. Tejedor, L. Fernández, F. Arroyo, A. Gutiérrez. “Algorithm of Active Rules Elimination for Evolution Rules Application” (submited). In 8th WSEAS Int. Conf. on Automation and Information, Vancouver (Canada), June 2007.

## Authors' Information

***F. Javier Gil Rubio –*** *Dpto. de Organización y Estructura de la Información, e-mail: jgil@eui.upm.es*

***Luis Fernández Muñoz -*** *Dpto. de Lenguajes, Proyectos y Sistemas Informáticos, e-mail: setillo@eui.upm.es*

***Fernando Arroyo Montoro -*** *Dpto. de Lenguajes, Proyectos y Sistemas Informáticos, e-mail: farroyo@eui.upm.es*

***Jorge A. Tejedor Cerbel -*** *Dpto. de Organización y Estructura de la Información, e-mail: jtejedor@eui.upm.es*

*E.U. de Informática. Natural Computing Group, Universidad Politécnica de Madrid, Spain;*

# RESEARCHING FRAMEWORK FOR SIMULATING/IMPLEMENTATING P SYSTEMS

**Sandra Gómez, Luis Fernández, Iván García, Fernando Arroyo**

***Abstract****: Researching simulation/implementation of membranes systems is very recent. Present literature gathers new publications frequently about software/hardware, data structures and algorithms for implementing P system evolution. In this context, this work presents a framework which goal is to make tasks of researchers of this field easier. Hence, it establishes the set of cooperating classes that form a reusable and flexible design for the customizable evaluation with new data structures and algorithms. Moreover, it includes customizable services for correcting, monitoring and logging the evolution and edition, recovering, automatic generating, persistence and visualizing P systems.*

***Keywords****: P System, framework, simulation, implementation.*

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

## Introduction

P systems are a new computational model based on the membrane structure of living cells. This model has become, during last years, a powerful framework for developing new ideas in theoretical computation. ”P systems with simple ingredients (number of membranes, forms and sizes of rules, controls of using the rules) are Turing complete” [Păun, 1999]. Moreover, P systems are a class of distributed, massively parallel and non-deterministic systems.

”As there do not exist, up to now, implementations in laboratories (neither in vitro or in vivo nor in any electronical medium), it seems natural to look for software tools that can be used as assistants that are able to simulate computations of P systems” [Ciobanu, 2006]. ”An overview of membrane computing software can be found in literature , or tentative for hardware implementations , or even in local networks is enough to understand how difficult is to implement membrane systems on digital devices” [Păun, 2005]. Moreover, he says: ”we avoid to plainly say that we have ’implementations’ of P systems, because of the inherent non-determinism and the massive parallelism of the basic model, features which cannot be implemented, at least in principle, on the usual electronic computer -but which can be implemented on a dedicated, reconfigurable, hardware or on a local network” [Păun, 2005]. Thereby, there exists many simulators in bibliography but ”the next generation of simulators may be oriented to solve (at least partially) the problems of storage of information and massive parallelism by using parallel language programming or by using multiprocessor computers” [Ciobanu, 2006].

The goal of this work is to present a framework to make easier the tasks of researchers who develop simulators/implementations of P systems. It does not expect to be a new simulator/implementation. It presents a set of cooperating classes that form a reusable design for developing simulators/implementations of P systems. This framework provides an architectonical guide to divide the design in abstract classes and to define their responsibilities and collaborations. Researchers have to adapt the framework to a concrete simulator/implementation inheriting and compounding instances of framework classes.

This paper is structured as follows: next section presents related works then, they are presented the requisites and design guidelines for the framework. Finally, conclusions are presented.

## Related Works

Membrane system implementation is a very recent investigation field. First approaches were simulators [Ciobanu, 2006] that demonstrated the functionality of the membrane systems. But, they lacked distributed and massively character.

First distributed implementations are presented in [Syropoulos, 2003] and [Ciobanu, 2004]. In their distributed implementations of P systems use Java Remote Method Invocation (RMI) and the Message Passing Interface (MPI) respectively, on a cluster of PC connected by Ethernet. These last authors do not carry out a detailed analysis of the importance of the time used during communication phase in the total time of P system evolution, although Ciobanu affirms that “the response time of the program has been acceptable. There are however executions that could take a rather long time due to unexpected network congestion” [Ciobanu, 2003]. In [Tejedor, 2007a] [Bravo, 2007a] [Bravo, 2007b], It is determined that the problem in implementing P systems is the time necessary in the communication of multisets among membranes allocated in different devices (PCs, PICs, or chips). This fact, forces to resign parallelism to the maximum to as much reach a parallelism degree dependent of the speed of the communications and the application of the evolution rules. Therefore, it is necessary to develop faster application algorithms that adapt so much to the sequential technologies as to the parallel ones.

On the other hand, [Fernández, 2007] determines the appropriate software architecture that is executed over a given evolution P system hardware architecture. So, it pretends to determine the set of process and their relationships that are appropriate to be executed over a set of connected processors. Considered possibilities are: evolution rules oriented software architecture, membranes oriented software architecture and processors oriented software architecture.

Works of investigation about sequential and/or parallel algorithms designed for the different phases of the evolution of a P system are very varied: for the utility of the evolution rules: [Frutos, 2007]; for the applicability of evolution rules: [Fernández, 2006a]; for the application of evolution rules: [Fernández, 2006b] [Fernández, 2006c] [Tejedor, 2007b] [Tejedor, 2007c] [Gil, 2007].

With respect to the storage of the information of a P System, [Fernández, 2005a] defines a universal vocabulary with XML technology and [Gutiérrez, 2007b] presents new data structures that compress multisets of objects information without penalizing the basic operations on these.

Finally, it is possible to indicate the works on different technologies whose objective is to implant the different architectures, algorithms and previous data structures. Thus, we found a line about circuits hardware in [Petreska, 2003] [Arroyo, 2004a] [Arroyo, 2004b] [Arroyo, 2004c] [Fernández, 2005b] [Martínez, 2006a] [Martínez, 2006b], the new opened line about microcontrollers in [Gutierrez, 2006] [Gutierrez, 2007a], and the traditional line about personal computers in [Syropoulos, 2003] [Ciobanu, 2004].

## Requisites

In this context, pursued goals is to develop a highly reusable framework that is flexible enough for any researcher to be able of concentrating on developing the algorithm or data structure object of its investigation.

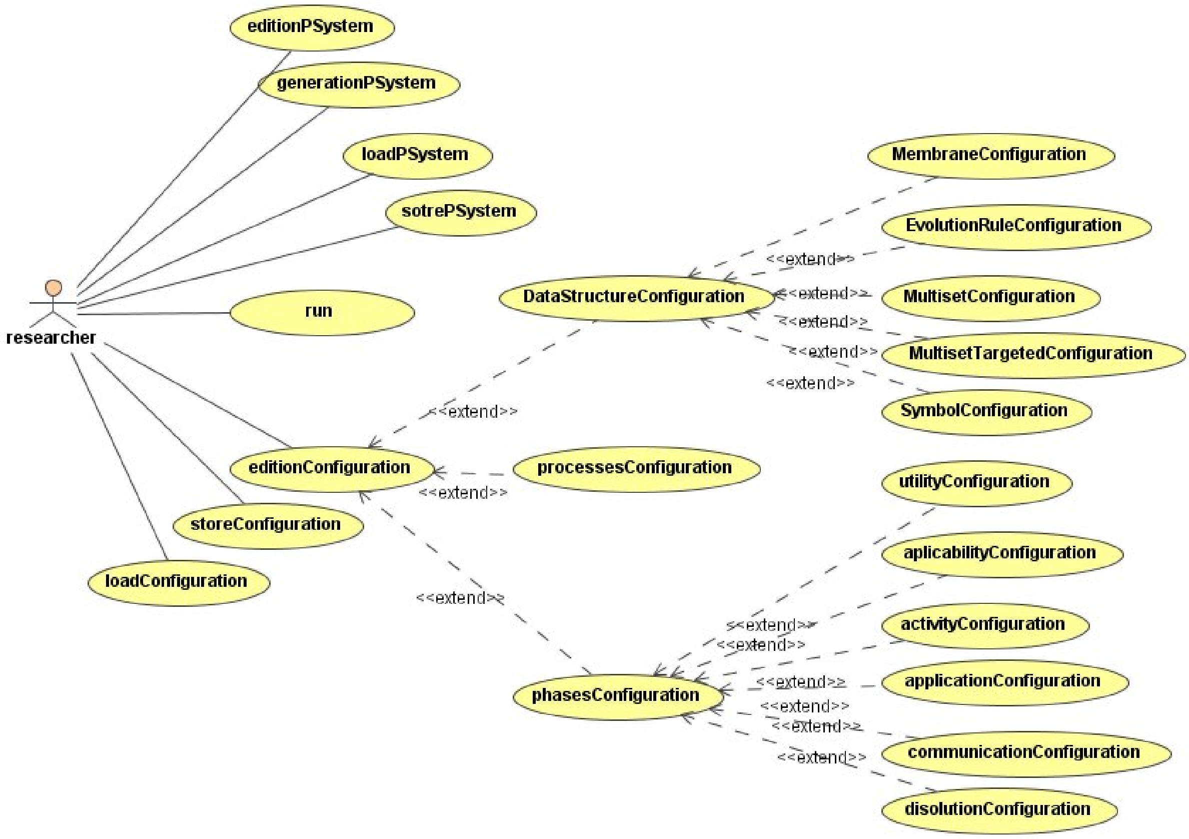


Figure 1: Use Case Diagram.

In this line, the framework provides to the researchers the following reusable modules:

1.1 Implementation of every standard data structure, agreed to the specification model, in order to equip framework with the total functionality of a simulator of membrane system. Hence, it is had the data structures for the symbols, multisets, evolution rules and membranes.

1.2 Implementation of every standard algorithm, agreed to the specification model, in order to equip framework with the total functionality of a simulator of membrane system. Hence, it is hadthe algorithms for utility phases, applicability, activity, application, communication and dissolution.

1.3 Process management and synchronization for the different software architectures: evolution rules oriented, membranes oriented and processor oriented.

1.4 Management of automatic detection of errors of any algorithm for functional tests.

1.5 Management of automatic monitoring (time, space, number of operations, …) of any algorithm for non functional tests.

1.6 Management of persistence, visualizing and logs for tracking the algorithms.

1.7 Management of P System for its edition, recovery, automatic generation of parameterized sets of tests

1.8 Management of configurations for P systems evolution.

1.9 Gestión de configuraciones para la evolución de P Systems.

On the other hand, the framework provides to the researchers the following flexibility for a given evolution of the P system:

2.1. Extension by inheritance of new data structures for the symbols, multisets, evolution rules and membranes.

2.2. Extension by inheritance of new algorithms for utility, applicability, activity, application, communication and dissolution phases.

2.3. Extension by inheritance of new functionalities over P systems (analiysis, compilation, …).

2.4. Architecture process, phases, algorithms and data structures configuration.

2.5. Evolution, visualization, monitorization, correction and logs configuration.

Figure 1 shows the use case diagram corresponding to the previous requisites.

## Framework

Figure 2 shows a class diagram of the domain model that has the most important object classes according to P system specification.

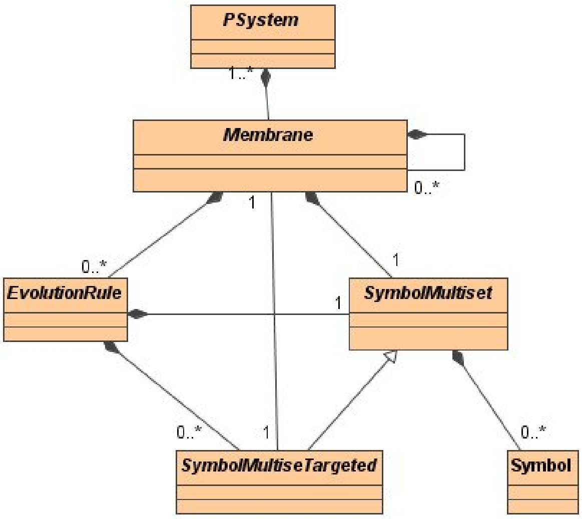


Figure 2: P system Domain Model Class Diagram.

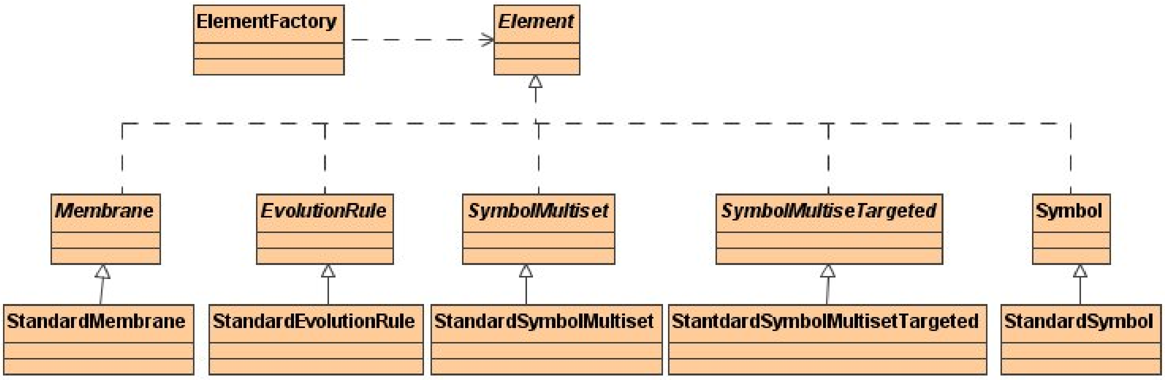


Figure 3: Data Structures Class Diagram.

Figure 3 shows the class diagram that was designed for covering requisites 1.1 and 2.1. This way, concrete classes in the third level of the inheritance hierarchy are contributed for every standard data structure. Also, it makes easy the incorporation of new data structures inheriting from the abstract classes of the second inheritance hierarchy level.

In particular, class *ElementFactory* is responsible of the configuration of the data structures for a given evolution of requisite 2.4.

Figure 4 shows the class diagram designed to cover requisites 1.2 and 2.2. This way, concrete classes in the fifth level of the inheritance hierarchy are contributed for every algorithm of evolution phases. Also, it makes easy the incorporation of new algorithms inheriting from forth level of inheritance hierarchy abstract classes.

In particular, class PhaseFactory is responsible of the configuration of the phases and algorithms for a concrete evolution of requisite 2.4.

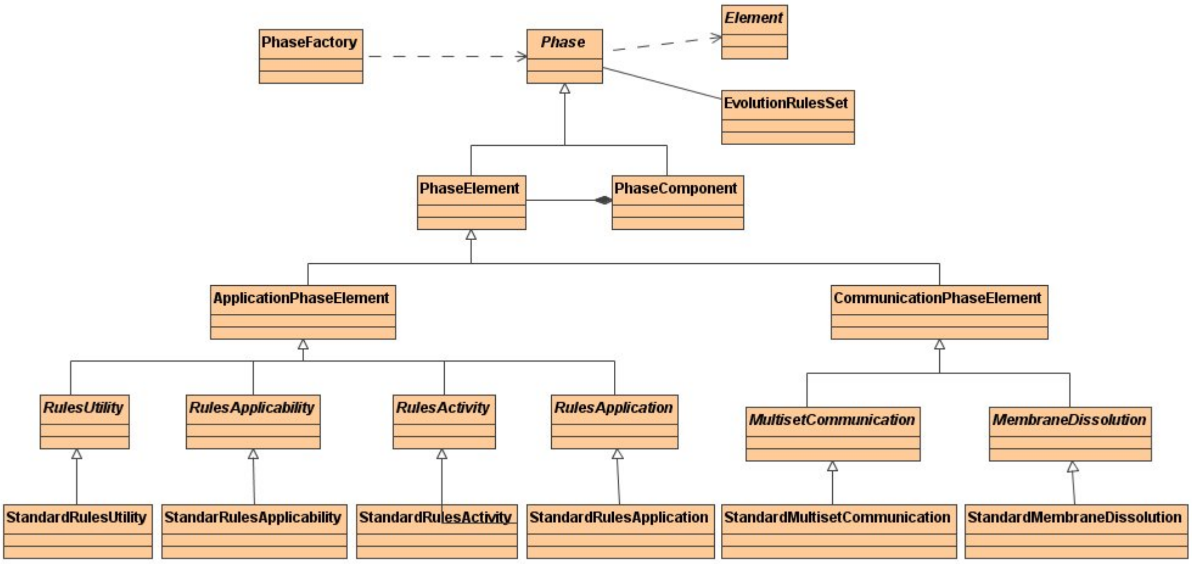


Figure 4: Algorithms Class Diagram.

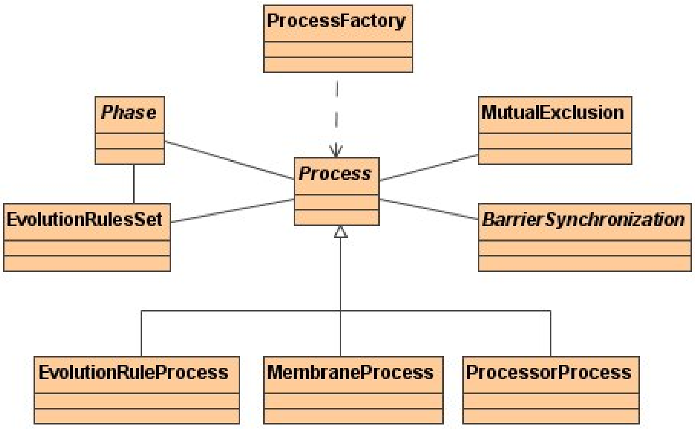


Figure 5: Software Architecture Class Diagram.

Figure 5 shows the class diagram designed to cover the requisite 1.3. This way, concrete classes in the second level of the inheritance hierarchy are contributed for every process architectures together with the classes for the process synchronization.

In particular, class *ProcessFactory* is responsible of the configuration of the process architectures for a concrete evolution of requisite 2.4.

Figure 6 shows the class diagram designed to cover requisites 1.4, 1.5, 1.6 and 2.3. This way, concrete classes in the forth level of inheritance hierarchy are contributed for the detection of errors and automatic monitorization of functional and non functional sets of tests respectively, and for the persistence, log and visualization of the results of a given evolution. Moreover, new functionalities (P system analysis, compilation, …) can be developed inheriting from *VisitorElement*.

In particular, class *VisitorFactory* is responsible of the configuration of a given evolution of the requisite 2.5.

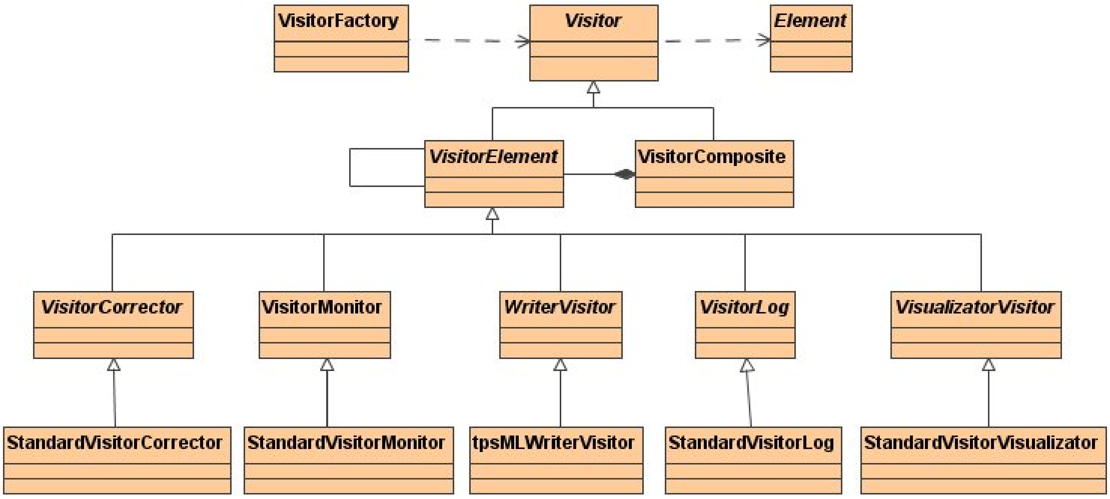


Figure 6: Visitors Class Diagram

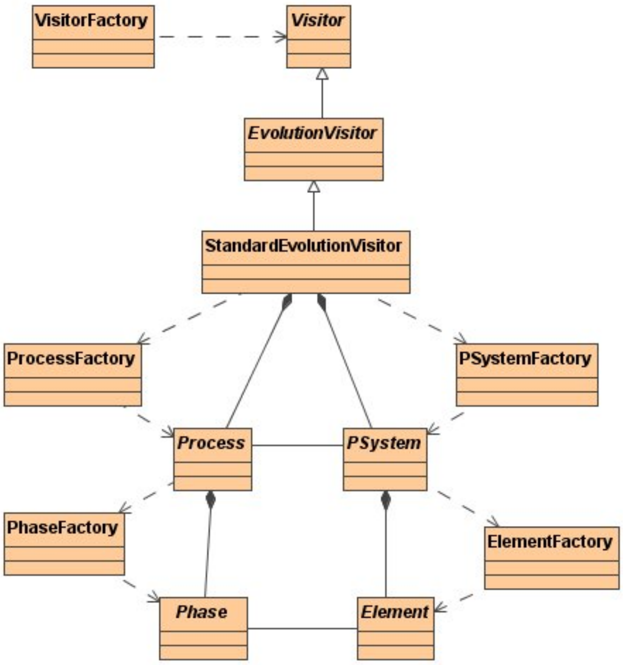


Figure 7: General Class Diagram.

Figure 7 shows the general class diagram that relates every class of previous class diagrams. In particular, class *PSystemFactory* is responsible of the edition, recovery and sets of parameterized tests automatic generation of requisite 1.7. Moreover, set of factory classes is responsible of managing the configurations for a given evolution of requisite 1.8.

## Conclusion

This work contributes a framework that makes investigation of developing new simulators and implementations of membrane systems easier. Its goal is to provide enough reusability and flexibility to get the researcher is concentrated in the goals of his investigation. This way, it is possible to reuse standard data structures and algorithms of the P system model, processes and synchronization management, error detection, monitorization and log for tests phase and the edition, recovery, automatic generation, persistence and visualization of P systems. On the other hand, the simple inheritance mechanism provides flexibility for the incorporation of new data structures and algorithms.

## Bibliography

[Alonso, 2007] S. Alonso, L. Fernández, F. Arroyo, J. Gil. *A Circuit Implementing Massive Parallelism in Transition P Systems*. Fifth International Conference Information Research and Applications (i.TECH-2007). Varna (Bulgary) june, 2007. (submitted).

[Arroyo, 2004a] F. Arroyo, C. Luengo, Castellanos, L.F. de Mingo. *A binary data structure for membrane processors: Connectivity Arrays*. Lecture Notes in Computer Science (A. Alhazov, C. Martin-Vide, G. Mauri, G. Paun, G.Rozenberg, A. Saloma, eds.) Springer Verlag (2933), 2004. 19-30.

[Arroyo, 2004b] F. Arroyo, C. Luengo, Castellanos, L.F. de Mingo. *Representing Multisets and Evolution Rules in Membrane Processors.* Preproceedings of the Fifth Workshop on Membrana Computing (WMC5). Milano (Italy) june, 2004. 126-137.

[Arroyo, 2004c] F. Arroyo, C. Luengo, L. Fernandez, L.F. de Mingo, J. Castellanos, *Simulating membrane systems in digital computers.* ITHEA-2004 International Journal Information Theories and Applications (vol.11 - 1) 2004. 29-34.

[Bravo, 2007a] G. Bravo, L. Fernández, F. Arroyo, J.A. Frutos. *A Hierarchical Architecture with parallel comunication for implementing P Systems.* Fifth International Conference Information Research and Applications (i.TECH-2007). Varna (Bulgary) june, 2007. (submitted).

[Bravo, 2007b] G. Bravo, L. Fernández, F. Arroyo, J. Tejedor. *Master/Slave Parallel Architecture for Implementing P Systems.* The 8th WSEAS International Conference on Mathematics and Computers in Business and Economics (MCBE'07). Vancouver (Canada) june, 2007. (submitted).

[Ciobanu, 2004] G.Ciobanu, W.Guo . *P Systems Running on a Cluster of Computers*. Workshop on Membrane Computing (Gh. Păun, G. Rozenberg, A. Salomaa Eds.), LNCS 2933, Springer, 123-139.

[Ciobanu, 2006] G. Ciobanu, M. Pérez-Jiménez, Gh. Păun. *Applications of Membrana Computing*". Natural Computing Series, Springer Verlag, october, 2006.

[Fernández, 2005a] L. Fernández, F. Arroyo, J. Castellanos, V.J. Martinez, L.F. Mingo. *Software Tools/ P System Simulators Interoperability.* (R. Freund, G. Lojka, M. Oswald, Gh. Paun, eds.) Preproceedings of Sixth International Workshop on Membrane Computing (WMC6), Vienna, (Austria) june, 2005. 147-161.

[Fernández, 2005b] L. Fernandez, V.J. Martinez, F. Arroyo, L.F. Mingo. *A Hardware Circuit for Selecting Active Rules in Transition P Systems.* (G. Ciobanu, Gh. Paun, eds.) Preproceedings of First International Workshop on Theory and Application of P Systems, Timisoara (Romania), september, 2005. 45-48.

[Fernández, 2006a] L. Fernández, F. Arroyo, I. García, G. Bravo. *Decision Trees for Applicability of Evolution Rules in Transition P System*. ITHEA-2006 Interantional Journal Information Theories and Applications (vol.11 - 1) 2006. 29-34.

[Fernández, 2006b] L. Fernández, F. Arroyo, J.A. Tejedor, J. Castellanos*, Massively Parallel Algorithm for Evolution Rules Application in Transition P Systems*. Preproceedings of Membrane Computing, International Workshop (WMC7), Leiden (The Netherlands) july, 2006. 337-343.

[Fernández, 2006c] L. Fernández, F. Arroyo, J. Castellanos, J.A. Tejedor, I. García, *New Algorithms for Application of Evolution Rules based on Applicability Benchmarks*. International Conference on Bioinformatics and Computational Biology(BIOCOMP06), Las Vegas (EEUU), july, 2006.

[Fernández, 2007] L. Fernández, F. Arroyo, I. Garcia, A. Gutierrez. *Parallel software architectures analysis for implementing P systems.* (M. Sugisaka, H. Tanaka, eds.), Proceedings of the 12th Int. Symposium on Artificial Life and Robotics (AROB07), Beppu (Japan) january, 2007. 494-499.

[Frutos, 2007] J.A.Frutos, L. Fernández, F.Arroyo, G.Bravo. *Static Analysis of Usefulness States in Transition P Systems.* Fifth International Conference Information Research and Applications (i.TECH-2007). Varna (Bulgary) june, 2007. (submitted).

[Gil, 2007] F.J. Gil, L. Fernández, F. Arroyo, J.A. Tejedor. *Delimited Massively Parallel Algorithm based on Rules Elimination for Application of Active Rules in Transition P Systems.* Fifth International Conference Information Research and Applications (i.TECH-2007). Varna (Bulgary) june, 2007. (submitted).

[Gutiérrez, 2006] A. Gutiérrez, L. Fernández, F. Arroyo, V. Martínez. *Design of a Hardware Architecture based on Microcontrollers for the Implementation of Membrane Systems*. Proceedings on 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC-2006). Timisoara (Romania) septiembre, 2006. 39-42.

[Gutiérrez, 2007a] A. Gutiérrez, L. Fernández, F. Arroyo, S. Alonso. *Hardware and Software Architecture for Implementing Membrane Systems: A case of study to Transition P Systems.* The International Meeting on DNA Computing (DNA13), Memphis (USA) June 3-8, 2007. (submitted)..

[Gutiérrez, 2007b] A. Gutiérrez, L. Fernández, F. Arroyo, G. Bravo. *Compression of Multisets and Evolution Rules Optimizing the Storage and Communication in Membrana System.* Eight Workshop on Membrane Computing (WMC8). Thessaloniki (Greece) june, 2007 (submitted).

[Martínez, 2006a] V. Martinez, L. Fernández, F. Arroyo, I. García, A. Gutierrez. *A HW circuit for the application of Active Rules in a Transition P System Region*. Proceedings on Fourth International Conference Information Research and Applications (i.TECH-2006). Varna (Bulgary) June, 2006. pp. 147-154. ISBN-10: 954-16-0036-0.

[Martínez, 2006b] V. Martínez, L. Fernández, F. Arroyo, A. Gutierrez. *HW Implementation of a Bounded Algorithm for Application of Rules in a Transition P-System*. Proceedings on 8th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC-2006). Timisoara (Romania) septiembre, 2006. pp. 32-38.

[Păun, 1999] Gh.Paun. *Computing with membranes*. An introduction. Bulletin of the EATCS, 67, 139-152, 1999.

[Păun, 2005] PGh. Păun. *Membrane computing. Basic ideas, results, applications.* Pre-Proceedings of First International Workshop on Theory and Application of P Systems. Timisoara (Romania), september , 2005. 1-8.

[Petreska, 2003] B. Petreska and C. Teuscher. *A hardware membrane system*. A. Alhazov, C. Martin-Vide, Gh. Paun (eds.):Pre-proceedings of the workshop on Membrane Computing Tarragona, July 17-22 2003, 343-355.

[Syropoulos, 2003] A. Syropoulos, E.G. Mamatas, P.C. Allilomes. *A distributed simulation of P systems.*(A. Alhazov, C. Martin-Vide and Gh. Păun, eds.) Preproceedings of the Workshop on Membrana Computing. Tarragona (Spain), july, 2003, 455-460.

[Tejedor, 2007a] A. Tejedor, L. Fernández, F. Arroyo, G. Bravo, *An architecture for attacking the bottleneck communication in P systems*. In: M. Sugisaka, H. Tanaka (eds.), Proceedings of the 12th Int. Symposium on Artificial Life and Robotics, Jan 25-27, 2007, Beppu, Oita, Japan, 500-505.

[Tejedor, 2007b] A. Tejedor, L. Fernández, F. Arroyo, A. Gutierrez. *Algorithm of Active Rule Elimination for Application of Evolution Rules.* The 8th WSEAS International Conference on Mathematics and Computers in Business and Economics (MCBE'07). Vancouver (Canada) june, 2007. (submitted).

[Tejedor, 2007c] A. Tejedor, L. Fernández, F. Arroyo, S. Gómez. *Application Algorithm based on Evolution Rules Competitivity.* Eight Workshop on Membrane Computing (WMC8). Thessaloniki (Greece) june, 2007 (submitted).

## Authors' Information

***Sandra Gómez Cannaval –*** *Natural Computing Group. Dpto. Organización y Estructura de la Información de la Escuela Universitaria de Informática de la Universidad Politécnica de Madrid, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: sgomez@eui.upm.es*

***Luis Fernández Muñoz –*** *Natural Computing Group. Dpto. Lenguajes, Proyectos y Sistemas Informáticos de la Escuela Universitaria de Informática de la Universidad Politécnica de Madrid, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: setillo@eui.upm.es*

***Iván García –*** *Natural Computing Group. Dpto. Lenguajes, Proyectos y Sistemas Informáticos de la Escuela Universitaria de Informática de la Universidad Politécnica de Madrid, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: igarcia@eui.upm.es*

***Fernando Arroyo –*** *Natural Computing Group. Dpto. Lenguajes, Proyectos y Sistemas Informáticos de la Escuela Universitaria de Informática de la Universidad Politécnica de Madrid, Ctra. de Valencia, km. 7, 28031 Madrid (Spain); e-mail: farroyo@eui.upm.es*

# Neural Nets

## NETWORKS OF EVOLUTIONARY PROCESSORS (NEP) AS DECISION SUPPORT SYSTEMS

**Miguel Angel Díaz, Nuria Gómez Blas, Eugenio Santos Menéndez, Rafael Gonzalo, Francisco Gisbert**

***Abstract**: This paper presents the application of Networks of Evolutionary Processors to Decision Support Systems, precisely Knowledge-Driven DSS. Symbolic information and rule-based behavior in Networks of Evolutionary Processors turn out to be a great tool to obtain decisions based on objects present in the network. The non-deterministic and massive parallel way of operation results in NP-problem solving in linear time. A working NEP example is shown.*

***Keywords**: Natural Computing, Networks of Evolutionary Processors, Decision Support Systems.*

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

### Introduction

There are many approaches to decision-making and because of the wide range of domains in which decisions are made; the concept of decision support system (DSS) is very broad. A DSS can take many different forms. In general, we can say that a DSS is a computerized system for helping make decisions. A decision is a choice between alternatives based on estimates of the values of those alternatives. Supporting a decision means helping people working alone or in a group gather intelligence, generate alternatives and make choices. Supporting the choice making process involves supporting the estimation, the evaluation and/or the comparison of alternatives. In practice, references to DSS are usually references to computer applications that perform such a supporting role [Alter, 1980].

Abbreviated DSS, the term refers to an interactive computerized system that gathers and presents data from a wide range of sources, typically for business purposes. DSS applications are systems and subsystems that help people make decisions based on data that is culled from a wide range of sources. For example: a national on-line book seller wants to begin selling its products internationally but first needs to determine if that will be a wise business decision. The vendor can use a DSS to gather information from its own resources (using a tool such as OLAP) to determine if the company has the ability or potential ability to expand its business and also from external resources, such as industry data, to determine if there is indeed a demand to meet. The DSS will collect and analyze the data and then present it in a way that can be interpreted by humans. Some decision support systems come very close to acting as artificial intelligence agents.

DSS applications are not single information resources, such as a database or a program that graphically represents sales figures, but the combination of integrated resources working together.

Using the mode of assistance as the criterion [Power, 2002] differentiates communication-driven DSS, data-driven DSS, document-driven DSS, knowledge-driven DSS, and model-driven DSS.

- A model-driven DSS emphasizes access to and manipulation of a statistical, financial, optimization, or simulation model. Model-driven DSS use data and parameters provided by users to assist decision makers in analyzing a situation; they are not necessarily data intensive. Dicodess is an example of an open source model-driven DSS generator [Gachet, 2004].
- A communication-driven DSS supports more than one person working on a shared task; examples include integrated tools like Microsoft's NetMeeting or Groove [Stanhope, 2002].
- A data-driven DSS or data-oriented DSS emphasizes access to and manipulation of a time series of internal company data and, sometimes, external data.
- A document-driven DSS manages, retrieves and manipulates unstructured information in a variety of electronic formats.
- A knowledge-driven DSS provides specialized problem solving expertise stored as facts, rules, procedures, or in similar structures.

By incorporating AI techniques in a decision support system, we make that DSS artificially intelligent - capable of displaying behavior that would be regarded as intelligent if observed in humans. Artificially intelligent DSSs are becoming increasingly common. Perhaps the most prominent of these are expert systems, which support decision making by giving advice comparable to what human experts would provide.

This paper is focused on knowledge-driven DSS using Artificial Intelligence models. Networks of Evolutionary Processors have rules, facts, and collaboration among processors to generate a final decision based on evolutionary steps that take place in processors. Next section describes the computational model of Networks of Evolutionary Processors. And finally, an example is shown.

## Networks of Evolutionary Processors

A network of evolutionary processors of size n is a construct NEP = $(V, N_1, N_2, \dots, N_n, G)$, where V is an alphabet and for each $1 \leq i \leq n$, $N_i = (M_i, A_i, PI_i, PO_i)$ is the i-th evolutionary node processor of the network. The parameters of every processor are:

- $M_i$ is a finite set of evolution rules of one of the following forms only:

  $a \prod b$, $a, b \in V$ (substitution rules)

  $a \prod \varepsilon$, $a \in V$ (deletion rules)

  $\varepsilon \prod a$, $a \in V$ (insertion rules)

  More clearly, the set of evolution rules of any processor contains either substitution or deletion or insertion rules.
- $A_i$ is a finite set of strings over V. The set $A_i$ is the set of initial strings in the i-th node. Actually, in what follows, we consider that each string appearing in any node at any step has an arbitrarily large number of copies in that node, so that we shall identify multisets by their supports.
- $PI_i$ and $PO_i$ are subsets of $V^*$representing the input and the output filter, respectively. These filters are defined by the membership condition, namely a string $w \in V^*$ can pass the input filter (the output filter) if $w \in PI_i$ ($w \in PO_i$).

G = $(N_1, N_2, \dots, N_n, E)$ is an undirected graph called the underlying graph of the network [Paun, 2002] [Paun, 2000]. The edges of G, that is the elements of E, are given in the form of sets of two nodes. Kn denotes the complete graph with n vertices. By a configuration (state) of an NEP as above we mean an n-tuple C = $(L_1, L_2, \dots,$

$L_n$)$, with $L_i \subseteq V^*$ for all $1 \leq i \leq n$. A configuration represents the sets of strings (remember that each string appears in an arbitrarily large number of copies) which are present in any node at a given moment; clearly the initial configuration of the network is $C_0 = (A_1, A_2, \ldots , A_n)$.

A configuration can change either by an evolutionary step or by a communicating step. When changing by an evolutionary step, each component $L_i$ of the configuration is changed in accordance with the evolutionary rules associated with the node i. When changing by a communication step, each node processor $N_i$ sends all copies of the strings it has which are able to pass its output filter to all the node processors connected to $N_i$ and receives all copies of the strings sent by any node processor connected with $N_i$ providing that they can pass its input filter [Manea, 2006] [Martin, 2005] [Garey, 1979].

**Theorem 1.** A complete NEP of size 5 can generate each recursively enumerable language.

**Theorem 2.** A star NEP of size 5 can generate each recursively enumerable language.

**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.

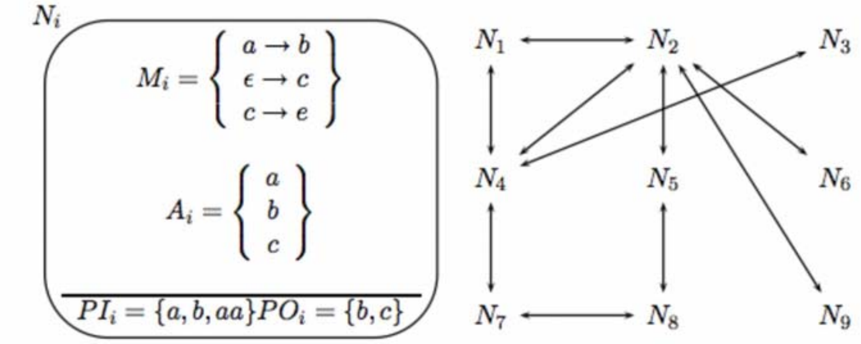

$N_i$
$M_i = \left\{ \begin{array}{c} a \to b \\ \epsilon \to c \\ c \to e \end{array} \right\}$
$A_i = \left\{ \begin{array}{c} a \\ b \\ c \end{array} \right\}$
$PI_i = \{a, b, aa\} PO_i = \{b, c\}$
$N_1$
$N_2$
$N_3$
$N_4$
$N_5$
$N_6$
$N_7$
$N_8$
$N_9$

**Figure 1.-** A sample Network of Evolutionary Processors

## An Example

Opportunities for building business expert systems abound for both small and large problems. In each case, the expert system is built by developing its rule set. The planning that precedes rule set development is much like the planning that would precede any project of comparable magnitude within the organization. The development process itself follows an evolutionary spiral composed of development cycles.

Each cycle picks up where the last ended, building on the prior rule set. For a developer, the spiral represents a continuing education process in which more and more of an expert's reasoning knowledge is discovered and formalized in the rule set. Here, each development cycle was presented in terms of seven consecutive stages. Other characterizations of a development cycle (involving different stages or sequences) may be equally valuable. Many aspects of traditional systems analysis and project management can be applied to the development of expert systems.

Rule set development is a process of discovery and documentation. Research continues in search of ways of automating various aspects of the process. It would not be surprising to eventually see expert systems that can assist in this process -- that is, an expert system that "picks the mind" of a human expert in order to build new expert systems. Until that time comes, the topics discussed in this chapter should serve as reminders to developers of expert decision support systems about issues to consider during the development process.

As it stands, rule-based systems are the most widely used and accepted AI in the world outside of games. The fields of medicine, finance and many others have benefited greatly by intelligent use of such systems. With the combination of rule-based systems and ID trees, there is great potential for most fields. Here it is an example of a rule-base expert system.

**Table 1.-** Rule-based decision support system applied to medical diagnosis (snapshoot)

| | |
|---|---|
| **Assertions**<br>A1: runny nose<br>A2: temperature=101.7<br>A3: headache<br>A4: cough<br><br>**Rules**<br>R1: if (nasal congestion)<br>(viremia)<br>then diagnose (influenza)<br>exit<br><br>R2: if (runny nose)<br>then assert (nasal congestion) | R3: if (body- aches)<br>then assert (achiness)<br><br>R4: if (temp >100)<br>then assert (fever)<br><br>R5: if (headache)<br>then assert (achiness)<br><br>R6: if (fever)<br>(achiness)<br>(cough)<br>then assert (viremia) |

It is necessary to define the underlying graph In order to simulate previous example with a Network of Evolutionary Processors. First of all, an initial processor containing the assertions and basic rules will forward important information to a second processor, which is in charge of a given disease, and forward its result to a container processor. This process can be shown in figure 2.

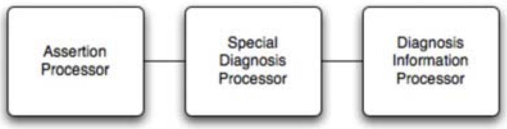


**Figure 2.-** Network of Evolutionary Processors Architecture sample

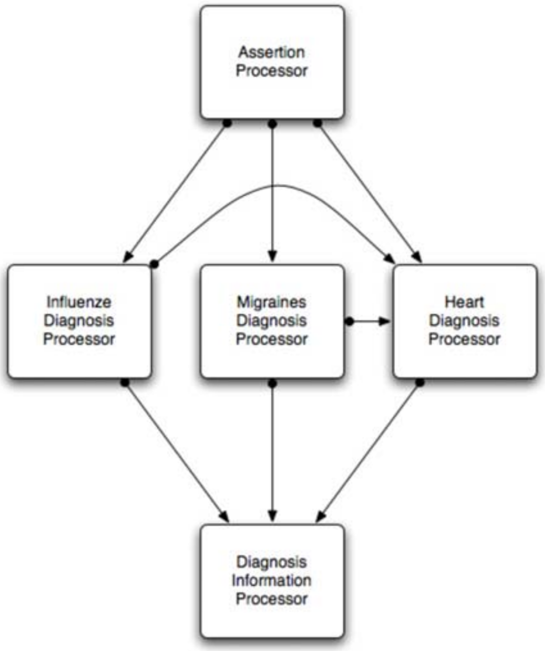

Assertion Processor
Influenze Diagnosis Processor
Migraines Diagnosis Processor
Heart Diagnosis Processor
Diagnosis Information Processor

**Figure 3.-** Network of Evolutionary Processors Architecture for medical diagnosis

Previous simple example can be extended to a more general diagnosis system as detailed in figure 3. There exists one processor or even more processors in charge of a located diagnosis problem such as: influenza, migraines, heart diseases, etc… These local diagnosis processors can communicate each other to auto complete information diagnosis. Finally, each result of diagnosis processors is sent to an information processor that can combine multiple diagnoses or just show them.

### Configuration of a Network of Evolutionary Processors

Next table shows an XML file with the NEP initial configuration corresponding to the medical diagnosis shown in table 1. There are three processors (see figure 2): assertion process (name = 0), specific diagnosis processor (name = 1) and diagnosis information processor (name = 2). Objects travel trough these processors until a final diagnosis is present in the last one. Obviously, this is a simple example, but according to figure 3 the NEP architecture could be complicated in order to obtain a more sophisticated diagnosis. Main idea is to put some assertions in one or more processors and then let them evolve using evolution steps (rules application) and communication steps.

This configuration file is parsed into JAVA objects and a separate thread for each processor is created; also each rule and filter are coded as threads in order to keep the massive parallelization defined in the theoretical model of Networks of Evolutionary Processors.

**Table 2.-** NEP initial configuration corresponding to table 1

```
<?xml version="1.0"?>
<NEP>
   <processor>
      <name>0</name>
      <object>runny nose</object>
      <object>high temperature</object>
      <object>headache</object>
      <object>cough</object>
      <rule>
         <antecedent>
            <object>runny nose</object>
         </antecedent>
         <consequent>
            <object>nasal congestion</object>
         </consequent>
      </rule>
      <rule>
         <antecedent>
            <object>body-aches</object>
         </antecedent>
         <consequent>
            <object>achiness</object>
         </consequent>
      </rule>
      <rule>
         <antecedent>
            <object>high temperature</object>
         </antecedent>
         <consequent>
            <object>fever</object>
         </consequent>
      </rule>
      <rule>
         <antecedent>
            <object>headache</object>
         </antecedent>
         <consequent>
            <object>achiness</object>
         </consequent>
      </rule>
      <inputfilter>
      </inputfilter>
      <outputfilter>
         <object>nasal congestion</object>
         <object>fever</object>
         <object>achiness</object>
         <object>cough</object>
      </outputfilter>
   </processor>
   <processor>
      <name>1</name>
      <rule>
         <antecedent>
            <object>nasal congestion</object>
            <object>viremia</object>
         </antecedent>
         <consequent>
            <object>influenza</object>
         </consequent>
      </rule>
      <rule>
         <antecedent>
            <object>fever</object>
            <object>achiness</object>
            <object>cough</object>
         </antecedent>
         <consequent>
            <object>viremia</object>
         </consequent>
      </rule>
      <inputfilter>
         <object>nasal congestion</object>
         <object>fever</object>
         <object>achiness</object>
         <object>cough</object>
      </inputfilter>
      <outputfilter>
         <object>influenza</object>
      </outputfilter>
   </processor>
   <processor>
      <name>2</name>
      <inputfilter>
         <object>influenza</object>
      </inputfilter>
      <outputfilter>
      </outputfilter>
   </processor>
   <conn>
      <from>0</from>
      <to>1</to>
   </conn>
   <conn>
      <from>1</from>
      <to>2</to>
   </conn>
</NEP>
```

Note that connections among processors are defined in a unidirectional way, but any type of connection can be expressed in the XML configuration file in order to have a more complex underlying graph in the network architecture.

### Simulation Results

Results concerning simulation of NEP configuration in table 2 can be seen in table 3. `Processor 2:3` , which is the output processor, has the object `influenza`, desired result. This object is generated in `Processor 1:2` using rules inside it. NEP behavior is totally non-deterministic since rules, filters and processors run together in parallel. This example only uses substitution rules, neither insertion nor deletion rules are coded.

**Table 3.-** Final configuration of NEP corresponding to described configuration on table 2

```
[----------
Processor 0 : 1
Rules: [[runny nose] --> [nasal congestion], [body-aches] --> [achiness], [high temperature] -
-> [fever], [headache] --> [achiness]]
Objects: [runny nose, high temperature, headache]
Output Filter: [nasal congestion, fever, achiness, cough]
Input Filter: []
----------
, ----------
Processor 1 : 2
Rules: [[nasal congestion, viremia] --> [influenza], [fever, achiness, cough] --> [viremia]]
Objects: [cough, fever, achiness, viremia, nasal congestion, influenza]
Output Filter: [influenza]
Input Filter: [nasal congestion, fever, achiness, cough]
----------
, ----------
Processor 2 : 3
Rules: []
Objects: [influenza]
Output Filter: []
Input Filter: [influenza]
----------
]
```

The great disadvantage is that a given NEP can only solve a given problem; if it is necessary to solve another problem (maybe a little variation) then another different NEP has to be implemented. The idea of learning tries to undertake such disadvantage proposing a model able to solve different kinds of problems (that is a general class of problems). Learning can be based on the self-organizing maps. There are a lot of open problems that need to be solved in order to show the computational power of this learning idea, but the possibility to compute NP-problems is promising apart from the massive parallelization and non-determinism of the model.

## Conclusion

This paper has introduced the computational paradigm Networks of Evolutionary Processors. NEPs can be easily applied to Knowlede-driven Decision Support Systems due to the inherent rule-based behavior of NEPs. JAVA implementation of this model works as defined by the theoretical background of NEPs: massive parallelization and non-deterministic behavior.

Connectionists' models such as Neural Networks can be taken into account to develop NEP architecture in order to improve behavior. As a future research, learning concepts in neural networks can be adapted in a NEP architecture provided the numeric-symbolic difference in both models. NEPs can be considered universal models since they are able to solve NP-problems.

## Bibliography

[Alter, 1980] Alter, S. L. Decision support systems: current practice and continuing challenges. Reading, Mass., Addison-Wesley Pub. (1980).

[Gachet, 2004] Gachet, A. Building Model-Driven Decision Support Systems with Dicodess. Zurich, VDF. (2004).

[Garey, 1979] M. Garey and D. Johnson. Computers and Intractability. A Guide to the Theory of NP-completeness. Freeman, San Francisco, CA, (1979).

[Manea, 2006] F. Manea, C. Martın-Vide, and V. Mitrana. All NP-problems can be solved in polynomial time by accepting networks of splicing processors of constant size. Proc. of DNA 12, in press. (2006).

[Martin, 2005] C. Martin-Vide and V. Mitrana. Networks of evolutionary processors: Results and perspectives. Molecular Computational Models: Unconventional Approaches. 78–114, Idea Group Publishing, Hershey. (2005).

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

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

[Power, 2002] Power, D. J. Decision support systems: concepts and resources for managers. Westport, Conn., Quorum Books. (2002).

[Stanhope, 2002] Stanhope, P. Get in the Groove: building tools and peer-to-peer solutions with the Groove platform. New York, Hungry Minds. (2002).

## Authors' Information

***Miguel Angel Díaz** – 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: mdiaz@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*

***Eugenio Santos Menéndez** – 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: esantos@eui.upm.es*

***Rafael Gonzalo** – Dept. Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, 28660 Madrid, Spain; e-mail: rgonzalo@fi.upm.es*

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

# NETWORKS OF EVOLUTIONARY PROCESSORS: JAVA IMPLEMENTATION OF A THREADED PROCESSOR

## Miguel Angel Díaz, Luis Fernando de Mingo López, Nuria Gómez Blas

***Abstract**: This paper is focused on a parallel JAVA implementation of a processor defined in a Network of Evolutionary Processors. Processor description is based on JDom, which provide a complete, Java-based solution for accessing, manipulating, and outputting XML data from Java code. Communication among different processor to obtain a fully functional simulation of a Network of Evolutionary Processors will be treated in future. A safe-thread model of processors performs all parallel operations such as rules and filters. A non-deterministic behavior of processors is achieved with a thread for each rule and for each filter (input and output). Different results of a processor evolution are shown.*

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

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

## Introduction

A network of evolutionary processors of size n is a construct NEP = $(V, N_1, N_2,\ldots, N_n, G)$, where V is an alphabet and for each $1 \le i \le n$, $N_i = (M_i, A_i, PI_i, PO_i)$ is the i-th evolutionary node processor of the network. The parameters of every processor are:

- $M_i$ is a finite set of evolution rules of one of the following forms only:

  $a \prod b$, $a, b \in V$ (substitution rules)

  $a \prod \varepsilon$, $a \in V$ (deletion rules)

  $\varepsilon \prod a$, $a \in V$ (insertion rules)

  More clearly, the set of evolution rules of any processor contains either substitution or deletion or insertion rules.

- $A_i$ is a finite set of strings over V. The set $A_i$ is the set of initial strings in the i-th node. Actually, in what follows, we consider that each string appearing in any node at any step has an arbitrarily large number of copies in that node, so that we shall identify multisets by their supports.

- $PI_i$ and $PO_i$ are subsets of $V^*$representing the input and the output filter, respectively. These filters are defined by the membership condition, namely a string $w \in V^*$ can pass the input filter (the output filter) if $w \in PI_i$ ($w \in PO_i$).

$G = (N_1, N_2,\ldots, N_n, E)$ is an undirected graph called the underlying graph of the network [Paun, 2002] [Paun, 2000]. The edges of G, that is the elements of E, are given in the form of sets of two nodes. Kn denotes the complete graph with n vertices. By a configuration (state) of an NEP as above we mean an n-tuple $C = (L_1, L_2,\ldots., L_n)\$$, with $L_i \subseteq V^*$ for all $1 \le i \le n$. A configuration represents the sets of strings (remember that each string appears in an arbitrarily large number of copies) which are present in any node at a given moment; clearly the initial configuration of the network is $C_0 = (A_1, A_2,\ldots, A_n)$.

A configuration can change either by an evolutionary step or by a communicating step. When changing by an evolutionary step, each component $L_i$ of the configuration is changed in accordance with the evolutionary rules associated with the node i. When changing by a communication step, each node processor $N_i$ sends all copies of the strings it has which are able to pass its output filter to all the node processors connected to $N_i$ and receives all copies of the strings sent by any node processor connected with $N_i$ providing that they can pass its input filter [Manea, 2006] [Martin, 2005] [Garey, 1979].

**Theorem 1.** A complete NEP of size 5 can generate each recursively enumerable language.

**Theorem 2.** A star NEP of size 5 can generate each recursively enumerable language.

**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.

Next sections deal with the JAVA implementation of a processor as the first step to achieve a fully functional simulation of NEPs. The non-deterministic behavior of NEPs must be taken into account, that is, a massive parallel implementation is reached having each rule in a thread and each filter in a thread. Objects in processor are locked to avoid mutual exclusion problems due to concurrent programming. [Fahlman, 1983] [Errico, 1994]

## JAVA Implementation

NEP processors must behave in a non-deterministic way. Configuration changes are outcome by a communication step or by an evolutionary step, but these two steps are accomplished with no order at all, that is, evolution or communication is chosen depending on the thread model of processor [Diaz, 2007]. Rules and filters (input and output) are implemented as threads extending `Runnable` interface. Therefore a processor is the parent of a set of threads, which use objects in processor in a mutual exclusion region.

Figure 1 shows an UML class diagram corresponding to all classes involved in the definition of a NEP processor. Rules, filters and objects are part of a processor. Filters can be either input filters or output filters, depending on their behavior, controlling how objects are sent or are received by different processors. Substitution rules have an antecedent and a consequent implemented as an object set. When a processor is run through the `start` method, it starts in a cascade way the rule threads and filter threads.

The whole system is prepared to a NEP implementation; only the communication classes must be coded in order to add the communication step to NEP since there exists methods to send a to receive objects in the processor class.

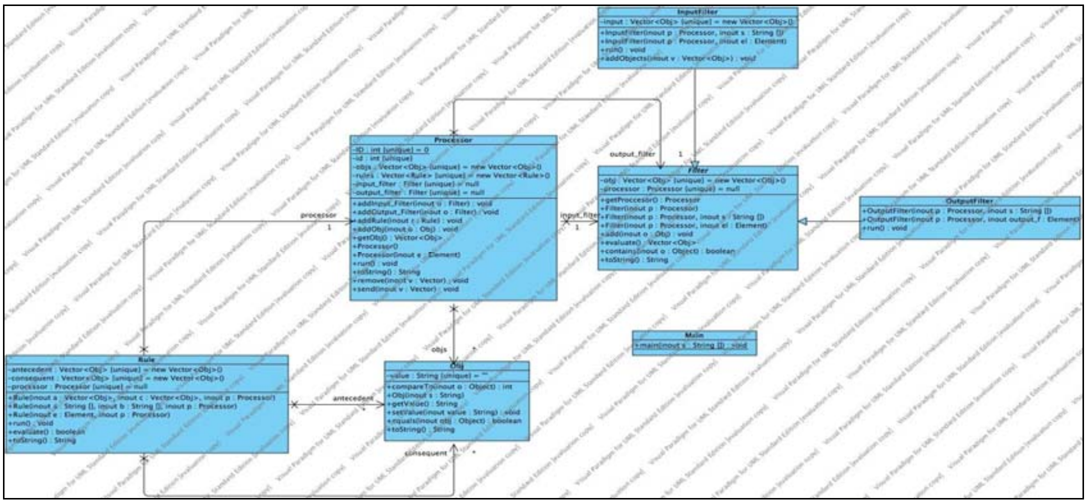

InputFilter
-input : Vector<Obj> {unique} = new Vector<Obj>()
+InputFilter(inout p : Processor, inout s : String [])
+InputFilter(inout p : Processor, inout el : Element)
+run() : void
+addObjects(inout v : Vector<Obj>) : void
Processor
-ID : int {unique} = 0
-id : int {unique}
-objs : Vector<Obj> {unique} = new Vector<Obj>()
-rules : Vector<Rule> {unique} = new Vector<Rule>()
-input_filter : Filter {unique} = null
-output_filter : Filter {unique} = null
+addInput_Filter(inout o : Filter) : void
+addOutput_Filter(inout o : Filter) : void
+addRule(inout r : Rule) : void
+addObj(inout o : Obj) : void
+getObj() : Vector<Obj>
+Processor()
+Processor(inout e : Element)
+run() : void
+toString() : String
+remove(inout v : Vector) : void
+send(inout v : Vector) : void
output_filter
1
Filter
-obj : Vector<Obj> {unique} = new Vector<Obj>()
-processor : Processor {unique} = null
+getProccesor() : Processor
+Filter(inout p : Processor)
+Filter(inout p : Processor, inout s : String [])
+Filter(inout p : Processor, inout el : Element)
+add(inout o : Obj) : void
+evaluate() : Vector<Obj>
+contains(inout o : Object) : boolean
+toString() : String
processor
1
input_filter
1
OutputFilter
+OutputFilter(inout p : Processor, inout s : String [])
+OutputFilter(inout p : Processor, inout output_f : Element)
+run() : void
Main
+main(inout s : String []) : void
objs
*
Rule
-antecedent : Vector<Obj> {unique} = new Vector<Obj>()
-consequent : Vector<Obj> {unique} = new Vector<Obj>()
-processor : Processor {unique} = null
+Rule(inout a : Vector<Obj>, inout c : Vector<Obj>, inout p : Processor)
+Rule(inout a : String [], inout b : String [], inout p : Processor)
+Rule(inout e : Element, inout p : Processor)
+run() : void
+evaluate() : boolean
+toString() : String
antecedent
*
Obj
-value : String {unique} = ""
+compareTo(inout o : Object) : int
+Obj(inout s : String)
+getValue() : String
+setValue(inout value : String) : void
+equals(inout obj : Object) : boolean
+toString() : String
consequent
*

**Figure 1.-** UML Class diagram of a NEP processor.

This is the basic composition of an evolutionary processor; nevertheless, there exist NEP architectures that have forbidden filters in the input and in the output. Differences in the implementation for the resolution of problems will be defined as types of the generic model for such given kind of problems.

## Processors

According to figure 1, each processor has a number of rules, objects and one input filter and one output filter. When the processor thread starts all rule threads are started and input/output threads to, see figure 2 and listing 1. Objects in processor are store using the `Vector` class that is thread-safe, so synchronization is guaranteed.

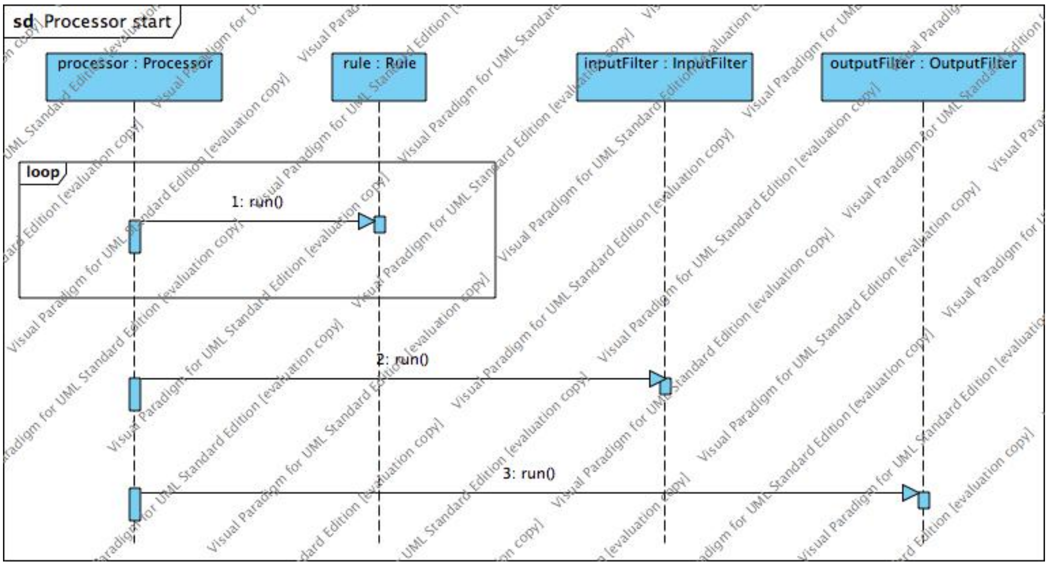


**Figure 2.-** UML Sequence diagram of processor.

```
public void run() {
     for (int i=0; i<this.rules.size(); i++)
      new Thread(this.rules.get(i)).start();
     new Thread((OutputFilter) this.output_filter).start();
     new Thread((InputFilter) this.input_filter).start();
     return;
}
```

**Listing 1.-** Processor behavior schema.

Processors can send and receive objects provided filter constraints are satisfied. This communication is perform in the following way:

- Sending objects. Output filter will check constraints and all objects that satisfy them will be removed from the object pool. In future they will be sent to processors connected to this one.

```
  public void run() {
    Vector v = null;
          while (true) {
                  v = super.evaluate();
                  super.getProccesor().remove(v);
                  super.send(v);
          }
          }
```

- Receiving objects. When the method `send` is invoked from another processor, some object will be located in the input filter pool to be checked, if it pass the constraints then it will be added to the processor if it is not present.

```
  synchronized public void send(Vector v) {
    ((InputFilter)this.input_filter).addObjects(v);
          }
```

Please note that all rules and filters are threaded, so the non-deterministic behavior is guaranteed.

## Rules

Rule threads are quite simple, they only check if the antecedent objects are in the processor object pool, if so objects in consequent are incorporated in processor pool, see figure 3. There is no order when applying rules, they all in separated threads, therefore they all check object pool at the "same time" and they will be applied at the "same time". Some random delay has been incorporated in order to make a more realistic simulation. There are no insertion and deletion rules in our system, but they can be easily added just defining a `null` object in the configuration file `config.xml` and the system will work in such NEP schema.

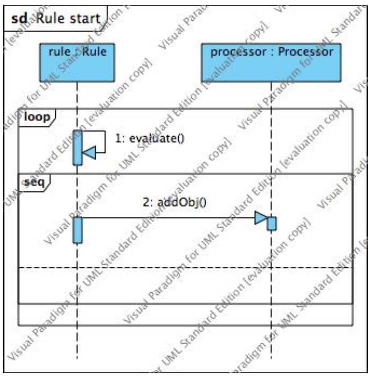


**Figure 3.-** UML Sequence diagram of rules.

Listing 2 shows an outline of the rule behavior according to figure 3. It can be noted that if the antecedent is satisfactory evaluated then all consequent objects will be added to the processor pool.

```
while(true) {
   if (this.evaluate()) {
     Enumeration<Obj> e = this.consequent.elements();
     while (e.hasMoreElements()) {
      Obj o = e.nextElement();
      if (!processor.getObj().contains(o)) this.processor.addObj(o);
     }
   }
}
```

**Listing 2.-** Rule behavior schema.

### Filters

Filters are implemented as threads. Therefore they run in parallel with rules. When a filter is satisfied then it will remove or add some objects to the processor pool. Main difference between input and output filters is that, see figure 4:

- Input filter just add objects to processor if they pass the constraints.
- Output filter evaluate object pool to guess which objects must be sent out.

Both of them use the functionality of a `Filter` class, which provides the evaluation of objects, see listing 3.

Several filters can be implemented in the evolutionary processors. A filter is a system that allows a symbol to go from one processor to another. Normally, the detection system is to compare a symbol with another one. Among possible filters that an evolutionary processor can have, most common filters are the PI or input filters, and the PO or output filters. A processor can have several filters of each type. This is the basic composition of an evolutionary processor, nevertheless, there exist NEP architectures that have forbidden filters in the input and in the output. Differences in the implementation for the resolution of problems will be defined as types of the generic model for such given kind of problems.

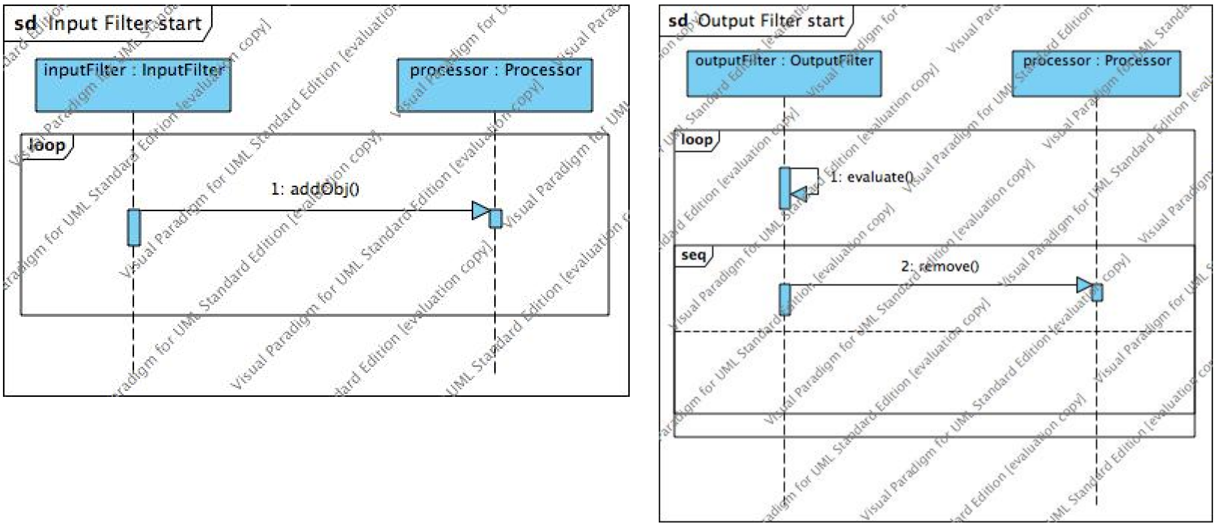


**Figure 4.-** UML Sequence diagram of Input and Output filters.

Listing 3 shows evaluation method corresponding to output filter to check if objects in pool must be sent out or not to connected processors, if so they will be removed from processor (see sending objects in processor subsection).

```java
public Vector<Obj> evaluate() {
   Vector<Obj> v = this.processor.getObj();
   Enumeration<Obj> e = v.elements();
   v = new Vector<Obj>();
   while(e.hasMoreElements()) {
     Obj o = e.nextElement();
     if (this.obj.contains(o)) if (!v.contains(o)) v.add(o);
   }
   return v;
}
```

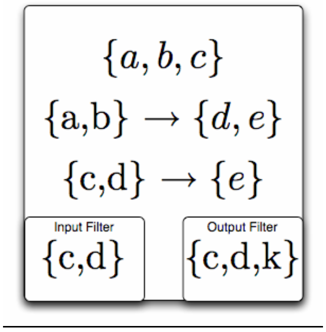
**Listing 3.-** Filter evaluation using object pool.

## Configuration

Configuration of a given processor is done with an XML file. In order to parse such configuration JDOM technology is used. There is no compelling reason for a Java API to manipulate XML to be complex, tricky, unintuitive, or a pain in the neck. JDOM is both Java-centric and Java-optimized. It behaves like Java, it uses Java collections, it is completely natural API for current Java developers, and it provides a low-cost entry point for using XML. While JDOM interoperates well with existing standards such as the Simple API for XML (SAX) and the Document Object Model (DOM), it is not an abstraction layer or enhancement to those APIs. Rather, it provides a robust, lightweight means of reading and writing XML data without the complex and memory-consumptive options that current API offerings provide.

Table 1 shows a processor with its corresponding XML configuration. Elements in such XML are: `processor`, `object`, `rule`, `antecedent`, `consequent`, `inputfilter` and `outputfilter`.

**Table 1.-** XML Configuration corresponding to sample processor in figure.



```xml
<NEP>
  <processor>
    <object>a</object>
    <object>b</object>
    <object>c</object>
    <rule>
      <antecedent>
        <object>a</object>
        <object>b</object>
      </antecedent>
      <consequent>
        <object>d</object>
        <object>e</object>
      </consequent>
    </rule>
    <rule>
     <antecedent>
          <object>c</object>
        <object>d</object>
      </antecedent>
      <consequent>
        <object>e</object>
      </consequent>
    </rule>
    <inputfilter>
      <object>c</object>
      <object>d</object>
    </inputfilter>
    <outputfilter>
      <object>c</object>
      <object>d</object>
      <object>k</object>
    </outputfilter>
  </processor>
</NEP>
```

This configuration file is extensible to a NEP system just adding as many processors as needed and adding a XML communication element to define the underlying graph in NEP architecture. As mentioned before, processors have methods to send and to receive objects and only the finalization condition must be taken into account to obtain a fully functional NEP system.

## Results: Non-deterministic behavior

This section shows results of evolution corresponding to processor in table 1. When the system starts some threads are create:

- One thread of rule `[a, b] --> [d, e]`
- One thread of rule `[c, d] --> [e]`
- One thread for output filter `[c, d, k]`
- One thread for input filter `[c, d]`

All four threads have access to objects `[a, b, c]` in processor. Depending on which thread access the object set it will be modified with new objects (rules) or some objects will be deleted (output filter) or some objects will be added (input filter). The input filter thread controls when new objects are sent from other processor to this one, this case is not yet implemented but it will be in future, when the NEP system will work as a whole not only isolated processors.

**Table 2.-** Initial and Final configuration of processor in table 1 (one possible evolution).

| Initial Configuration | Final Configuration |
|---|---|
| ---------- | ---------- |
| Processor 1 | Processor 1 |
| Rules: [[a, b] --> [d, e], | Rules: [[a, b] --> [d, e], |
| [c, d] --> [e]] | [c, d] --> [e]] |
| Objects: [a, b, c] | Objects: **[a, b, e]** |
| Output Filter: [c, d, k] | Output Filter: [c, d, k] |
| Input Filter: [c, d] | Input Filter: [c, d] |
| ---------- | ---------- |

Another execution of same processor outputs results in table 3. Please note objects in final configuration, they are different to those in table 2. This is due to the fact that first rule produces objects d and e but output filter was not activated (the processor stops). In table 2, object d is sent out by output filter.

**Table 3.-** Initial and Final configuration of processor in table 1 (another possible evolution).

| Initial Configuration | Final Configuration |
|---|---|
| ---------- | ---------- |
| Processor 1 | Processor 1 |
| Rules: [[a, b] --> [d, e], | Rules: [[a, b] --> [d, e], |
| [c, d] --> [e]] | [c, d] --> [e]] |
| Objects: [a, b, c] | Objects: **[a, b, e, d]** |
| Output Filter: [c, d, k] | Output Filter: [c, d, k] |
| Input Filter: [c, d] | Input Filter: [c, d] |
| ---------- | ---------- |

If this processor receives an object it will take part of object set. With this implementation the communication and evolution steps have a non-deterministic way, that is, both steps have no priority. In some cases the communication will take place before the evolution and in other cases evolution will take place before communication.

## Conclusion and Future Work

This paper has introduced the novel computational paradigm Networks of Evolutionary Processors that is able to solve NP-problems in linear time. The implementation of such model in a traditional computer is being performed and this paper shows an UML architecture. This architecture is a generic representation of a NEP processor behavior. The non-deterministic behavior is performed using JAVA threads accessing the object pool in processor; depending on the Java Virtual Machine a thread will run faster than another. Tables 2 and 3 show such non-deterministic behavior on a given processor.

Future work includes the simulation of a NEP system; only communication must be added to presented model. Such communication will be expressed in the XML configuration file. A separate thread for each processor will be created in final model together with a communication matrix to open communication channels with other processors.

## Bibliography

[Diaz, 2007] Miguel Angel Diaz, Miguel Angel Peña, and Luis F. de Mingo: Simulation of Networks of Evolutionary Processors with Filtered Connections. WESAS Transactions on Information, Science and Applications. Issue 3, Vol. 4. ISSN: 1709-0832. Pp.: 608-616. (2007).

[Errico, 1994] L. Errico and C. Jesshope. Towards a new architecture for symbolic processing. Artificial Intelligence and Information-Control Systems of Robots 94, 31–40, World Scientific, Singapore. (1994).

[Fahlman, 1983] S. Fahlman, G. Hinton, and T. Seijnowski. Massively parallel architectures for AI: NETL, THISTLE and Boltzmann machines. Proc. AAAI National Conf. on AI, 1983:109–113, William Kaufman, Los Altos. (1983).

[Garey, 1979] M. Garey and D. Johnson. Computers and Intractability. A Guide to the Theory of NP-completeness. Freeman, San Francisco, CA, (1979).

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

[Manea, 2006] F. Manea, C. Martın-Vide, and V. Mitrana. All NP-problems can be solved in polynomial time by accepting networks of splicing processors of constant size. Proc. of DNA 12, in press. (2006).

[Martin, 2005] C. Martin-Vide and V. Mitrana. Networks of evolutionary processors: Results and perspectives. Molecular Computational Models: Unconventional Approaches. 78–114, Idea Group Publishing, Hershey. (2005).

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

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

## Authors' Information

***Miguel Angel Díaz** – 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: mdiaz@eui.upm.es*

***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*

# A LEARNING ALGORITHM FOR FORECASTING ADAPTIVE WAVELET-NEURO-FUZZY NETWORK

Yevgeniy Bodyanskiy, Iryna Pliss, Olena Vynokurova

***Abstract:*** *The architecture of forecasting adaptive wavelet-neuro-fuzzy-network and its learning algorithm for the solving of nonstationary processes forecasting tasks are proposed. The learning algorithm is optimal on rate of convergence and allows to tune both the synaptic weights and dilations and translations parameters of wavelet activation functions. The simulation of developed wavelet-neuro-fuzzy network architecture and its learning algorithm justifies the effectiveness of proposed approach.*

***Keywords***: *wavelet, adaptive wavelet-neuro-fuzzy network, recurrent learning algorithm, forecasting, emulation.*

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

## Introduction

At present time the neuro-fuzzy systems have been an increasingly popular technique of soft computing [1-4] successfully applied for the processing of information containing complex nonlinear regularities and distortions of all kinds. These systems combine the linguistic interpretability and the approximation properties of the fuzzy inference systems [5, 6] with the learning and universal approximation capabilities of artificial neural networks [7, 8]. This means, that they can be used in forecasting of stochastic and chaotic signals and sequences with complex nonlinear trends and nonstationary parameters, described by the difference nonlinear autoregression equations (NAR) in the form

$$x(k) = F(x(k)) + \xi(k),$$

where $x(k)$ is a signal value in $k$-th instant of discrete time $k = 0,1,2,\ldots,$ $X(k) = (x(k-1), x(k-2),\ldots, x(k-n))^T$ is $(n\times 1)$ the prehistory vector, which determines present state $x(k)$, $F(\bullet)$ is an arbitrary nonlinear function, unknown in the general case, $\xi(k)$ is a stochastic disturbance with unknown characteristics but with bounded second moment.

Along with the neuro-fuzzy systems for processing the signals of all kinds, the wavelet transform has been an increasingly popular technique [9-11] which provides a compact local signal representation in both time and frequency domain. At the turn of the artificial neural network and wavelets theories the wavelet neural networks have evolved for the analysis of nonstationary processes with considerably nonlinear trends [12-18].

The natural step is to combine the transparency and the interpretability of fuzzy inference systems, powerful approximation and learning capabilities of artificial neural networks and compact description and the flexibility of wavelet transform in the context of hybrid systems of computational intelligence, which further we shall call as the adaptive wavelet-neuro-fuzzy networks (AWNFN).

The key point, defining effectiveness of such systems, is the choice of learning algorithm, which is usually based on the gradient procedures of the accepted criterion minimization. Combination of the gradient optimization with the error backpropagation essentially reduces the rate of learning hybrid systems [19] and leads to necessity of using rather large training samples. In the case when the data processing has to be carried out in real time, forecasted sequence is nonstationary and distorted, conventional gradient descent learning algorithms (let alone genetic algorithms) appeared to be ineffective.

The paper is devoted to the tasks of synthesis of forecasting adaptive wavelet-neuro-fuzzy network, which has higher rate of learning in comparison with systems using conventional backpropagation gradient algorithm.

## Architecture of the forecasting adaptive wavelet-neuro-fuzzy network

Let us introduce into consideration the five-layers architecture, shown on fig. 1, someway similar to the well-known ANFIS [2] which is in turn the learning system of Takagi-Sugeno-Kang fuzzy inference [20, 21].

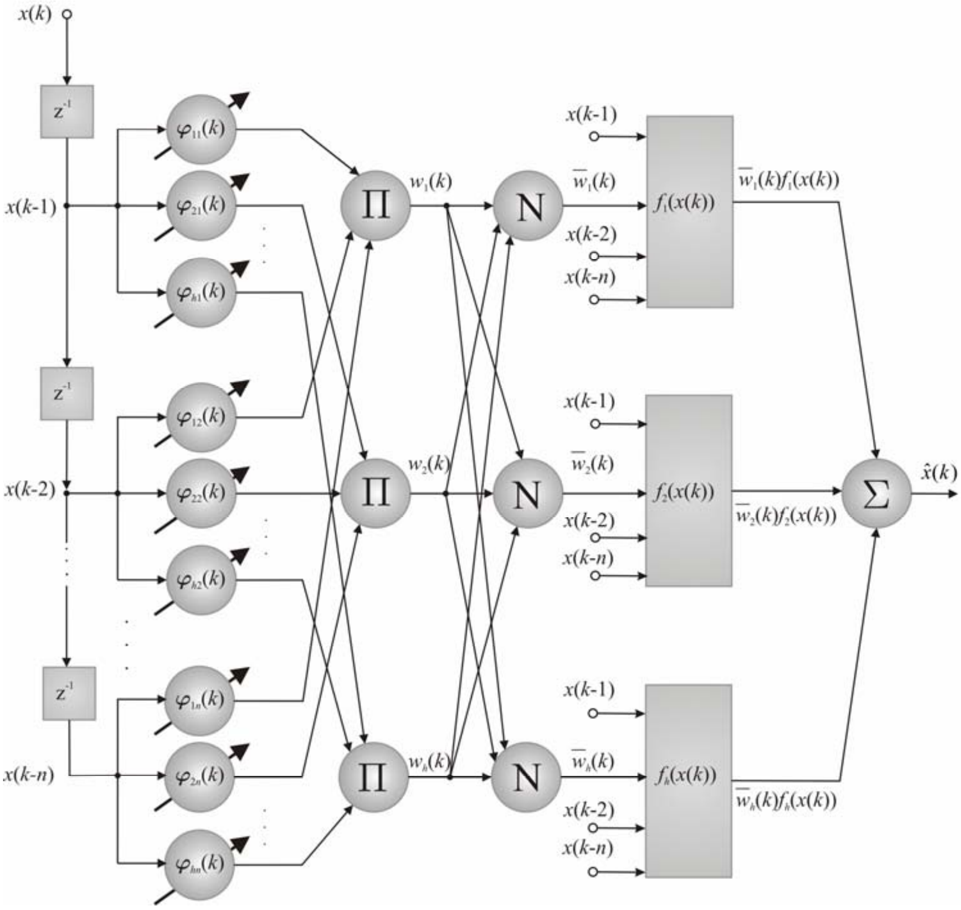


Fig. 1 – Forecasting adaptive wavelet-neuro-fuzzy network

The input layer of the architecture is formed of the time-delay elements $z^{-1}$ ($z^{-1}x(k) = x(k-1)$) and under the input of current signal value $x(k)$ the prehistory vector $X(k) = (x(k-1), x(k-2), \ldots, x(k-n))^T$ is formed as an output of this layer.

The first hidden layer unlike the neuro-fuzzy systems is formed not of conventional non-negative membership functions, but of $hn$ wavelets ($h$ wavelets for each input) $\varphi_{ji}(x(k-i)) = \varphi_{ji}(x(k-i), c_{ji}, \sigma_{ji}) = \varphi_{ji}(k)$ with $2hn$ tuning parameters of dilation (center) $c_{ji}$ and translation (width) $\sigma_{ji}$.

Various kinds of analytical wavelets can be used as the activation functions in adaptive wavelet-neuro-fuzzy network, for example: Morlet wavelets, “Mexican hat” wavelets, Polywog wavelets, Rasp wavelets [12], the generator of analytic wavelets [22], the triangular wavelets [23].

Here it can be noticed, that the oscillation character of wavelet function doesn’t contradict the unipolarity of membership functions as negative values $\varphi_{ji}$ can be interpreted in terms of the small membership or nonmembership levels [24, 25].

The second hidden layer performs the operation similar to computing of fuzzy $T$-norm

$$w_j(k)=\prod_{i=1}^{n}\varphi_{ji}(x(k-i)),\ \ j=1,2,\ldots,h,$$

after that the normalization is performed in the third hidden layer

$$\overline{w}_j(k)=\frac{w_j(k)}{\sum_{j=1}^{h}w_j(k)}=\frac{\prod_{i=1}^{n}\varphi_{ji}(x(k-i))}{\sum_{j=1}^{h}\prod_{i=1}^{n}\varphi_{ji}(x_i(k-i))},$$

providing fulfillment of the condition

$$\sum_{j=1}^{h}\overline{w}_j(k)=1.$$

The fourth hidden layer performs an operation similar to computing of the consequent in the fuzzy inference systems. The most often used function $f_j(x(k))$ in fuzzy inference systems is linear form (in our case local autoregression model):

$$f_j(X(k))=p_{j0}+\sum_{i=1}^{n}p_{ji}x(k-i).$$

In this case in the fourth layer signal values are computed

$$\overline{w}_j(k)(p_{j0}+\sum_{i=1}^{n}p_{ji}x(k-i))=\overline{w}_j(k)p_j^T\overline{X}(k),$$

where $\overline{X}(k)=(1,X^T(k))^T$, $p_j=(p_{j0},p_{j1},\ldots,p_{jn})^T$, and $h(n+1)$ parameters $p_{ji}$, $j=1,2,\ldots,h$, $i=0,1,2,\ldots,n$ are to be determined.

And at last output signal (forecast $\hat{x}(k)$ ) of network is computed in the fifth output layer

$$\hat{x}(k)=\sum_{j=1}^{h}\overline{w}_j(k)f_j(X(k))=\sum_{j=1}^{h}\frac{w_j(k)}{\sum_{j=1}^{h}w_j(k)}f_j(X(k))=\sum_{j=1}^{h}\frac{\prod_{i=1}^{n}\varphi_{ji}(x_i(k-i),c_{ji},\sigma_{ji})}{\sum_{j=1}^{h}\prod_{i=1}^{n}\varphi_{ji}(x_i(k-i),c_{ji},\sigma_{ji})}f_j(X(k)),$$

which, introducing the variables vectors $f(X(k))=(\overline{w}_1(k),\overline{w}_1(k)x(k-1),\ldots,\overline{w}_1(k)x(k-n),\overline{w}_2(k),$ $\overline{w}_2(k)x(k-1),\ldots,\overline{w}_2(k)x(k-n),\ldots,\overline{w}_h(k),\overline{w}_h(k)x(k-1),\ldots,\overline{w}_h(k)x(k-n))^T$, $p=(p_{10},p_{11},\ldots,$ $p_{1n},p_{20},p_{21},\ldots,p_{2n},p_{h0},p_{h1},\ldots,p_{hn})^T$ of dimensionality $h(n+1)$, can be rewritten in the compact form

$$\hat{x}(k)=p^Tf(X(k)).$$

The tunable parameters of this network are located only in the first and fourth hidden layers. These are $2hn$ wavelets parameters $c_{ji}$ and $\sigma_{ji}$, and $h(n+1)$ parameters of the linear local autoregression models $p_{ji}$. Namely they must be determined during the learning process.

## The learning of forecasting adaptive wavelet-neuro-fuzzy network

As far as tunable vector of parameters $p$ is contained in the network description linearly, for its refinement any of the algorithms used in adaptive identification [26] will operate, primarily the exponentially weighted recurrent least squares method (this method is the second order optimization procedure and has both filtering and following properties) in the form

$$\begin{cases} p(k+1) = p(k) + \dfrac{P(k)(x(k) - p^T(k) f(X(k)))}{\alpha + f^T(X(k))P(k) f(X(k))} f(X(k)), \\ P(k+1) = \dfrac{1}{\alpha}\left( P(k) - \dfrac{P(k) f(X(k+1)) f^T(X(k+1))P(k)}{\alpha + f^T(X(k+1))P(k) f(X(k+1))} \right) \end{cases} \quad (1)$$

where $x(k) - p^T(k) f(x(k)) = x(k) - \hat{x}(k) = e(k)$ is the forecasting error, $0 < \alpha \le 1$ is the out-dated information forgetting factor; optimal on operation rate one-step gradient Kaczmarz algorithm [27, 28], having the following properties

$$p(k+1) = p(k) + \frac{x(k) - p^T(k) f(X(k))}{f^T(X(k)) f(X(k))} f(X(k)), \quad (2)$$

or Goodwin-Ramadge-Caines algorithm [29]

$$\begin{cases} p(k+1) = p(k) + r^{-1}(k)(x(k) - p^T f(X(k))) f(X(k)), \\ r(k+1) = r(k) + \| f(X(k+1)) \|^2, \end{cases} \quad (3)$$

which is the stochastic approximation procedure.

Here it should be mentioned, that exponentially weighted recurrent least squares method (1), having filtering and following properties, can be unstable under small values of parameter $\alpha$; convergence of the algorithm (2) under the intensive disturbance $\xi$ is disrupted, and stochastic approximation procedures, including (3), operate only in the stationary conditions.

For tuning of the first hidden layer parameters in AWNFN backpropagation learning algorithm based on the chain rule of differentiation and gradient descend optimization of local criterion

$$E(k) = \frac{1}{2} e^2(k) = \frac{1}{2}(x(k) - \hat{x}(k))^2$$

is used.

In the general case learning procedure in this layer has the form

$$\begin{cases} c_{ji}(k+1) = c_{ji}(k) - \eta_c(k) \dfrac{\partial E(k)}{\partial c_{ji}(k)}, \\ \sigma_{ji}(k+1) = \sigma_{ji}(k) - \eta_\sigma(k) \dfrac{\partial E(k)}{\partial \sigma_{ji}(k)}, \end{cases}$$

and its properties are completely determined by the learning rate parameter $\eta_c(k)$, $\eta_\sigma(k)$, selected according to the empirical reasons. It should be noticed that if the parameters of the fourth layer can be tuning most rapidly, the operation rate is lost in the first layer.

Increasing of the convergence rate can be achieved with more complex than gradient procedures, such as Hartley [30] or Marquardt [31] algorithms which can be written in general form [32]

$$\Phi(k+1) = \Phi(k) + \lambda (J(k) J^T(k) + \eta I)^{-1} J(k) e(k), \quad (4)$$

where $\Phi(k) = (c_{11}(k), \sigma_{11}^{-1}(k), c_{21}(k), \sigma_{21}^{-1}(k), \ldots, c_{ji}(k), \ \sigma_{ji}^{-1}(k), \ldots, c_{hn}(k), \sigma_{hn}^{-1}(k))^T$ is the $(2hn \times 1)$ tunable parameter vector (at that for the computation complexity reduction it includes not the width parameter $\sigma_{ji}$, but its inverse value $\sigma_{ji}^{-1}$), $J(k)$ is the $(2hn \times 1)$ gradient vector of output signal $\hat{x}(k)$ on the tunable parameters, $I$ is the $(2hn \times 2hn)$ identity matrix, $\eta$ is a scalar regularizing parameter, $\lambda$ is the positive scalar gain.

To compute elements of gradient vector

$$J(k)=\left(\frac{\partial\hat{x}(k)}{\partial c_{11}},\frac{\partial\hat{x}(k)}{\partial\sigma_{11}^{-1}},\frac{\partial\hat{x}(k)}{\partial c_{21}},\frac{\partial\hat{x}(k)}{\partial\sigma_{21}^{-1}},\ldots,\frac{\partial\hat{x}(k)}{\partial c_{ji}},\frac{\partial\hat{x}(k)}{\partial\sigma_{ji}^{-1}},\ldots,\frac{\partial\hat{x}(k)}{\partial c_{hn}},\frac{\partial\hat{x}(k)}{\partial\sigma_{hn}^{-1}}\right)^T$$

the chain rule can be used, at that

$$\begin{cases}\dfrac{\partial\hat{x}(k)}{\partial c_{ji}}=\dfrac{\partial\hat{x}(k)}{\partial\overline{w}_j}\cdot\dfrac{\partial\overline{w}_j}{\partial w_j}\cdot\dfrac{\partial w_j}{\partial\varphi_{ji}}\cdot\dfrac{\partial\varphi_{ji}}{\partial c_{ji}}=f_j(X(k))\overline{w}_j(k)(1-\overline{w}_j(k))\dfrac{1}{\varphi_{ji}(x_i(k),c_{ji},\sigma_{ji}^{-1})}\cdot\dfrac{\partial\varphi_{ji}}{\partial c_{ji}},\\ \dfrac{\partial\hat{x}(k)}{\partial\sigma_{ji}^{-1}}=\dfrac{\partial\hat{x}(k)}{\partial\overline{w}_j}\cdot\dfrac{\partial\overline{w}_j}{\partial w_j}\cdot\dfrac{\partial w_j}{\partial\varphi_{ji}}\cdot\dfrac{\partial\varphi_{ji}}{\partial\sigma_{ji}^{-1}}=f_j(X(k))\overline{w}_j(k)(1-\overline{w}_j(k))\dfrac{1}{\varphi_{ji}(x_i(k),c_{ji},\sigma_{ji}^{-1})}\cdot\dfrac{\partial\varphi_{ji}}{\partial\sigma_{ji}^{-1}}.\end{cases}$$

where $\partial\varphi_{ji}/\partial c_{ji}$, $\partial\varphi_{ji}/\partial\sigma_{ji}^{-1}$ is partial derivatives of concrete wavelet activation function.

To reduce the computational complexity of the learning algorithm we can use the matrix inversion lemma in the form

$$(JJ^T+\eta I)^{-1}=\eta^{-1}I-\frac{\eta^{-1}IJJ^T\eta^{-1}I}{1+J^T\eta^{-1}IJ},$$

using which it is easy to obtain the relation

$$\lambda(JJ^T+\eta I)^{-1}J=\lambda\frac{J}{\eta+\|J\|^2}.$$

Substituting this relation to the algorithm (4), we obtain first hidden layer parameters learning algorithm in the form

$$\Phi(k+1)=\Phi(k)+\lambda\frac{J(k)e(k)}{\eta+\|J(k)\|^2}. \tag{5}$$

It is easy to see, that algorithm (5) is the nonlinear additive-multiplicative modification of Kaczmarz algorithm, and under $\lambda=1,\eta=0$ coincides with it structurally.

To provide the filtering properties to the learning algorithm (5) let us introduce additional tuning procedure of the regularizing parameter $\eta$ in the form [33, 34, 35]

$$\begin{cases}\Phi(k+1)=\Phi(k)+\lambda\dfrac{J(k)e(k)}{\eta(k)},\\ \eta(k+1)=\alpha\eta(k)+\|J(k+1)\|^2.\end{cases} \tag{6}$$

If $\alpha=0$, then this procedure coincides with (5) and has the highest rate of convergence, and if $\alpha=1$, then this procedure obtains properties of stochastic approximation, and serves as generalization of procedure (3) in the nonlinear case.

Here it should be noticed, that the algorithm (6) is stable at any value of forgetting factor $\alpha$, what favorably differs it from the exponentially weighted recurrent least squares method (1). As a result this procedure can be used too in the form

$$\begin{cases}p(k+1)=p(k)+\lambda_p\eta_p^{-1}(k)(x(k)-p^T(k)f(X(k)))f(X(k)),\\ \eta_p(k+1)=\alpha\eta_p(k)+\|f(x(k))\|^2\end{cases} \tag{7}$$

for the fourth layer parameters tuning. One can notice close relation of the algorithms (1) and (**Error! Reference source not found.**), as

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

However algorithm (**Error! Reference source not found.**) is much simpler in the computing implementation and easily reconstructs its properties from the most following to the most filtering ones.

## Simulation results

To demonstrate the effectiveness of the proposed adaptive wavelet-neuro-fuzzy-network and its learning algorithm (6), (7), AWNFN was trained to forecast the Mackey-Glass chaotic time series. Forecasting of the Mackey-Glass time series is a standard test, widely used to evaluate and compare the performance of neural and neuro-fuzzy systems for nonlinear system modeling and time series forecasting. The Mackey-Glass time series is generated by the time-delay differential equation [36]

$$\dot{x}(t) = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t). \tag{8}$$

The values of the time series (8) were obtained at each integer point by means of the fourth-order Runge-Kutta method. The time step used in this method was set 0.1, while initial condition $x(0)=1.2$, delay $\tau=17$ and $x(t)$ was obtained for $t=0\ldots 51000$. The values $x(t-18), x(t-17), x(t-6), x(t)$ were used to forecast $x(t+6)$. In the online mode of learning, AWNFN was trained with procedure (6), (7) for 50000 iteration (50000 training samples for $t=118,\ldots 50117$). The parameters of the learning algorithm were $\alpha=0.95$, $\lambda_p=2$, $\lambda=1$. Initial values were $\eta(0)=1$ and $\eta_p(0)=10000$. After 50000 iterations the training was stopped, and the next 500 points for $t=50118\ldots 50617$ were used as the testing data set to compute forecast. As the activation function “Maxican hat” wavelet is used. Initial values of synaptic weights were generated in a random way from $-0.1$ to $+0.1$.

The root mean-square error (RMSE) was used as criterion for the quality of forecasting

$$RMSE = \frac{1}{N}\sum_{k=1}^{N}(x(k)-\hat{x}(k))^2 .$$

Fig. 5 shows the results of chaotic time series forecasting. The two curves, representing the actual (dot line) and forecasting (solid line) values, are almost indistinguishable.

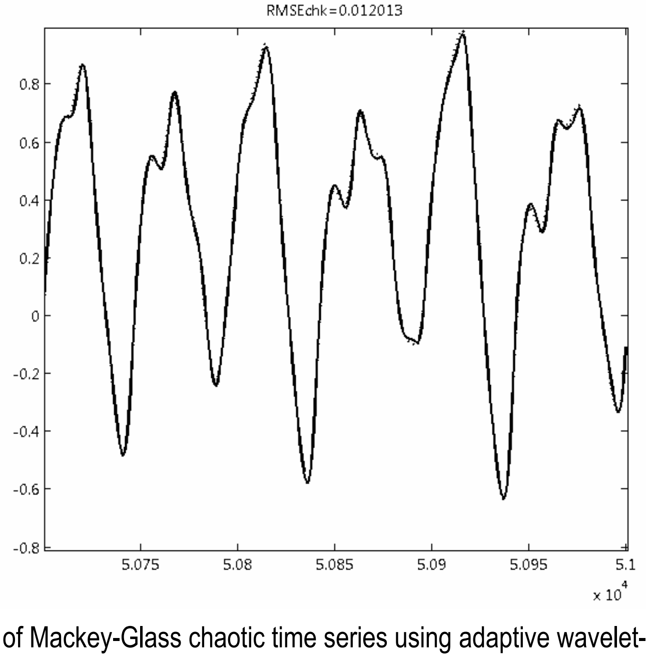


Fig. 5 – Forecasting of Mackey-Glass chaotic time series using adaptive wavelet-neuro-fuzzy network (online-learning, 50000 iteration)

Table 1 shows results of forecasting process on the basis of the adaptive wavelet-neuro-fuzzy-network compared the results of forecasting process on the basis of standard ANFIS with the backpropagation learning algorithm.

Table 1: The results of chaotic time series forecasting

| Neural network/ Learning algorithm | RMSE |
|---|---|
| Adaptive wavelet-neuro-fuzzy-network / Proposed learning algorithm (6), (7) | 0.0120 |
| Backpropagation ANFIS | 0.2312 |

Thus as it can be seen from experimental results the proposed forecasting adaptive wavelet-neuro-fuzzy-network with the learning algorithm (6), (7) having the same number of adjustable parameters ensures the best quality of forecast and high learning rate in comparison with conventional ANFIS architecture.

## Conclusions

Computationally simple learning algorithms for the adaptive wavelet-neuro-fuzzy network in the forecasting of nonlinear nonstationary signals tasks are proposed. The simulation of developed approach justifies the effectiveness of AWNFN using for solving wide category of emulation, forecasting and diagnostics problems.

## Bibliography

[1]. Jang J.-S. R. Neuro-Fuzzy Modeling: Architectures, Analyses and Applications. Berkeley, CA: University of California. 1992, 155 p.

[2]. Jang J.-S. R. ANFIS: Adaptive network-based fuzzy inference systems. IEEE Trans. on Systems, Man, and Cybernetics. 23, 1993, P. 665-685.

[3]. Jang J.-S. R., Sun C.-T. Neuro-fuzzy modeling and control. Proc. IEEE. 83 (3), 1995, P. 378-406.

[4]. Jang J.-S. R., Sun C.-T., Muzutani E. Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence. Upper Saddle River, N.J.: Prentice Hall, Inc., 1997, 614 p.

[5]. Kosko B. Fuzzy systems as universal approximators. Proc. 1-st IEEE Int. Conf. on Fuzzy Systems. San Diego, CA. 1992, P. 1153-1162.

[6]. Kreinovich V., Mouzouris Y. C., Nguen H. T. Fuzzy rule based modeling as a universal approximation tool. In “Fuzzy Systems: Modeling and Control”. Eds.: H. T. Nguen, M. Sugeno. Boston: Kluwer Academic Publishers. 1998, P. 135-195.

[7]. Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators. Neural Networks. 2, 1982, P. 359-366.

[8]. Scarcelli F., Tsoi A. S. Universal approximation using feedforward neural networks: a survey of some existing methods and some new results. Neural Networks. 11, 1998, P. 15-37.

[9]. Chui C. K. An Introduction to Wavelets. New York: Academic. 1992, 264 p.

[10]. Daubechies I. Ten Lectures on Wavelets. Philadelphia, PA: SIAM. 1992, 228 p.

[11]. Meyer Y. Wavelets: Algorithms and Applications. Philadelphia, PA: SIAM. 1993, 133 p.

[12]. Lekutai G., VanLandingham H. F. Self-tuning control of nonlinear systems using neural network adaptive frame wavelets. Proc. IEEE Int. Conf. on Systems, Man and Cybernetics. Piscataway, N.J. 2, 1997, P. 1017-1022.

[13]. Billings S. A., Wei H.-L. A new class of wavelet networks for nonlinear system identification. IEEE Trans. on Neural Networks. 16 (4), 2005, P. 862-874.

[14]. Zhang Q. H., Benveniste A. Wavelet networks. IEEE Trans. on Neural Networks. 3 (6), 1992, P. 889–898.

[15]. Oussar Y., Dreyfus G. Initialization by selection for wavelet network training. Neurocomputing. 34, 2000, P. 131–143.

[16]. Zhang J., Walter G. G., Miao Y., Lee W. N. W. Wavelet neural networks for function learning. IEEE Trans. on Signal Process. 43(6), 1995, P. 1485–1497.

[17]. Zhang Q. H. Using wavelet network in nonparametric estimation. IEEE Trans. on Neural Networks. 8 (2), 1997, P. 227–236.

[18]. Bodyanskiy Ye., Lamonova N., Pliss I., Vynokurova O. An adaptive learning algorithm for a wavelet neural network. Blackwell Synergy: Expert Systems. 22 (5), 2005, P. 235-240.

[19]. Avci E., Turkoglu I., Poyraz M. Intelligent target recognition based on wavelet adaptive network based fuzzy inference system. Lecture Notes on Computer Science. Berlin-Heidelberg: Springer-Verlag, 3522, 2005, P. 594-603.

[20]. Takagi T., Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. on Systems, Man, and Cybernetics. 15, 1985, P. 115-132.

[21]. Sugeno M., Kang G. T. Structure identification of fuzzy model. Fuzzy Sets and Systems, 28, 1998, P. 15-33.

[22]. Vynokurova O., Lamonova N., Pliss I. Analytic wavelets generator. Problemy Bioniki, 60, 2004, P. 104-109. (in Russian)

[23]. Bodyanskiy Ye., Vynokurova O. Variable form of triangular wavelet in the wavelet neural networks. Proc. of 13th International Conference on Automatic Control “Automatics-2006”, Vinnica, 2006, P. 393. (in Russian)

[24]. 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”, 2, 1996, P. 1237-1213.

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

[26]. Ljung L. System Identification: Theory for the user. PTR Prentice Hall, Upper Saddle River, N.J., 1999

[27]. Kaczmarz S. Angenaeherte Ausloesung von Systemen linearer Gleichungen. Bull. Int. Acad. Polon. Sci., Let. A , 1973, S. 355-357.

[28]. Kaczmarz S. Approximate solution of systems of linear equations. Int. J. Control, 53, 1993, P. 1269-1271.

[29]. Goodwin Y. C., Ramadge P. J., Caines P. E. A globally convergent adaptive predictor. Automatica, 17, 1981, P. 135-140.

[30]. Hartley H. The modified Gauss-Newton method for the fitting of nonlinear regression functions of least squares. Technometrics, 3, 1961, P. 269-280.

[31]. Marquardt D. An algorithm for least squares estimation of nonlinear parameters. SIAM J. Appl. Math., 11, 1963, P. 431-441.

[32]. Bodyanskiy Ye. Adaptive identification algorithm of nonlinear control object. Automatic control system and devices of automatic, Kharkiv: Vyshcha shk., 81, 1987, P. 43-46. (in Russian)

[33]. Bodyanskiy Ye., Pliss I., Solovyova T. Multistage optimal predictors multidimensional nonstationary stochastic process. Docl. AN USSR. 12, 1986, P. 41-49.

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

[35]. Bodyanskiy Ye., Kolodyazhniy V., Otto P. A new learning algorithm for a forecasting neuro-fuzzy network. Integrated Computer-Aided Engineering. Amsterdam: IOS Press, 10 (4), 2003, P. 399-409.

[36]. Mackey M. C., Glass L. Oscillation and chaos in physiological control systems. Science, 1977, 197, P. 287-289

## 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 Electronic, Lenina av. 14, Kharkiv, Ukraine 61166, Tel +380577021890, bodya@kture.kharkov.ua*

***Pliss Iryna** - 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, Ukraine, 61166, Tel +380577021890, pliss@kture.kharkov.ua*

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

# GENERALIZED REGRESSION NEURO–FUZZY NETWORK

Yevgeniy Bodyanskiy, Nataliya Teslenko

***Abstract:*** *Generalized Regression Neuro-Fuzzy Network, which combines the properties of conventional Generalized Regression Neural Network and Adaptive Network-based Fuzzy Inference System is proposed in this work. This network relates to so-called "memory-based networks", which is adjusted by one-pass learning algorithm.*

***Keywords:*** *memory-based networks, one-pass learning, Fuzzy Inference Systems, fuzzy-basis membership functions, neurons at data points, nonlinear identification.*

***ACM Classification Keywords****: F.1 Computation by abstract devices - Self-modifying machines (e.g., neural networks), I.2.6 Learning - Connectionism and neural nets, G.1.2. Approximation – Nonlinear approximation.*

## Introduction

Nowadays neural networks have wide spreading for identification, prediction and nonlinear objects control problems solving. Neural networks possess universal approximating abilities and capabilities for learning by the data that characterize the functioning of investigating systems. The situation becomes sharply complicated in the case, when the data are fed in real time, their processing must be simultaneous with functioning of the plant, and the plant is nonstationary. It's clear, that conventional multilayer perceptron, that is universal approximator, isn't effective in this case, so Radial Basis Functions Networks (RBFN) can be used as its alternative [1-3]. These networks are also universal approximators, and their output is linearly dependent on tuned synaptic weights. In this case, recurrent least squares method or its modifications can be used for their real time learning. These procedures are second-order optimization algorithms, which provide quadratic convergence to the optimal solution. At the same time, practical application of RBFN is bounded by so-called curse of dimensionality as well as appearance of "gaps" in the space of radial-basis functions (RBF) that lead to appearance of regions where all neurons of the network are inactive.

So-called, space partition of unity, implemented by Normalized Radial Basis Functions Networks (NRBFN), in which output signal is normalized by the sum of outputs of all neurons, is used to avoid such a "gaps" [4]. Given networks are learned using recurrent gradient algorithms that have slow rate of convergence and possibility of getting to the local minima as their common drawback.

Thus, these neural networks and many others that use recurrent learning procedures and united by general name "optimization-based networks" may be inefficient in problems of adaptive identification, prediction and real-time control, when the information is fed for processing with sufficiently high frequency. In this case, these networks have not time to learn and are unable to follow changing parameters of a plant.

The so-called "memory-based networks" are the effective alternative to "optimization-based networks" and Generalized Regression Neural Network (GRNN), proposed by D. F Specht [5], is the brightest representative of these networks.

At the basis of this network lies the idea of Parzen windows [6], kernel estimates of Nadaraya-Watson [7-9] and nonparametric models [10]. Its learning consists of one-time adjustment of multidimensional radial-basis functions (RBF) at points of unit centered hypercube, which are specified unambiguously by the learning set. Therefore, these networks can be referred to, so-called, just-in-time models [4], which are adjusted by one-pass learning algorithm. Being similar to NRBFN by the architecture, GRNN learns much faster, placing the centers of RBF at the points with coordinates that are determined by input signals of a plant using principle "neurons at the data points" [11] and with RBF heights, which coincide with corresponding values of plant output signal. High learning rate of GRNN provides their effective using in the real-time problems solving [12,13].

For the solving of nonlinear plant identification problem

$$y(k) = F(x(k))$$

where $y(k)$, $x(k)$– scalar and $(nx1)$-vector of output and input signals correspondingly in the instant time $k=1,2,\ldots$, $F(\bullet)$ – unknown nonlinear operator of the plant, it is necessary to form learning sample $\left\{x^*(k), y^*(k)\right\}$, $k=1,2,\ldots,l$, whereupon it is possible to get the estimate $\widehat{y}(k)$ of the plant response $y(k)$ to arbitrary input signal $x$ in the form

$$\widehat{y}(x) = \frac{\sum_{k=1}^{l} y^*(k)\varphi(D(k))}{\sum_{k=1}^{l} \varphi(D(k))} \tag{1}$$

where $D(k)$ – distance measure in accepted metrics between $x$ and $x^*(k)$, $\varphi(\bullet)$ – some kernel function, usually, Gaussian. Conventionally the Euclidean metrics is used as a distance

$$D^2(k) = \sum_{i=1}^{n} \left( \frac{x_i(k) - x_i^*(k)}{\sigma(k)} \right)^2$$

(here $\sigma(k)$ – scalar parameter, which determines the receptive field radius of kernel function $\varphi(\bullet)$), although in more common case it is possible to use Minkowski metrics

$$D^p(k) = \sum_{i=1}^{n} \left| \frac{x_i - x_i^*(k)}{\sigma(k)} \right|^p, \ p \geq 1 .$$

Thus, GRNN converges asymptotically to optimal nonlinear regression surface with the growing of learning sample size [9].

GRNN learning process can be organized easily in real time. In this case the learning pairs $x^*(k)$, $y^*(k)$ are fed to the network sequentially, forming new radial-basis function-neurons. At the same time, the distance between newly formed and already existing functions is estimated gradually. If this distance is smaller than threshold value $r$, that is defined in advance, new neuron isn't included in the network. The main problems concerned with GRNN using are defined by possible curse of dimensionality. Growing of the learning sample size $l$ and the difficulties with correct definition of parameter $r$, which is sufficiently difficult to choose and interpret in multidimensional space, are the causes of it.

Neuro-Fuzzy Systems (NFS) are the natural expansion of artificial neural networks [14-15]. They combine the neural networks learning abilities with transparence and interpretability of the Fuzzy Inference Systems (FIS). Generally, FIS represents fuzzy models, which are learned by observations data of plant inputs and outputs, using univariate Fuzzy Basis Functions (FBF) instead of multidimensional RBF. In common case FBF are bell-shaped (usually Gaussian) membership functions, which are used in Fuzzy Logic. Using of bell-shaped FBF allows us to combine local features of the kernel functions with the properties of sigmoidal activation functions that provide global approximation properties [16]. Having the approximating abilities of RBFN [15], NFS subject to curse of dimensionality with less degree, that provides them advantage in comparison with neural networks.

Among Neuro-Fuzzy Systems (NFS) Adaptive Network-based Fuzzy Inference System (ANFIS) have got wide spread [17]. ANFIS has five-layer architecture, whose synaptic weights are tuned similarly to RBFN. The adjusting possibility of FBFs using error back-propagation algorithm is provided in this system too. ANFIS and many other similar neuro-fuzzy systems [4,15,16] are typical representatives of the optimization-based networks family, which are characterized by insufficient learning rate.

Lattice-based Associative Memory Networks (LAMN) [18, 19] are the representatives of memory-based networks, whose output signal is formed on basis of univariate bell-shaped functions uniformly distributed on axes of n-dimensional input space. As a result of aggregation operation multidimensional FBFs are formed, whose centers are also uniformly distributed in multidimensional space, and their layout doesn't depend on characteristics of learning sample.

The goal of this work is the development of Generalized Regression Neuro-Fuzzy Network (GRNFN), which represents by itself NFS and learns as GRNN that provides it approximating properties of ANFIS with learning rate of memory-based networks.

## The Generalized Regression Neuro-Fuzzy Network architecture

The architecture of Generalized Regression Neuro-Fuzzy Network is illustrated on Fig. 1 and consists of five sequentially connected layers. First hidden layer is composed of *l* blocks with *n* FBF in each and realizes fuzzification of the input variables vector. Second hidden layer implements aggregation of membership levels that are computed in first layer, and consists of *l* multiplication blocks. Third hidden layer – the layer of synaptic weights that are defined in special way. Fourth layer is formed by two summation units and computes the sums of output signals from the second and third layers. Finally, normalization takes place in fifth (output) layer, as a result of which, the output network signal is computed.

One can see, that the architecture of GRNFN coincides with the architecture of L.-X. Wang–-J.M. Mendel neuro-fuzzy system [20], which, in turn, is the modification of zero-order T. Takagi–M. Sugeno fuzzy inference system [21]. However, if NFS is learned using one or another optimization procedures, GRNFN is adjusted using one-pass learning algorithm.

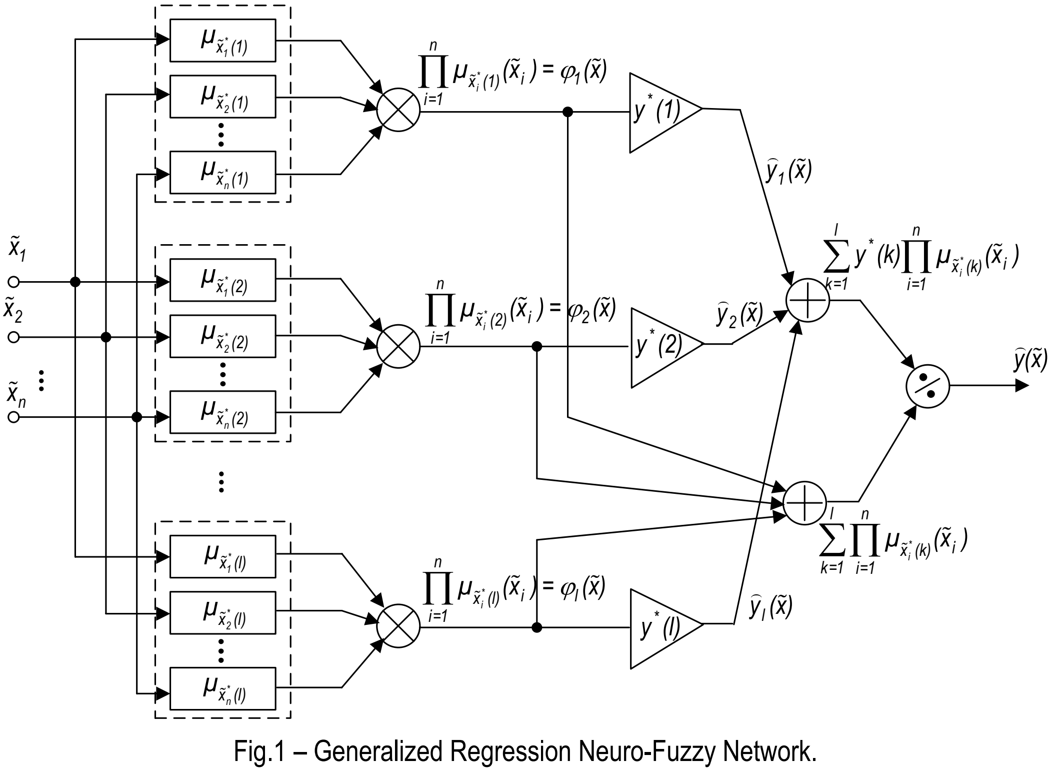


Fig.1 – Generalized Regression Neuro-Fuzzy Network.

## Generalized Regression Neuro-Fuzzy Network learning

Since GRNFN belongs to memory-based networks, its learning is based on principle "neurons at data points" that makes it extremely easy and fast.

Learning sample vectors $x^*(1),...,x^*(k),...,x^*(l)$ are normalized in advance on unit centered hypercube so, that

$$x_i^{*min} \le x_i^*(k) \le x_i^{*max}\ ,\ \ i=1,2,...,n\ , \qquad 1$$

$$-0,5 \le \tilde{x}_i^*(k) \le 0,5 \ .$$

Mutual recalculation is made according to the next expressions

$$\tilde{x}_i^*(k) = \frac{x_i^*(k) - x_i^{*min}}{x_i^{*max} - x_i^{*min}} - 0,5 \ ,$$

$$x_i^*(k) = (\tilde{x}_i^*(k) + 0,5)(x_i^{*max} - x_i^{*min}) + x_i^{*min} \ .$$

For each vector from the learning sample $\tilde{x}^*(k) = (\tilde{x}_1^*(k), \tilde{x}_2^*(k), ..., \tilde{x}_n^*(k))^T$ in the first hidden layer own set of fuzzy-basis membership functions $\mu_{\tilde{x}_1^*(k)}, \mu_{\tilde{x}_2^*(k)}, ..., \mu_{\tilde{x}_n^*(k)}$ is formed, so that centers of $\mu_{\tilde{x}_i^*(k)}$ coincide with $\tilde{x}_i^*(k)$, $k=1,2,...,l$. The process of FBF formation is illustrated on Fig. 2. Note that GRNFN contains $nl$ fuzzy-basis functions, that can't lead to the curse of dimensionality.

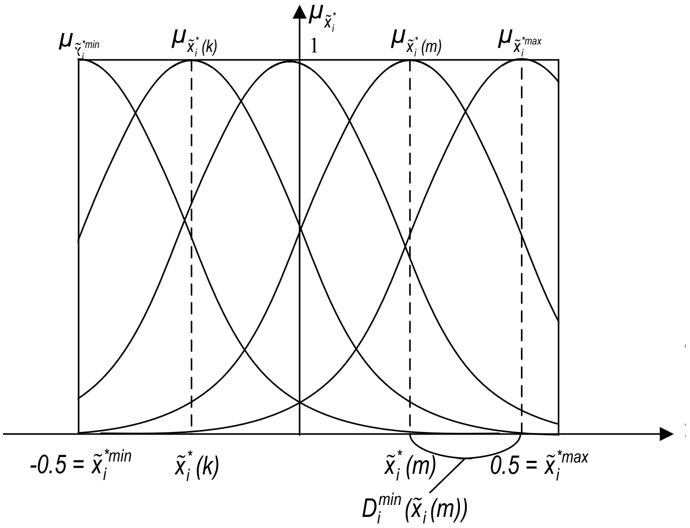

$\mu_{\tilde{x}_i^{*min}}$
$\mu_{\tilde{x}_i^*(k)}$
$\mu_{\tilde{x}_i^*}$
1
$\mu_{\tilde{x}_i^*(m)}$
$\mu_{\tilde{x}_i^{*max}}$
$-0.5 = \tilde{x}_i^{*min}$
$\tilde{x}_i^*(k)$
$\tilde{x}_i^*(m)$
$0.5 = \tilde{x}_i^{*max}$
$D_i^{min}(\tilde{x}_i(m))$

Fig.2 – Fuzzy-basis membership functions.

Theoretically, any kernel function with non-strictly local support can be used as FBF. It allows avoiding of appearance of "gaps" [9]. As such a function one can recommend generalized Gaussian

$$\mu_{\tilde{x}_i^*(k)}(\tilde{x}_i) = \left( 1 + \left| \frac{\tilde{x}_i^*(k) - \tilde{x}_i}{\sigma_i(k)} \right|^{2b} \right)^{-1}, \ b \ge 0,5 \ , \qquad (2)$$

that is the bell-shaped function, whose shape is defined by the scalar parameter $b$ [15]. Let's also note, that $b$ defines the metrics $D^{2b}(k)$ too. As for choosing of the width parameter $\sigma_i(k)$, standard recommendation leads to the idea [8], that it must ensures small overlapping of neighboring FBFs. Easy to see, that for Gaussian this recommendation leads to estimate

$$\sigma_i(k) < \frac{l-1}{2 \div 3} \ .$$

At the same time with FBFs forming in first hidden layer, the synaptic weights are being tuned in the third hidden layer and they are supposed to be equal to the signals of learning sample $y^*(k)$.

Thus, when arbitrary signal $\tilde{x}$ is fed to the input of GRNFN in the first hidden layer membership levels $\mu_{\tilde{x}_i^*(k)}(\tilde{x}_i)$, $i=1,2,...,n$, $k=1,2,...,l$ are computed, in the second layer their aggregation is made by forming multidimensional FBFs

$$\varphi_k(\tilde{x}) = \prod_{i=1}^{n} \left( 1 + \left| \frac{\tilde{x}_i^*(k) - \tilde{x}_i}{\sigma_i(k)} \right|^{2b} \right)^{-1}, \quad k = 1,2,...,l,$$

in the third layer products $\widehat{y}(\tilde{x}) = y^*(k)\varphi_k(\tilde{x})$ are determined, fourth layer computes the values of signals $\sum_{k=1}^{l} y^*(k)\varphi_k(\tilde{x})$ and $\sum_{k=1}^{l} \varphi_k(\tilde{x})$, and, finally, in the output layer the estimate

$$\widehat{y}(\tilde{x}) = \frac{\sum_{k=1}^{l} y^*(k)\varphi_k(\tilde{x})}{\sum_{k=1}^{l} \varphi_k(\tilde{x})} = \frac{\sum_{k=1}^{l} y^*(k) \prod_{i=1}^{n} \mu_{\tilde{x}_i^*(k)}(\tilde{x}_i)}{\sum_{k=1}^{l} \prod_{i=1}^{n} \mu_{\tilde{x}_i^*(k)}(\tilde{x}_i)},$$

is forming, which coincides with (1) with the only difference, that instead of radial-basis functions multidimensional fuzzy-basis functions are used, that were formed of univariate FBF.

The scheme of fuzzy inference, which is realized by GRNFN can be presented as a logic equations system

$$IF(\tilde{x}_1.IS.A_1(1)).AND.(\tilde{x}_2.IS.A_2(1)).AND.....AND.(\tilde{x}_n.IS.A_n(1)), \quad THEN \quad \widehat{y}_1(\tilde{x}) = y^*(1)$$

$$\vdots$$

$$IF(\tilde{x}_1.IS.A_1(k)).AND.(\tilde{x}_2.IS.A_2(k)).AND.....AND.(\tilde{x}_n.IS.A_n(k)), \quad THEN \quad \widehat{y}_k(\tilde{x}) = y^*(k)$$

$$\vdots$$

$$IF(\tilde{x}_1.IS.A_1(l)).AND.(\tilde{x}_2.IS.A_2(l)).AND.....AND.(\tilde{x}_n.IS.A_n(l)), \quad THEN \quad \widehat{y}_l(\tilde{x}) = y^*(l)$$

where the operator $A_i(k)$ is represented by the membership function (2). Hence, using of neuro-fuzzy approach allows ensuring of obtained results interpretation.

The GRNFN learning process can proceed both in batch mode, when learning sample $\left\{ x^*(k), y^*(k) \right\}$ is specified apriori and in real time, when pairs *x*(k), y*(k)* are given sequentially, forming multidimensional FBFs $\varphi_k$. It is sufficiently easy to organize the exclusion process of slight information pairs. If for some observation $\tilde{x}^*(m)$ next condition is held

$$\max_i D_i^{min}(\tilde{x}_i(m)) < r < (l-1)^{-1} \tag{3}$$

(here $D_i^{min}(\tilde{x}_i(m))$ – the least distance between $\tilde{x}_i(m)$ and earlier formed neighboring centers of FBFs), then $\tilde{x}^*(m)$ doesn't form function $\varphi_m$ and is removed from the consideration. Note, that for univariate situation the threshold parameter *r* and the distance $D_i^{max}$ are significantly easier to define, then in multidimensional case of GRNN.

Operation of GRNFN can be organized simply in the continuous adaptation mode that is essentially important for nonstationary objects identification and control. Here it is possible to use two approaches. The first is – on the sliding window of *l* observations, when while learning pairs *x*(l+1), y*(l+1)* are being fed to the input of the network, in the first and third layers the pair of $\mu_{\tilde{x}_i^*(1)}$ and *y*(1)* is removed, and instead of it the membership function $\mu_{\tilde{x}_i^*(l+1)}$ and weight *y*(l+1)* are formed. The second approach is based on inequality (3). In this case newly received pair *x*(m), y*(m)* isn't removed, but replaces the nearest to it in the “old” data.

As far as the learning process operates almost immediately, there is no problem with following properties of tuning algorithm at all.

## Numerical experiment

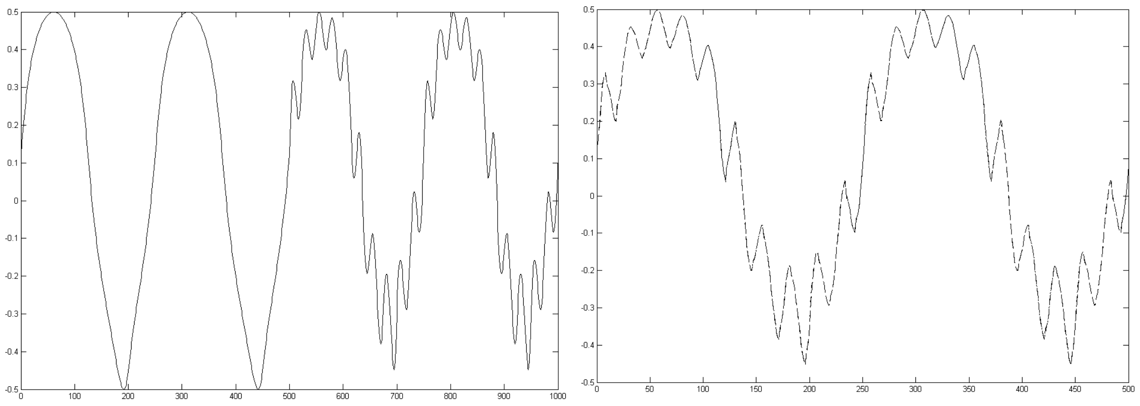
In this experiment, the plant is assumed to be of the form [22]:

$$y(k+1) = f(y(k), y(k-1), y(k-2), u(k), u(k-1)),$$

where the unknown function *f* has the form

$$f(x_1, x_2, x_3, x_4, x_5) = \frac{x_1 x_2 x_3 x_5 (x_3 - 1) + x_4}{1 + x_2^2 + x_3^2}.$$

The input to the plant is given by $u(k) = sin(2\pi k/250)$ for $k \leq 500$ and $u(k) = 0.8sin(2\pi k/250) + 0.2sin(2\pi k/25)$ for $k > 500$, in all 1000 signals. Fig.3(a) shows the output of the plant.



(a) (b)

Fig. 3. (a) Outputs of the plant. (b) Outputs of the GRNFN (dash-dot line) and GRNN (dashed line) practically coincide.

Two experiments were made. In the first experiment, GRNFN was constructed and learned by first 500 signals, which organized learning sample. After that, the next 500 signals were fed to the network for testing its performance. In addition, this problem was solved using conventional GRNN. The results are shown in Fig.3(b) for last 500 instants. One can see that output signals of GRNFN and GRNN practically agree with test signals and with each other, but numerical analysis shows that GRNFN has accuracy higher by 2%. In the second experiment the distances between all learning signals were computed and compared with threshold value. Only 378 of 500 signals exceeded preassigned threshold value, and they organized learning sample. In this case, GRNFN has the same accuracy. Hence, it is logically to conclude that GRNFN needs less number of signals to be learned in comparison with GRNN.

## Conclusions

Generalized Regression Neuro-Fuzzy Network, that is generalization of conventional GRNN and adaptive fuzzy inference systems, is proposed in this work. Network is characterized by computational simplicity, interpretability of the results and ensures high accuracy in the nonlinear nonstationary systems prediction and identification problems.

## Bibliography

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

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

[3] Schilling R.J., Carrol J.J., Al-Ajlouni A.F. Approximation of nonlinear systems with radial basis function neural networks// IEEE Trans. on Neural Networks.-2001.-12.-P.1-15.

[4] Nelles O. Nonlinear System Identification.-Berlin: Springer, 2001.-785p.

[5] Specht D.E. A general regression neural network// IEEE Trans. on Neural Networks.-1991.-2.-P.568-576.

[6] Parzen E. On the estimation of a probability density function and the mode//Ann. Math. Stat.-1962.-38.-P.1065-1076.

[7] Nadaraya E.A. About nonparametric probability density and regression estimates// Probability theory and its Application.-1965.-10.-№1.-P199-203.

[8] Bishop C.M. Neural Networks for Pattern Recognition.- Oxford: Clarendon Press, 1995.-482p.

[9] Friedman J., Hastie T., Tibshirani R. The Elements of Statistical Learning. Data Mining, Inference, and Prediction.- Berlin: Springer, 2003.-552p.

[10] Zhivoglyadov V.G., Medvedev A.V. Nonparametric algorithms of adaptation.-Frunze: Ilim, 1974.-135p. (in Russian).

[11] Zahirniak D.R., Capman R., Rogers S.K., Suter B.W., Kabrisky M., Pyati V. Pattern recognition using radial basis function network// Proc. 6-th Ann. Aerospace Application of AI Conf.- Dayton, OH, 1990.-P.249-260.

[12] Seng T.L., Khalid M., Yusof R., Omatu S. Adaptive neuro-fuzzy control system by RBF and GRNN neural networks// J. of Intelligent and Robotic Systems.- 1998.-23.-P.267-289.

[13] Guo X.-P., Wang F.-L., Jia M.-X. A sub-stage moving window GRNN quality prediction method for injection molding process// "Lecture Notes in Computer Science"- V3973.- Berlin-Heidelberg: Springer-Verlag, 2006.-P.1138-1143.

[14] Jang J.-S. R., Sun G.-T. Neuro-fuzzy modeling and control// Proc. IEEE.-1995.-83.-P.378-406.

[15] Jang J.-S. R., Sun G.-T., Mizutani E. Neuro-Fuzzy and Soft Computing.- Upper Saddle River, NJ: Prentice Hall, 1997.-614p.

[16] Cios K.J., Pedrycz W. Neuro-Fuzzy algorithms// In: "Handbook on Neural Computation" – Oxford: University Press, 1997.-D1.3:1-D1.3:7.

[17] Jang J.-S. R. ANFIS: Adaptive-Network-based Fuzzy Inference Systems// IEEE Trans. on Systems, Man, and Cybernetics.-1993.-23.-P.665-685.

[18] Brown M., Harris C.J. Neural networks for modeling and control/ In: Eds. by C.J. Harris "Advances in Intellectual Control".- London: Taylor and Francis, 1994.-P.17-55.

[19] Wang H., Liu G.P., Harris C.J., Brown M. Advanced Adaptive Control.- Oxford: Pergamon, 1995.- 262p.

[20] Wang L.-X., Mendel J.M. Fuzzy basis functions, universal approximation, and orthogonal least squares learning//IEEE Trans. on Neural Networks.-1992.-3.-P.807-814.

[21] Takagi T., Sugeno M. Fuzzy identification of systems and its applications to modeling and control// IEEE Trans. on Systems, Man, and Cybernetics.-1985.-15.-P.116-132.

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

## Authors' Information

***Yevgeniy Bodyanskiy –*** *Kharkiv National University of Radio Electronics, Doctor of Technical Sciences, Professor of Artificial Intelligence Department, Head of Control Systems Research Laboratory, IEEE Senior Member; Postal address: CSRL, Ofiice 511, Lenin Av., 14, Kharkiv, 61166, Ukraine; e-mail: bodya@kture.kharkov.ua.*

***Nataliya Teslenko -*** *Kharkiv National University of Radio Electronics, post-graduate student, research scientist of Control Systems Research Laboratory; Postal address: CSRL, Ofiice 511, Lenin Av., 14, Kharkiv, 61166, Ukraine; e-mail: ntntp@ukr.net.*

# A STATISTICAL CONVERGENCE APLICATION FOR THE HOPFIELD NETWORKS

**Víctor Giménez-Martínez, Gloria Sánchez–Torrubia, Carmen Torres–Blanc**

***Abstract***: *When Recurrent Neural Networks (RNN) are going to be used as Pattern Recognition systems, the problem to be considered is how to impose prescribed prototype vectors* $\xi^1, \xi^2, ..., \xi^p$ *,as fixed points. The synaptic matrix* $W$ *should be interpreted as a sort of sign correlation matrix of the prototypes, In the classical approach. The weak point in this approach, comes from the fact that it does not have the appropriate tools to deal efficiently with the correlation between the state vectors and the prototype vectors The capacity of the net is very poor because one can only know if one given vector is adequately correlated with the prototypes or not and we are not able to know what its exact correlation degree. The interest of our approach lies precisely in the fact that it provides these tools. In this paper, a geometrical vision of the dynamic of states is explained. A fixed point is viewed as a point in the Euclidean plane* $R^2$*. The retrieving procedure is analyzed trough statistical frequency distribution of the prototypes. The capacity of the net is improved and the spurious states are reduced. In order to clarify and corroborate the theoretical results, together with the formal theory, an application is presented*

***Keywords***: *Learning Systems, Pattern Recognition, Graph Theory, Recurrent Neural Networks.*

## 1. Introduction

As is well known, a RNN is a discrete time, discrete-valued dynamic system which at any given instant of time *t* is characterized by a binary *state vector* $x(t) = \left[x_1(t), ..., x_i(t), ..., x_n(t)\right] \in \{1, -1\}^n$ . The behavior of the system is described by a *dynamic equation* of the type

$$x_i(t+1) = Sgn\left[\sum_{j=1}^{n} w_{ij} x_j(t) - \theta_i\right] \quad i = 1,2,...,n \tag{1}$$

A point $x$ is a fixed point if all its components remain unchanged when (1) is applied. The aim is to get the network parameters, namely the synaptic matrix $W$ and the threshold vector $\theta$, for which the prototype vectors $\xi^1, \xi^2, ..., \xi^p$, are fixed points. In our approach $x(t) = \{0,1\}^n$ and, as the components of $x(t)$ may only be zero or one, we will refer to them as the null and unit components in $x(t)$. Associated to the network there will be a complete graph $G$ , with $n$ vertices $\{v_1, ..., v_n\}$, and *one* bi-directional *edge* $a_{ij}$ for every possible pair of different vertices (1). Initially, at the *training stage* a null value $w_{ij}$ is assigned to every edge $a_{ij}$ in the graph; afterwards, when $\xi^\mu$ is presented to the net; the weight $w_{ij}$ is updated by:

$$\Delta w_{ij} = \begin{cases} +1 & if \quad \xi_i^\mu = \xi_j^\mu = 1,\ i \neq j, \\ -1 & if \quad \xi_i^\mu = \xi_j^\mu = 0,\ i \neq j, \\ 0 & \qquad otherwise. \end{cases} \tag{2}$$

This idea may be much more easily understood using the next *graphical interpretation* of the training algorithm: At the first step, a null value is assigned to all the edges $a_{ij}$ , then, when a learning pattern $\xi^\mu$ is acquired by the net, it is superposed over the graph *G.* The components $\{\xi_1^\mu, ..., \xi_n^\mu\}$, are going to be mapped over the vertices $\{v_1, ..., v_n\}$ of *G.* This mapping may be interpreted as a *coloring* of the edges in *G*, in such a way that, if

$\xi_i^\mu = \xi_j^\mu = 1$, the edge $a_{ij}$ (whose ending vertices are $v_i$ and $v_j$) will be colored with a certain color, for example *red*. On the other hand, if $\xi_i^\mu = \xi_j^\mu = 0$, then $a_{ij}$ will be colored with a different color, as for example *blue*. The rest of the edges in *G* remain uncolored. Once this coloring has been done, the value assigned over the, also *complete*, graph of *red* edges are positively reinforced and the value assigned over the edges of the *blue* graph are negatively reinforced. The value over the rest of the edges remains unchanged. Once the pattern $\xi^\mu$ is acquired, the colors are erased and we repeat the same color assignation with the next pattern to be acquired by the net, and so on. When every vector in the training pattern set has been integrated in the net, the training stage is finished, the *resulting graph G* has become edge-valued and its weight matrix is the *synaptic matrix* $W$ of the net.

## 2. Parameters of the Net

According with the theorem proved in [1], If any set $\xi^1, \xi^2, ..., \xi^p$ of prototype vectors are acquired by the net, then for any possible four different components $"i", "j", "r" \; y \; "s",$ then the relation: $w_{ij} + w_{rs} = w_{is} + w_{rj}$ is satisfied, and solving the system

$$\{w_{ij} = p_i + p_j, \; i \neq j \qquad (3)$$

(in *n* unknown $p_1, p_2, ..., p_n$) a solution and only a solution is obtained [1].. The *training algorithm* could be revisited in order to obtain the *weight vector* $\vec{p}$ without the necessity of obtaining the *weight matrix* $W$ first and then solving the system (3). In the *graphical interpretation* of the training algorithm, we may consider $\{p_i, p_j\}$ as the ending vertices of the generic edge $a_{ij}$. Just when the pattern $\xi^\mu$ has been acquired, if $a_{ij}$ has been colored by red, its weight $w_{ij}$ has been incremented by one. As $w_{ij} = p_i + p_j$ *j*, we may consider that $p_i$ and $p_j$ have both been incremented by *"½"*. If $n_1$ and $n_2$ are the number of *unit* and *null* components of $\xi^\mu$ and if $\xi_i^\mu = 1$, then obviously the number of red edges with one end in $p_i$ is equal to $(n_1 - 1)$. Consequently just when the pattern $\xi^\mu$ has been acquired $p_i$ has been incremented by $\frac{1}{2}(n_1 - 1)$. In the same way, it could be proved that if $\xi_i^\mu = 0$, then $p_i$ is incremented by $-\frac{1}{2}(n_0 - 1)$. The training algorithm in (3) could then be designed as follows:

$$\Delta p_i = \begin{cases} \frac{1}{2}(n_1 - 1) & if \quad \xi_i^\mu = 1 \\ -\frac{1}{2}(n_0 - 1) & if \quad \xi_i^\mu = 0 \end{cases} \qquad (4)$$

($n_1$ and $n_0$ are the number of *unit* and *null* components of $\xi^\mu$). When all the learning patterns have been acquired, the training is finished. Considering ½ a scale factor and since the inner product $\xi^\mu . \xi^\mu$ is equal to $n_1$ and the inner product $\bar{\xi}^\mu . \bar{\xi}^\mu$ is equal to $n_0$,, it can be interpreted that when the learning pattern $\xi^\mu$ is acquired the *weight vector* $\vec{p}$ is modified as follows:

$$\vec{p} \leftarrow \vec{p} + (\xi^\mu \; . \; \xi^\mu \; - \; 1) \cdot I - (\bar{\xi}^\mu \; . \; \bar{\xi}^\mu \; - \; 1) \cdot I \text{, where } \boldsymbol{I} \text{ is the unitary vector } (1,1,.., 1) \qquad (5)$$

The above expression realizes the *updating* of the weights $p_i$, for *i* from *1* to *n*, when $\xi^\mu$ is acquired. The computational *time* of the training algorithm, is then highly optimized.

### 2.1. Energy

The state vector *x* at time *t* could also be interpreted as a *coloring* of the edges in *G,* but now this coloring is going to be used to retrieve the stored data. If the graph *G* is colored with the coloring associated with the pattern $x(t)$, it is easy to understand (taking into account how the training algorithm was designed), that the bigger the summation of all the edges in the *red graph* and the lower the summation of all the edges in the *blue graph* are, then the more correlated the pattern $x(t)$ must be, with those that were used during the training stage. So, if *W* is the weight matrix of *G*, the *energy point EP* of the net is defined as a pair of numbers. The first of them represents the summation of all the values on the edges of the *red* graph, and the second one represents the same summation, but on the *blue* ones. So if *G* is colored with the color associated to *x(t)*, then *{I (t), O(t)}* may be defined as the pair of quadratic forms:

$$\begin{cases} I(t) = \tfrac{1}{2} x(t) \cdot W \cdot x(t)^t \\ O(t) = \tfrac{1}{2} \bar{x}(t) \cdot W \cdot \bar{x}(t)^t \end{cases} \tag{6}$$

If $n_1$ is the number of *unit* components of $x(t)$ and $n_0$ is the number of the *null* ones (in other words $n_1$ is the *Hamming distance* from *x(t)* to the *zero* vector). By other hand, it is obvious [1], that if $\{I_i(t), O_i(t)\}$, is the *EP*, when $W = (w_{ij})$, is the matrix with all its values equal to zero except those in file or the row *l*, then

$$I_i(t)\begin{cases} n_1 - 1 & \text{if } x_i(t) = 1 \\ 0 & \text{if } x_i(t) = 0 \end{cases}, \quad and \quad O\ (t)\begin{cases} n_0 - 1 & \text{if } x_i(t) = 0 \\ 0 & \text{if } x_i(t) = 1 \end{cases} \tag{7}$$

So, if $i_1, i_2, ..., i_{n_1}$ and $j_1, j_2, ..., j_{n_0}$ are the places where the *unit* and *null* components of *x(t)* are respectively located, the equations (13) could be written as

$$I(t) = (n_1 - 1)\left(p_{i_1} + ... + p_{i_{n1}}\right) \text{ and } O(t) = (n_0 - 1)\left(p_{j_1} + ... + p_{j_{n0}}\right) \tag{8}$$

Which means that

$$\frac{I(t)}{(n_1 - 1)} = \left(p_{i_1} + ... + p_{i_{n1}}\right) \ \ and \ \ \frac{O(t)}{(n_0 - 1)} = \left(p_{j_1} + ... + p_{j_{n0}}\right) \text{ and} \tag{9}$$

in other words

$$\frac{I(t)}{(n_1 - 1)} + \frac{O(t)}{(n_0 - 1)} = K \qquad being \ K = (p_1 + ... + p_n) \tag{10}$$

As all *state vectors* $x(t)$ with the same *Hamming distance "i"* to the zero vector contains the same numbers $n_1$ and $n_0$ of *unit* and *null* components, the *energy points* $\{I(t), O(t)\}$, associated to all of them, will be placed in the same line $r_i$ $r_{i,}$ whose equation expressed in *(x,y)* is

$$r_i \equiv (n_0 - 1)x + (n_1 - 1)y - (n_0 - 1)\cdot(n_1 - 1)\cdot K = 0 \tag{11}$$

In other words, all the *EP´s* associated with state vectors with the same number of *unit* components are placed in the same line of the *energy field*, and the equation of this line is the one represented in (11). In this way the *state vector space* is classified in as many classes as the dimension *n* of the space.

### 2.2. Dynamics

On the other hand, and as we said in the introduction, the nature of the algorithm here proposed let to know how the value of $x(t)$ affects the whole energy of the state $x(t)$. We may define the relative weight of the neuron i when the net is in state x(t) as the contribution of this neuron to the component $I(t)$, if $x_i(t)=1$; or as the contribution of this neuron to the component $O(t)$, if $x_i(t)=0$. So, if $x_i(t)=1$, we define the relative weight $w(t)$ of the neuron i when the net is in state $x(t)$ as:

$$w_i(t)=\frac{1}{n_1-1}+\frac{n_1-2}{n_1-1}\cdot\frac{p_i}{p.x(t)}=\frac{1}{x(t)\cdot x(t)-1}+\frac{x(t)\cdot x(t)-2}{x(t)\cdot x(t)-1}\cdot\frac{p_i}{p.xx(t)} \tag{12}$$

If in time *t* the state vector $x(t)$ is in class $[j]$, then for any *i* from *1* to *n*, the dynamic equation is defined as

$$x_i(t+1)=f_h\left[f_b\left(x_i(t)\right)\ .\ \left(w_i(t)-\theta_j\right)\right] \tag{13}$$

where $f_h$ is the Heaviside step function and $f_b$ is the function defined as $f_b(x)=2x-1$, which achieves the transformation from the domain $\{0,1\}$ to the domain $\{1,-1\}$

It can also be stated that the sum of the relative weights $w_i(t)$ for the unit components of $x(t)$ is equal to *2*. The same could be proved for the null components. We have then that the relative weight vector $w(t)$ associated to any state vector $x(t)$ may also be interpreted as a sort of frequency distribution of probabilities [2]. The reason is that

$$\sum_{i=1}^{n} w_i(t)=4 \quad\Rightarrow\quad \sum_{i=1}^{n}\tfrac{1}{4}w_i(t)=1 \tag{14}$$

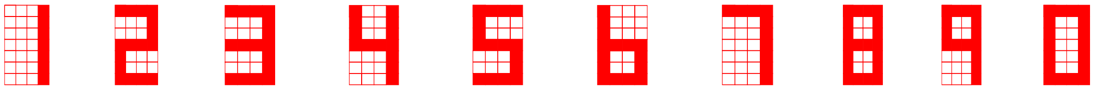
For any relative weight vector $w(t)$ The "uniform distribution vector" would be the one with all its components equal to $4\!/\!_n$. For any state $x(t)$ we could then define its deviation $D(x(t))$ as

$$D(x(t))=\sqrt{\sum_{i=1}^{n}\left[x_i(t)-\frac{4}{n}\right]^2} \tag{15}$$

The deviation of a given vector to the prototypes has been used for avoiding the parasite fixed points.

## 3. Application

We take, as an example for validating the performance of the algorithm we propose, the problem of the recognition of the Arabian digits as the prototype vectors:



Where the dimension *n*, of the pattern space is *28*, and

$$\begin{cases}\xi^1=[0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,1]\\ \xi^2=[1,1,1,1,0,0,0,1,0,0,0,1,1,1,1,1,1,0,0,0,1,0,0,0,1,1,1,1]\\ .\\ .\\ \xi^{10}=[1,1,1,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,,1,1,1,1]\end{cases}$$

and $p = 1/14\ \{53, 25, 25, 53, 11, -73, -73, 39, 11, -73, -73, 39, 39, 25, 25, 67, -17, -73, -73, 53, -17, -73, -73, 53, 11, 11, 11, 67\}$ .

In figure 1 the reader may see the energy lines and theirs associated PE´s. The Arabian digits are in this way placed on the lines: $r_7,\ r_{16},\ r_{16},\ r_{13},\ r_{16},\ r_{15},\ r_{10},\ r_{20},\ r_{15},\ r_{18}$. The associated PE´s are $1/7\{1113, -3710\}, 1/7\{3420, -2508\}$, $1/7\{4470, -3278\}, 1/7\{3210, -3745\}$, $1/7\{4050, -2970\}$, $1/7\{2821, -2418\}, 1/7\{2133, -4029\}$, $1/7\{5548, -2044\}$, $1/7\{4095, -3510\}$, $1/7\{4539, -2403\}$ , The problem now is how to obtain in an adaptive way the capacity parameters $\theta_1, \theta_2, \ldots, \theta_{28}$, in order to obtain the Arabian digits as fixed points with the least number of parasitic points as possible.

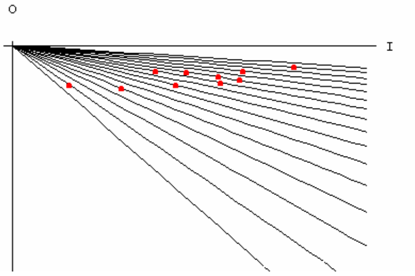


Figure 1 Arabian Digits Projections

When the dynamic equation in (13) is considered, a point $x(t)$ whose energy projection belongs to the $r_j$ line, is a fixed point if, and only if, the (capacity) parameter $\theta_j$ is an upper bound for all the relative weights $w_i(t)$ associated to the components of $x(t)$. Once the training has finished, the relative weight vector of the prototypes could then be calculated. If the energy projection of the prototype $\xi^\mu$ belongs to $r_j$ and the largest of the components of $w_i(t)$ is taken as $\theta_j$: it is clear that the prototype $\xi^\mu$ will be a fixed point. But the problem is how to avoid that points with high degree of correlation with a prototype but with all its relative weights components lower than the capacity parameter to skip away from this prototype. The idea proposes in this paper, made use of the deviation defined in (15). When, in time *t*, the dynamic equation is applied to a component of the vector $x(t)$, this component will change its state not only if the relative weight $w_i(t)$ is lower that the capacity parameter of its class. The deviation of the new state, in the case of change of sate, must be similar to the deviation of the prototypes in the new class. The degree of similarity may be measured by a coefficient □. The coefficient □□ is handled in a dynamical way (the more is the time the higher is the coefficient).

Besides the weight vector, there are other set of parameters of the net. For every one class $r_i$, the capacity parameter □i and the deviation of the prototypes in this class are obtained. So the algorithm control not only if the new state is strongly correlate with some prototype in its class, the algorithm also control that the components in the new state must, with a high degree of probability, be placed in similar places as some prototype of the class. We have applied with to our example, obtaining that almost all the points inside a neighborhood of radius 1, of the prototypes, are attracted by these prototypes. The 10 Arabian digits are fixed points of the system, and almost all the 28 neighbor of any one of them were attracted by its attractor prototype. In figure 2, the number of points inside a neighborhood of radius 1, of the prototypes are expressed.

$$\begin{cases} \boxed{24 \to 1}\ \boxed{23 \to 2}\ \boxed{25 \to 3}\boxed{22 \to 4}\boxed{27 \to 5} \\ \boxed{22 \to 6}\ \boxed{25 \to 7}\ \boxed{25 \to 8}\ \boxed{22 \to 9}\ \boxed{21 \to 0} \end{cases}$$

Figure 2. Prototypes belonging also to $r_{15}$

## 4. Conclusion

The weight parameters in the *Hopfield* network are not a free set of variables. They must fulfill a set of constrains which have been deduced trough a new re-interpretation of the net as *Graph Formalisms*. Making use of this constrains the *state-vector* has been classified in *n* classes according to the *n* different possible distances from any of the state-vectors to the *zero* vector. The $(n \times n)$ matrix of weights may also be reduced to a *n*-vector of weights. In this way the computational time and the memory space, required for obtaining the weights, is optimized and simplified. The degree of correlation from a pattern with the prototypes may be controlled by the dynamical value of two parameters: the capacity parameter $\theta$ which is used for controlling the capacity of the net (it may be proved that the bigger is the $\theta_j$ component of $\theta$, the lower is the number of fixed points located in the $r_j$ energy line) and the parameter $\mu$ which measures the deviation to the prototypes. A typical example has been exposed, the obtained results have proved to improve the obtained when the classical algorithm is applied.

## Bibliography

[1] V. Giménez-Martínez, *A Modified Algoritm Hopfield Auto-Associative Memory with Improved Capacity*, IEEE Transactions on Neural Networks, (in press), 2000.

[2] N. K. Bose and P. Liang. *Neural Network Fundamentals with Graphs, algorithms and Applications.* McGraw Series in Electrical and Computer Engineering, 1996.

[3] V. Giménez-Martínez, P. Gómez-Vilda, E. Torrano and M. Pérez-Castellanos, *A New Algorithm for Implementing a Recursive Neural Network,* Proc. of the IWANN´95 Torremolinos, Málaga, Spain, June, pp. 252-259, 1995.

[4] V. Giménez-Martínez, P. Gómez, M. Pérez- Castellanos, and E. Torrano, *A new approach for controlling the capacity of a Recursive Neural Network,* Proc of AMS´94. IASTED, Lugano, Suise, June, pp 90-93, 1994

## Authors' Information

***Víctor Giménez-Martínez*** *– Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: vgimenez@fi.upm.es*

***M Gloria Sánchez–Torrubia*** *– Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: gsanchez@fi.upm.es*

***Carmen Torres–Blanc*** *– Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo s.n., 28660 Boadilla del Monte, Madrid, Spain; e-mail: ctorres@fi.upm.es*

# THE FUZZY-NEURO CLASSIFIER FOR DECISION SUPPORT

## Galina Setlak

***Abstract:*** *This paper aims at development of procedures and algorithms for application of artificial intelligence tools to acquire, process and analysis various types of knowledge. The proposed environment integrates techniques of knowledge and decision process modeling such as neural networks and fuzzy logic-based reasoning methods. The problem of an identification of complex processes with the use of neuro-fuzzy systems is solved. The proposed classifier has been successfully applied for building one decision support systems for solving managerial problem.*

***Keywords****: artificial intelligence, artificial neural networks, fuzzy inference systems, classification, decision support.*

***ACM Classification Keywords****: I. Computing Methodologies, I.2 Artificial Intelligence*

## Introduction

A managerial decision support system must harness the information embedded in corporate data and apply this information to problem-solving processes of managers. Information systems required by the factories of the future must be capable of managing, maintaining and processing all forms of information required in the factory. Information is considered as being a vital resource of the enterprise because it represents its mind, because it is the basis for decision making and communication and it forms the basis for new designs.

The traditional artificial intelligence systems, mainly based on the symbolic paradigm, showed to be efficient tools for solving exactly and completely stated problems. However, they were ineffective for solving the real life problems that are described or represented by the imprecise, incomplete, uncertain and linguistic knowledge or by large amounts of numerical data collected in databases. The foregoing drawbacks of the symbolic paradigm based artificial intelligence systems have motivated many researches for creating new tools for designing intelligent decision support systems. As a result of those efforts the techniques named "computational intelligence" have been developed. They have been worked out as a joint of three methodologies: artificial neural networks, fuzzy logic and genetic algorithms. The artificial neural networks bring in the resulting system the ability for learning, generalizing and processing large amount of numerical data, the fuzzy logic allows the follow-on systems to represent and process inexact and uncertain information [Zadeh L.A., Kacprzyk J., 1992], and the genetic algorithm - as a global optimization tool - is used for strengthening the learning abilities of the resulting tool [Rutkowska D., M.Pilinski, L. Rutkowski, 1997]. As a result of joining the artificial neural networks (ANN), fuzzy logic, and genetic algorithms we get the system that is a synergistic combination of the three complementary technologies [Takagi H., 2000], [Rutkowska D., 2000].

Neural computing, genetic algorithms, and fuzzy systems are effective ways to deal with complex problems efficiently. Each method handles uncertainty and ambiguity differently, and these technologies can often be blended to utilize the features each, achieving impressive results. A combination of artificial neural networks and fuzzy logic can result in synergy that improves speed, fault tolerance, and adaptive ness. Fusion of neural networks and fuzzy inference systems have attracted the growing interest of researchers in various scientific and engineering areas due to the growing need of intelligent decision support systems to solve the real world problems. There are many real-world applications of intelligent systems integration [Takagi H., 2000], [Li S., 2000], [F. Wong, 1992], [D.Nauck, F. Klawonn, R.Kruse, 1997]. Each intelligent system can be a valuable component in a decision support system in which each technology can be used in series or in parallel. For instance, the neural network can identify classes of membership function for the fuzzy system [Jang R., Sun C.T., Mizutani E., 1997], [Takagi H., 2000].The genetic learning method can perform rule discovery in large databases, with the rules fed into the conventional expert system [Takagi H., 2000].

There are two directions of researches on systems that are built as a combination of neural networks and fuzzy logic based systems. The first one gives as results so-called fuzzy neural networks build of fuzzy neurons [Rutkowska D., M.Pilinski, L.Rutkowski, 1997]. The results of the second research course are neuro-fuzzy systems that use the artificial neural networks within the fuzzy logic systems framework. The most advanced types of the neuro-fuzzy systems are hybrid ones [Li S., 2000], [Rutkowska D., 2000].

The paper presents the neuro-fuzzy technologies which can be used for designing the rule-based intelligent decision support systems. In the paper a connectionist neuro-fuzzy system designed for classification problems is presented. The proposed classifier has been successfully applied for building one decision support systems for solving managerial problem. Example of classification problems solved by means of this hybrid intelligent system is illustrated.

## The Neuro-Fuzzy Method for Knowledge Modelling

Neural Networks can be used in constructing fuzzy inference systems in ways other than training. They can also be used for rule selection, membership function determination and in what we can refer to as hybrid intelligent systems.

Fuzzy systems that have several inputs suffer from the curse of dimensionality. In this paper will investigate and apply the Takagi-Hayashi method [Takagi H., Hayashi I., 1991] for the construction and tuning of fuzzy rules, this is commonly referred to as neural network driven fuzzy reasoning – NDF – method (see Fig.1). The NDF method is an automatic procedure for extracting rules and can greatly reduce the number of rules in a high dimensional problem, thus making the problem tractable.

The NDF method performs three major functions:

- Partitions the decision hyperspace into a number of rules. It performs this with a clustering algorithm.
- Identifies a rule's antecedent values (left hand side - LHS membership function). It performs this with a neural network.
- Identifies a rule's consequent values (right hand side - RHS membership function) by using a neural network with supervised training. This part necessitates the existence of target outputs.

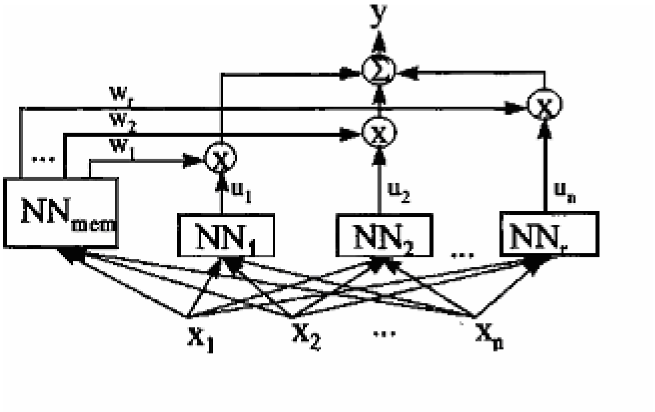


Fig.1. Neural Network Driven Fuzzy Reasoning [Takagi H., Hayashi I., 1991].

The above block diagram represents the NDF method of fuzzy rule extraction. This method uses a variation of the Sugeno fuzzy rule:

$$\textbf{IF } x_i \textbf{ is } A_i \textbf{ AND } x_2 \textbf{ is A2 AND ....AND } x_n \textbf{ is } A_n \textbf{ THEN } y=f(x_1, x_2, ..., x_n), \qquad (1)$$

where f(.) is a neural network model rather than a mathematical function. This results in a rule of the form:

$$\textbf{IF } x_i \textbf{ is } A_i \textbf{ AND } x_2 \textbf{ is } A_2 \textbf{ AND ....AND } x_n \textbf{ is } A_n \textbf{ THEN } y=NN(x_1, x_2, ..., x_n). \qquad (2)$$

The $NN_{mem}$ calculates the membership of the input to the LHS membership functions and outputs the membership values. The other neural networks form the RHS of the rules. The LHS membership values weigh the RHS neural network outputs through a product function. The altered RHS membership values are aggregated to calculate the NDF system output. The neural networks are standard feed forward multilayer perceptron designs.

The following 4 steps implement the NDF method. The NDF method also implements methods for reducing the neural network inputs to a small set of significant inputs and checking them for overfitting during training.

Step 1: The training data x is clustered into r groups: $R^1$, $R^2$,..., $R^s$ {s=l,2,...,r} with $n_t^s$ terms in each group. Note that the number of inference rules will be equal to r.

Step 2: The $NN_{mem}$ neural network is trained with the targets values selected as:

$$w_i^s = \begin{cases} 1, & x_i \in R^s \\ 0, & x_i \notin R^s \end{cases} \quad \textbf{i =1,2,..., } n_t^s \textbf{, S=1,2,...,r.} \qquad (3)$$

The outputs of $NN_{mem}$ for an input $x_i$ are labeled $w_{is,}$ and are the membership values of $x_i$ to each antecedent set $R^S$.

Step 3: The NNS networks are trained to identify the consequent part of the rules. The inputs are $\{x_{i1}^s, x_{i2}^s, ..., x_{im}^s\}$, and the outputs are $y_s$ , i = l,2,...,$n_{t.}$

Step 4: The final output value y is calculated with a weighted sum of the $NN_S$ outputs:

$$y_i = \frac{\sum_{s=1}^{r} \mu_{A^s}(x_i).\{U_s(x_i)\}_{inf}}{\sum_{s=1}^{r} \mu_{A^s}(x_i)}, \quad \textbf{i=1,2,...,n.,} \qquad (4)$$

where $u_s(x_i)$ is the calculated output of $NN_S$.

## The Neuro-Fuzzy Classifier for Decision Support

Neural networks are widely used as classifiers; see e.g. [Jang R., Sun C.T., Mizutani E., 1997], [Moon Y.B., Divers C.K., and H.-J.Kim, 1998], [Takagi H., 2000]. Classification and clustering problems has been addressed in many problems and by researchers in many disciplines like statistics, machine learning, and data bases. The basic algorithms of the classification methods are presented in [D.Nauck, F. Klawonn, R.Kruse, 1997], [Setlak G., 2004]. The application of the clustering procedure can be classified into one of the following techniques [Jang R., Sun C.T., Mizutani E., 1997] partition in which a set is divided into m subsets, when m is the input parameter:

- hierarchical form trees in which the leaves represent particular objects, and the nodes represent their groups. The higher level concentrations include the lower level concentrations. In terms of hierarchical methods, depending on the technique of creating hierarchy classes (agglomerative methods and divisive methods);
- graph-theoretic clustering,
- fuzzy clustering,
- methods based on evolutionary methods,
- methods based on artificial neural networks.

In this work two approaches have been applied to clustering of parts and assembly units. As basic method it was used Self Organizing Map (SOM) of Kohonen, a class of unsupervised learning neural networks, to perform direct clustering of parts families and assembly units. Self Organizing Maps are unsupervised learning neural networks which were introduced by T. Kohonen [Kohonen T., 1990] in the early ’80s. This type of neural network is usually a two-dimensional lattice of neurons all of which have a reference model weight vector. SOM are very well suited to organize and visualize complex data in a two dimensional display, and by the same effect, to create abstractions or clusters of that data. Therefore neural networks of Kohonen are frequently used in data exploration applications [Kohonen T., 1990], [Takagi H., 2000]. SOM have been applied to classification of machine elements in group technology [Setlak G., 2004].

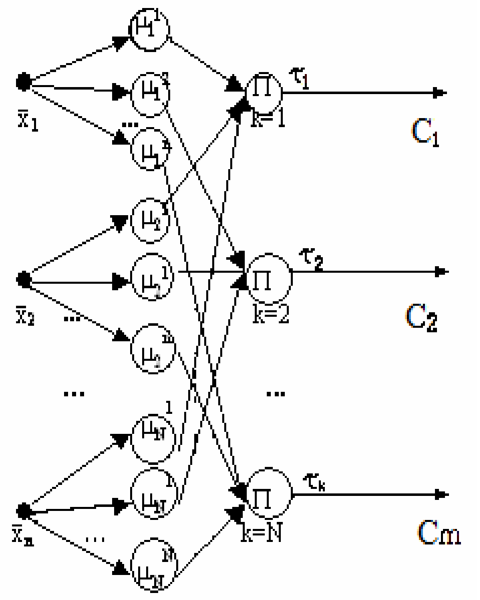


Fig.2. A neuro-fuzzy system for classyfication

The other approach applies fuzzy logic and fuzzy neural systems for classification problems. However, neural networks work as a "black box", which means that they produce classification results but do not explain their performance. Thus, we do not know the rules of classification. Neural network weights have no physical interpretation. Fuzzy and fuzzy neural systems can be employed in order to solve classification problems [Setlak G., 2000]. The neural-fuzzy systems are rule-based systems that realize fuzzy IF-THEN rules. Some of the major woks in this area are ANFIS [Jang, 1992], [Jang 1997], NEFCLASS [D.Nauck, F. Klawonn, R.Kruse, 1997], CANFIS

ANFIS (Adaptive Neuro-Fuzzy Inference System) [Jang, 1992], [Jang R., Sun C.T., Mizutani E., 1997] (Fig.1) is a network-structured adaptive fuzzy inference system which has found various applications including control, system identification, time series prediction, and noise cancellation. A common version of ANFIS uses normalized input fuzzy membership functions, product fuzzification, product inference, sum composition and Sugeno-type linear output functions (and thus needs no defuzzification). The system parameters are tuned using stochastic gradient descent method for the premise parameters and recursive least square method for the consequent parameters.

The CANFIS (Co-Active Neuro-Fuzzy Inference System) model integrates fuzzy inputs with modular neural network to quickly solve poorly defined problems. Fuzzy inference systems are also valuable as they combine the explanatory nature of rules (membership functions) with the power of "black box" neural networks.

Proposed hybrid intelligent system for classification can be presented and is shown in Fig. 2. They make a fuzzy inference based on a collection of fuzzy IF-THEN rules, called the rule base, described as follows:

A neuro-fuzzy system for classification can be presented in the form of the connectionist network shown in Fig. 2. The proposed Neuro-Fuzzy Classifier (NFC) is a hybrid neuro-fuzzy system that has a feed-forward network-like structure. The structure of the system expresses the fuzzy rules base that models the process of decision-making.

The classifier illustrated in Fig.2 reflects the following fuzzy classification rules:

$$\mathbf{R^{(k)}:}\ \text{IF}\ \mathbf{x_1\ is\ G_1^k}\ \text{and}\ \mathbf{x_2\ is\ G_2^k}\ \text{and. ....and}\ \mathbf{x_n\ is\ G_n^k}\ \text{THEN}\ (\mathbf{x} \in \mathbf{C_l}) \qquad \mathbf{(5)}$$

where $\boldsymbol{x} = [x_1, x_2 ..., x_n]^T$, and $x_i$, for $i$ = 1,2,..., $n$, are linguistic variables, $G_i^k$ is fuzzy sets for i-th input and k-th fuzzy rule, $C_l$, for l=1,2,...,$m$, are classes, N denotes the number of rules $R^{(k)}$, for $k$ = 1,..., $N$.

The crisp input values, presented in Fig.2, constitute the input vector: $\bar{x}=[\overline{x_1},\overline{x_2},...\overline{x_n}]^T$ .

The output values, $\tau_k$, for $k$ = 1, 2,..., N, represent degrees of rule activation [6], expressed as follows:

$$\tau_k = \prod_{i=1}^{n} \mu_i^k(\overline{x_i}), \qquad \mathbf{(6)}$$

where

$$\mu_i^k(x_i) = exp\left[-\left(\frac{x_i - \overline{X}_i^k}{\sigma_i^k}\right)^2\right] \qquad \mathbf{(7)}$$

is the Gaussian membership function, characterized by the center and width parameters, $\overline{x}_i^k$ and $\sigma_i^k$, respectively. The neuro-fuzzy network portrayed in Fig.2 performs a classification task based on the values of $\tau_k$, for $k$ = 1,..., $N$ . Each input vector $\bar{x} = [\overline{x_1},\overline{x_2},...,\overline{x_n}]^T$ is classified to the class $C_l$ (where l = 1,2,…,m), which is associated with the maximal degree of rule activation, that is $\max_k \{\tau_k\}$ .

There are five phases of designing the NFC system:

- Each input attribute is described by a number of fuzzy sets;
- The initial fuzzy rules base is determined;
- System training;
- Testing the system against test data;
- Pruning the system – removing "weak", superfluous fuzzy rules in order improve system's transparency.

An example of implementing this neuro-fuzzy classifier is given below.

## Example: international stock selection

The presented hybrid neuro-fuzzy system has been applied for building Intelligent Decision Support System (IDSS). As example of a hybrid neuro-fuzzy system we have chosen a method for deriving a stock portfolio plan.

An international investment company uses a hybrid neuro-fuzzy system to forecast the expected returns from stocks, cash, bonds, and other assets to determine the optimal allocation of assets. Because the company invests in global markets, it is first necessary to determine the creditworthiness of various countries, based on past and estimated performances of key socio-economic ratios, and then select specific stocks based on company, industry, and economic data. The final stock portfolio must be adjusted according to the forecast of foreign exchange rates, interest rates, and so forth, which are handled by a currency exposure analysis. The IDSS includes the following technologies:

- **Expert system.** The system provides the necessary knowledge for both country and stock selection (rule-based system).
- **Neural network.** The neural network conducts forecasting based on the data included in the database.
- **Fuzzy logic.** The fuzzy logic component supports the assessment of factors for which there are no reliable data. For example, the *credibility* of rules in the rule base is given only as a probability. Therefore, the conclusion of the rule can be expressed either as a probability or as a fuzzy membership degree.

The rule base feeds into IDSS along with data from the database. IDSS is composed of three modules: membership function generator (MFG), neuro-fuzzy inference system (NFIS), and neural network (NN). The modules are interconnected, and each performs a different task in the decision process.

Performance of a IDSS has been tested on the following input data:

- There are three input nodes (n = 3):

$X_1$ – risk of investment, it is defined by a linguistic term $G_i^k$, such as "high", "medium", "low".

$X_2$ – clear profit, also it is defined by a linguistic term $G_i^k$, such as "high", "medium" and "low".

$X_3$ – period refund of investment, it is defined by a linguistic term $G_i^k$, such as "long", "medium" and "short".

- The output values: are three classes $C_l$ , where $C_1$ – is defined by a linguistic term "very good investment", $C_2$ – "poor investment" and $C_3$ – "resign".
- The fuzzy set is characterized by a membership function $\mu_i^k(x_i)$: R → [0,1]. The membership functions for the fuzzy set, are expressed as (7).

Following the procedure described in section 2, the initial shapes of the fuzzy sets describing the input attributes were defined and the initial fuzzy rules base, containing 144 rules, was generated. The NFC method of using neural networks to generate the antecedent and consequent membership functions has been found to be useful and easily implemental with

Table 1. Results of the classification problem obtained using NFC and agglomerative methods

| N | Price | Advertising | Volume of sales | Stimulus of sale | Neuro-fuzzy Classifier | Agglomerative methods | Profit |
|---|---|---|---|---|---|---|---|
| 1 | 385,65 | 12000 | 227180 | 10000 | P1 | P1 | 87589967 |
| 2 | 397,24 | 10000 | 235090 | 6000 | P1 | P1 | 93371151,6 |
| 3 | 452,20 | 10000 | 217340 | 10000 | P4 | P4 | 98261148 |
| 4 | 478,92 | 12000 | 261280 | 8000 | P3 | P3 | 125112217,6 |
| 5 | 493,10 | 10000 | 184380 | 5000 | P2,P3 | P3 | 90902778 |
| 6 | 526,35 | 8000 | 147180 | 4000 | P2 | P2 | 77456193 |
| 7 | 583,24 | 5000 | 149300 | 3000 | P2,P3 | P2 | 87069732 |
| 8 | 594,93 | 5000 | 156520 | 4000 | P1 | P1 | 93109443,6 |
| 9 | 620,70 | 5000 | 121280 | 2000 | P2,P3 | P2 | 75271496 |
| 10 | 634,56 | 10000 | 116530 | 0 | P3,P4 | P4 | 73935276,8 |
| 11 | 663,20 | 2000 | 102160 | 0 | P2,P3 | P3 | 67750512 |
| 12 | 672,35 | 0 | 112510 | 0 | P1,P2,P3 | P3 | 75646098,5 |

## Conclusions

The single-layer neural network, proposed in this paper for classification, has the following features:

- Each neuron represents one fuzzy IF-THEN rule.
- The number of neurons equals to the number of rules in the rule base.
- Weights of the neurons have an interpretation concerning parameters of the membership functions of the corresponding neuro-fuzzy system.
- It is easy to modify the network architecture when a rule is added or removed (by addition or removal, respectively, the neuron that represents this rule).

Thus, in contrast to classical neural networks, the network proposed in this paper does not work as a "black box". This network is a rule-based neural network.

Single-layer neural networks can contain many neurons, so it is no problem to increase the number of rules in order to achieve better performance of the classifier. In this way, it is easy to incorporate learning from mistakes [Li S., 2000], by formulating new rules to avoid the mistakes. This method was applied in the classification problems depicted latter.

It is worth emphasizing that, in contrast to classical learning of neural networks, the method proposed in this paper allows the network to work without mistakes (based on the data set) and do not lose its generalization ability.

In the paper we have applied basic soft techniques for extracting rules and classification in a high dimensional managerial problem. The hybrid neuro-fuzzy system briefly presented in the paper was successfully applied for designing intelligent decision support system.

By using several advanced technologies (combination of fuzzy logic and neural networks) it is possible to handle a broader range of information and solve more complex problems. The research conducted proves that fuzzy neural networks are a very effective and useful instrument of implementation of intelligent decision support systems in management.

## Bibliography

[Jang, 1992] Jang R.: Neuro-Fuzzy Modeling: Architectures, Analyses and Applications, PhD Thesis, University of California, Berkeley, 1992.

[Jang R., Sun C.T., Mizutani E., 1997] Jang S.R., Sun C.T., Mizutani E.: Neurofuzzy and Soft Computing, Prentice-Hall, Upper Saddle River 1997, p. 245.

[Kohonen T., 1990] Kohonen T.: Self-organizing Maps, Proc. IEEE, 1990, 78, NR.9, pp. 1464-1480.

[Li S., 2000] Li S.: The Development of a Hybrid Intelligent System for Developing Marketing Strategy, Decision Support Systems, 2000, Vol 27, N4,

[Moon Y.B., Divers C.K., and H.-J. Kim, 1998] Moon Y.B., Divers C.K., and H.-J. Kim: AEWS: An Integrated Knowledge-based System with Neural Network for Reliability Prediction // Computers in Industry, 1998, Vol.35, N2, pp.312-344.

[D.Nauck, F. Klawonn, R.Kruse, 1997] D.Nauck, F. Klawonn, R.Kruse: Foundations of Neuro-Fuzzy Systems, J.Wiley&Sons, Chichester, 1997.

[Rutkowska D., 2000] Rutkowska D.: Implication-based neuro-fuzzy architectures.- Applied mathematics and computer science, V.10, N4, 2000, Technical Unieversity Press, Zielona Gora, 675-701.

[Rutkowska D., M.Pilinski, L. Rutkowski, 1997] Rutkowska D., M.Pilinski, L. Rutkowski: Sieci neuronowe, algorytmy genetyczne i systemy rozmyte, PWN, Warszawa, 1997 r., pp.411.

[Setlak G., 2004] Setlak G.: Intelligent Decision Support System, // LOGOS, Kiev, 2004, (in Rus.), pp. 250.

[Setlak G., 2000] Setlak G.: Neural networks in intelligent decision support systems for management // Journal of Automation and Information Sciences, Kiev, N1, 2000r., pp. 112-119.

[Takagi H., Hayashi I., 1991] Takagi H., Hayashi I.: NN-Driven Fuzzy Reasoning, //International Journal of Approximate Reasoning, 1991, Vol.3, p.1376-1389.

[Takagi H., 2000] Takagi H. Fusion technology of neural networks and fuzzy systems //International Journal of Applied mathematics and computer science, Zielona Gora, 2000, Vol.10, №4, pp.647-675.

[F. Wong, 1992] F. Wong: "Neural Networks, Genetic Algorithms, and Fuzzy Logic for Forecasting," Proceedings, International Conference on Advanced Trading Technologies, New York, July 1992, pp.504-532.

[Zadeh L.A., Kacprzyk J., 1992] Zadeh L.A., Kacprzyk J.(ed.): Fuzzy logic for the Management of Uncertainty, Wiley, New York, 1992, p. 492.

## Authors' Information

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

# A VARIANT OF BACK-PROPAGATION ALGORITHM FOR MULTILAYER FEED-FORWARD NETWORK

**Anil Ahlawat, Sujata Pandey**

***Abstract**: In this paper, a variant of Backpropagation algorithm is proposed for feed-forward neural networks learning. The proposed algorithm improve the backpropagation training in terms of quick convergence of the solution depending on the slope of the error graph and increase the speed of convergence of the system. Simulations are conducted to compare and evaluate the convergence behavior and the speed factor of the proposed algorithm with existing Backpropagation algorithm. Simulation results of large-scale classical neural-network benchmarks are presented which reveal the power of the proposed algorithm to obtain actual solutions.*

***Keywords**: Backpropagation, convergence, feed-forward neural networks, training.*

## Introduction

Feed-forward neural networks (FNN) have been widely used for various tasks, such as pattern recognition, function approximation, dynamical modeling, data mining, and time series forecasting, [1-3]. The training of FNN is mainly undertaken using the back-propagation (BP) based learning. The back-propagation algorithm has been investigated many times with minor variations [4-7]. However, even to date, there are still a great number of problems that cannot be solved efficiently by the majority of the training algorithms that have been proposed over the years, using standard simple feed-forward network architectures. A number of different kind of BP based learning algorithms, such as an on-line neural-network learning algorithm for dealing with time varying inputs [8], fast learning algorithms based on gradient descent of neuron space [9], second derivative based non-linear optimization methods [10], conjugate gradient methods [11] and genetic algorithms [12,13] avoid use of any gradient information. Levenberg–Marquardt algorithm [14-16] is the most powerful and a popular second derivative based algorithm that have been proposed for the training of feed-forward networks which combines the excellent local convergence properties of Gauss-Newton method near a minimum with the consistent error decrease provided by (a suitably scaled) gradient descent faraway from the solution. In first-order methods (such as gradient descent), a local minimizer problem is overshooted with the inclusion of momentum term. The momentum term actually inserts second-order information in the training process and provides iterations whose form is similar to the conjugate gradient (CG) method. The major difference of Backpropagation with the conjugate gradient method is that the coefficients regulating the weighting between the gradient and the momentum term are heuristically selected in BP, whereas in the CG algorithm these coefficients are adaptively determined. However, these algorithms also share problems [17] present in the standard Backpropagation algorithm and may converge faster in some cases and slower in others. Comparison of the speeds of convergence of different schemes for implementing Backpropagation is not clear-cut, though a discussion on benchmarking of the algorithms can be found [18].

In this paper, a proposal for a variant of back-propagation algorithm for FNN with time-varying inputs has been presented which is capable of overcoming the shortcomings of the BP as discussed above. In the experimental section, the proposed algorithm is compared with the existing Backpropagation algorithm for training multilayer feed-forward networks on training tasks that are well known for their complexity. It was observed that the proposed algorithm have shown to solve these tasks with exceptionally high success rates and converged much faster than the original BP algorithm and showed greater accuracy.

## Backpropagation Algorithm

### *Overview of the Algorithm*

The Backpropagation training algorithm [1] for training feed-forward networks was developed by Paul Werbos [7], and later by Parker [4] and Rummelhart [5]. This type of network configuration is the most common in use, due to its ease of training [19]. It is estimated that over 80% of all neural network projects in development use back-propagation. In back-propagation, there are two phases in its learning cycle, one to propagate the input pattern through the network and the other to adapt the output, by changing the weights in the network. It is the error signals that are back propagated in the network operation to the hidden layer(s). The portion of the error signal that a hidden-layer neuron receives in this process is an estimate of the contribution of a particular neuron to the output error. Adjusting on this basis the weights of the connections, the squared error, or some other metric, is reduced in each cycle and finally minimized, if possible.

A Back-Propagation network consists of at least three layers of units: an input layer, at least one intermediate hidden layer, and an output layer. Typically, units are connected in a feed-forward fashion with input units fully connected to units in the hidden layer and hidden units fully connected to units in the output layer. When a Back-Propagation network is cycled, an input pattern is propagated forward to the output units through the intervening input-to-hidden and hidden-to-output weights.

### *Training in Backpropagation Algorithm*

The feed-forward back-propagation network undergoes supervised training [20], with a finite number of pattern pairs consisting of an input pattern and a desired or target output pattern. An input pattern is presented at the input layer. The neurons here pass the pattern activations to the next layer neurons, which are in a hidden layer. The outputs of the hidden layer neurons are obtained by using a bias, and also a threshold function with the activations determined by the weights and the inputs. These hidden layer outputs become inputs to the output neurons, which process the inputs using an optional bias and a threshold function. The final output of the network is determined by the activations from the output layer.

The computed pattern and the input pattern are compared, a function of this error for each component of the pattern is determined, and adjustment to weights of connections between the hidden layer and the output layer is computed. A similar computation, still based on the error in the output, is made for the connection weights between the input and hidden layers. The procedure is repeated with each pattern pair assigned for training the network. Each pass through all the training patterns is called a cycle or an epoch. The process is then repeated as many cycles as needed until the error is within a prescribed tolerance. The adjustment for the threshold value of a neuron in the output layer is obtained by multiplying the calculated error in the output at the output neuron and the learning rate parameter used in the adjustment calculation for weights at this layer.

After a Back-Propagation network has learned the correct classification for a set of inputs from a training set, it can be tested on a second set of inputs to see how well it classifies untrained patterns. Thus, an important consideration in applying Back-Propagation learning is how well the network generalizes.

### *Mathematical Analysis of the Algorithm*

Assume a network with N inputs and M outputs. Let $x_i$ be the input to $i^{th}$ neuron in input layer, $B_j$ be the output of the $j^{th}$ neuron before activation, $y_j$ be the output after activation, $b_j$ be the bias between input and hidden layer, $b_k$ be the bias between hidden and output layer, $w_{ij}$ be the weight between the input and the hidden layers, and $w_{jk}$ be the weight between the hidden and output layers. Let ƞ be the learning rate and ƍ the error. Also, let i, j and k be the indexes of the input, hidden and output layers respectively.

The response of each unit is computed as:

$$B_j = b_j + \sum_{i=1}^{n} x_i \cdot w_{ij} \tag{1}$$

$$y_j = \left(1/\left(1 + \exp(-B_j)\right)\right) \tag{2}$$

Weights and bias between input and hidden layer are updated as follows:

For input to hidden layer, for i = 1 to n,

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_j y_i \tag{3}$$

$$b_j(t+1) = b_j(t) + \eta\delta_j \tag{4}$$

Where, $\delta_j$ is the error between the input and hidden layer and calculated as:

$$\delta_j = y_j \cdot (1 - y_j) \cdot \sum_k \delta_k w_{jk} \tag{5}$$

Weights and bias between hidden and output layer are updated as follows:

For hidden to output layer, for j = 1 to h,

$$w_{jk}(t+1) = w_{jk}(t) + \eta\delta_k y_j \tag{6}$$

$$b_k(t+1) = b_k(t) + \eta\delta_k \tag{7}$$

and $\delta_k$ is the error between the hidden and output layer and calculated as:

$$\delta_k = y_k \cdot (1 - y_k) \cdot (d_k - y_k) \tag{8}$$

## Proposed variant of Backpropagation Algorithm

The Backpropagation algorithm described above has many shortcomings [17]. The time complexity of the algorithm is high and it gets trapped frequently in sub-optimal solutions. It is also difficult to get an optimum step size for the learning process, since a large step size would mean faster learning, which may miss an optimal solution altogether, and a small step size would mean a very high time complexity for the learning process. The proposed variant of the Backpropagation algorithm aims to overcome some of these shortcomings.

*Overview of the Proposed Algorithm*

The Backpropagation Algorithm described above is modified by following changes:

1 Momentum: A simple change to the training law that sometimes results in much faster training is the addition of a momentum term [21]. With this change, the weight change continues in the direction it was heading. This weight change, in the absence of error, would be a constant multiple of the previous weight change. The momentum term is an attempt to try to keep the weight change process moving, and thereby not gets stuck in local minima’s. In some cases, it also makes the convergence faster and the training more stable.

2 Dynamic control for the learning rate and the momentum: Learning parameters such as Learning rate and momentum serve a better purpose if they can be changed dynamically during the course of the training [21]. The learning rate can be high when the system is far from the goal, and can be decreased when the system gets nearer to the goal, so that the optimal solution cannot be missed.

3 Gradient Following: Gradient Following has been added to enable quick convergence of the solution depending on the slope of the error graph. When the system is far away from the solution, the learning rate is further increased by a constant parameter C1 and when the system is close to a solution, the learning rate is decreased by a constant parameter C2. The farness or closeness of the system from the solution was determined from the slope of the Error graph [22-26].

4 Speed Factor: To increase the speed of convergence of the system, a speed factor S has been used as determined by a mathematical formula derived from the study of graphs.

*Mathematical Analysis of the Algorithm*

1. Momentum: Let the momentum term be α. Then equation (3) and equation (4) would be modified as:

For input to hidden layer, for i = 1 to n,

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_j y_i + \alpha \cdot (w_{ij}(t) - w_{ij}(t-1)) \quad (9)$$

$$b_j(t+1) = b_j(t) + \eta\delta_j + \alpha \cdot (b_j(t) - b_j(t-1)) \quad (10)$$

The term $\delta_j$ would be calculated as in equation (5).

The equation (6) and equation (7) would be modified as:

For hidden to output layer, for j = 1 to h,

$$w_{jk}(t+1) = w_{jk}(t) + \eta\delta_k y_j + \alpha \cdot (w_{jk}(t) - w_{jk}(t-1)) \quad (11)$$

$$b_k(t+1) = b_k(t) + \eta\delta_k + \alpha \cdot (b_k(t) - b_k(t-1)) \quad (12)$$

The term $\delta_k$ would be calculated as in equation (8).

2. Dynamic Control for learning rate and momentum: If changing the weight decreases the cost function (mean squared error), then the learning rate is given by

$$\eta = \eta + 0.05 \quad (13)$$

Else

$$\eta = \eta - 0.05 \quad (14)$$

Similar conditions were placed for the momentum term α.

3. Gradient Following: Let C1 and C2 be two constants, such that C1 > 1 and 0 < C2 < 1 and $\Delta$max and $\Delta$min be the maximum and minimum change permissible for the weight change. If $\frac{\partial E}{\partial w}$ is the gradient following term then three cases need to be considered:

a) Case I

If $\frac{\partial E}{\partial w}(t) \cdot \frac{\partial E}{\partial w}(t-1) > 0$

Then

$$\Delta(t) = \min(\Delta(t) \,.\, C1, \Delta\text{max}) \quad (15)$$

$$\Delta w = -\frac{\partial E}{\partial w}(t) \cdot \Delta(t) \quad (16)$$

$$w(t+1) = w(t) + \Delta w \quad (17)$$

b) Case II

If $\frac{\partial E}{\partial w}(t) \cdot \frac{\partial E}{\partial w}(t-1) < 0$

Then

Δ(t) = min(Δ(t) . C2, Δmin) (18)

$$\frac{\partial E}{\partial w}(t) = 0 \tag{19}$$

w(t+1) = w(t) - Δw (20)

c) Case III

If $\frac{\partial E}{\partial w}(t) \cdot \frac{\partial E}{\partial w}(t-1) = 0$

Then

$$\Delta w = -\frac{\partial E}{\partial w}(t) \cdot \Delta(t) \tag{21}$$

w(t+1) = w(t) + Δw (22)

4. Speed Factor: Let S be the speed factor. Then equation (9) and equation (10) would further be modified to:

For input to hidden layer, for i = 1 to n,

$$w_{ij}(t+1) = w_{ij}(t) + S\left[\eta\delta_j y_i + S \cdot \alpha \cdot (w_{ij}(t) - w_{ij}(t-1))\right] \tag{23}$$

$$b_j(t+1) = b_j(t) + S\left[\eta\delta_j + S \cdot \alpha \cdot (b_j(t) - b_j(t-1))\right] \tag{24}$$

Similarly, equation (11) and (12) would be modified as:

For hidden to output layer, for j=1 to h,

$$w_{jk}(t+1) = w_{jk}(t) + S\left[\eta\delta_k y_j + S \cdot \alpha \cdot (w_{jk}(t) - w_{jk}(t-1))\right] \tag{25}$$

$$b_k(t+1) = b_k(t) + S\left[\eta\delta_k + S \cdot \alpha \cdot (b_k(t) - b_k(t-1))\right] \tag{26}$$

## Experimental Study

The algorithm proposed in this paper were tested on the training of standard multilayer feed forward networks (FNNs) and applied to several problems. The FNN simulator was implemented in Visual Basic .NET. The performance of the proposed algorithm was compared to existing Backpropagation algorithm. All simulations were carried out on a Pentium IV 2 GHz with 128 MB RAM PC using the FNN simulator developed by our team.

| Number of cycles | BP (time in msec) | Speed1 (time in msec) for momentum = 0.1 and speed = 0.1 | Speed2 (time in msec) for momentum = 0.1 and speed = 0.2 | Speed3 (time in msec) for momentum = 0.2 and speed = 0.1 |
|---|---|---|---|---|
| 100 | 42781.52 | 4626.66 | 4927.08 | 5668.15 |
| 300 | 123968.25 | 9333.43 | 9754.03 | 10094.51 |
| 500 | 206146.42 | 15442.20 | 15452.21 | 15552.36 |
| 800 | 330385.06 | 24204.80 | 25666.91 | 24585.36 |
| 1000 | 414546.10 | 30173.39 | 31054.64 | 30383.69 |
| 1200 | 496964.60 | 35671.28 | 36612.64 | 36712.79 |
| 1500 | 617187.47 | 44954.65 | 46096.28 | 46076.26 |

Table 1: Comparison of training time between Backpropagation algorithm and proposed algorithm for different momentum and speed for 8-bit parity problem.

The selection of initial weights is important in feed-forward neural network training. If the initial weights are very small, the backpropagated error is so small that practically no change takes place for some weights, and therefore more iteration are necessary to decrease the error. If the error remains constant, then the learning stops in an undesired local minimum. Large values of weights, results in speed up of learning, but they can lead to saturation and to flat regions of the error surface where training is slow. Keeping these in consideration, the experiments were conducted using the same initial weight vectors that have been randomly chosen from a uniform distribution in (-1,1). Sensitivity of the algorithm in some other intervals (-0.1,0.1) was also studied to investigate its convergence behavior.

| Number of cycles | BP (time in msec) | Speed1 (time in msec) for momentum = 0.1 and speed = 0.1 | Speed2 (time in msec) for momentum = 0.1 and speed = 0.2 | Speed3 (time in msec) for momentum = 0.2 and speed = 0.1 |
|---|---|---|---|---|
| 100 | 2072.9856 | 1482.1376 | 1412.0320 | 1442.0736 |
| 300 | 4256.1152 | 2022.9120 | 1832.6400 | 1592.2944 |
| 500 | 6399.2064 | 2022.9120 | 2563.6864 | 2363.3920 |
| 800 | 9183.2064 | 2022.9120 | 3815.4880 | 3585.1648 |
| 1000 | 11156.0448 | 2022.9120 | 3945.6768 | 3895.6032 |
| 1200 | 13038.7584 | 2022.9120 | 4746.8288 | 5097.3184 |
| 1500 | 16714.0352 | 2022.9120 | 5808.3584 | 5588.0320 |

Table 2: Comparison of training time between Backpropagation algorithm and proposed algorithm for different momentum and speed for Hurst Motor.

The initial learning rate was kept constant for both algorithms. It was chosen carefully so that the Backpropagation training algorithm rapidly converges without oscillating toward a global minimum. Then all the other learning parameters were tuned by trying different values and comparing the number of successes exhibited by five simulation runs that started from the same initial weights.

To obtain the best possible convergence, the momentum term and the speed constant are normally adjusted by trial and error or even by some kind of random search. Since the optimal value is highly dependent on the learning task, no general strategy has been developed to deal with this problem. Thus, the optimal value of these two terms is experimental but depends on the learning rate chosen. In our experiments, we have tried eight different values for the momentum ranging from 0.1 to 0.8 and for speed constant, we have tried five different values ranging from 0.1 to 0.5 and we have run five simulations combining all these values with the best available learning rate for BP. But it was shown that some combinations give better results, which is shown in Table 1 for 8-bit parity problem and in Table 2 for Hurst motor. On the other hand, it is well known that the “optimal” learning rate must be reduced when momentum is used. Thus, we also tested combinations with reduced learning rates.

Table 1 shows the results of training on 8-8-1 network (eight inputs, one hidden layer with eight nodes and one output node) on the 8-bit parity problem. It can be observed that training is considered successful for the given dataset for speed constant and momentum in table 1. It can be seen from the table 1 that training time is drastically reduced in the proposed algorithm. Figure 1 shows the variation of training time with number of cycles (epoch) for the Backpropagation and the three different cases for the proposed algorithm for 8-bit parity problem. In BP for increase in number of cycles, the training time increases rapidly but in all the cases for the proposed speed algorithm the training time increases gradually. Also for the change in the momentum and speed term, there was not much difference in the training time.

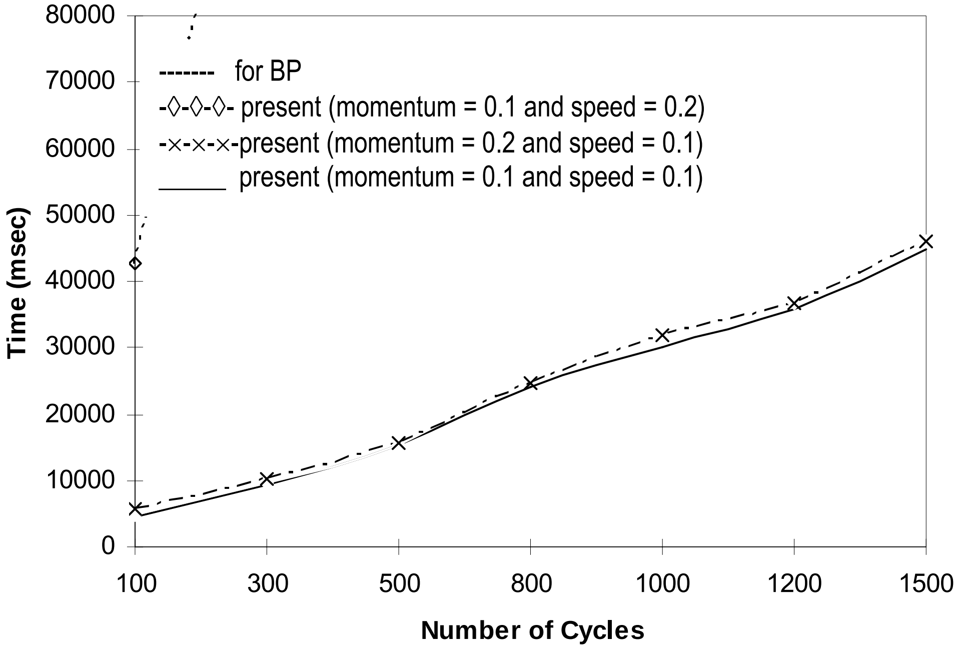


Fig. 1. Variation of time with number of cycles for 8-bit parity problem.

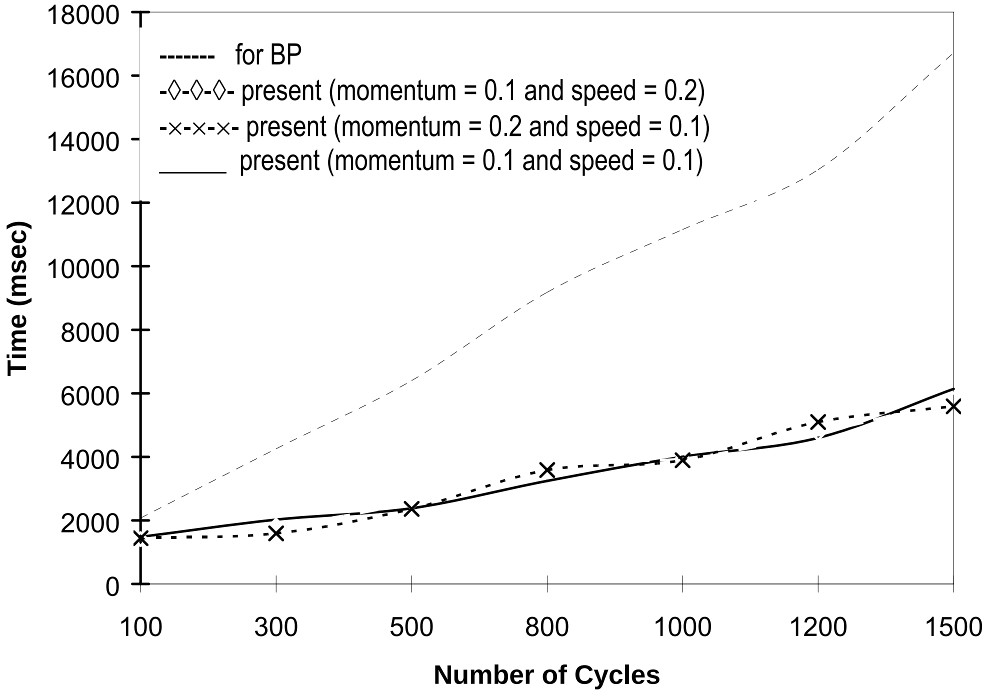


Fig. 2. Variation of time with number of cycles for Hurst Motor.

Table 2 shows the results of training on 5-8-2 network (five inputs, one hidden layer with eight nodes and two output node) for the Hurst motor. It was observed that training is also considered successful for the given dataset for speed constant and momentum in table 2. It can also be seen from the table 2 that training time is drastically reduced in the proposed algorithm. Figure 2 shows the variation of training time with number of cycles (epoch) for the Backpropagation and the three different cases for the proposed algorithm for Hurst motor. In BP for increase in number of cycles, the training time increases rapidly but in all the cases for the proposed speed algorithm the training time increases gradually. Also for the change in the momentum and speed term, there was not much difference in the training time.

## Conclusion

The variant in BP has been proposed for the training of feed forward neural networks. The convergence properties of both algorithm have been studied and the conclusion was reached that new algorithm is globally convergent. The proposed algorithm was tested on available training tasks. These results point to the conclusion that the proposed methods stand as very promising new tools for the efficient training of neural networks in terms of time. It also proves to be much more accurate than the existing Backpropagation algorithm. In addition the error correction rate achieved is much faster and training time is also much faster as shown in the results.

The proposed variant has a lower slope signifying a faster training compared to the Backpropagation algorithm and it converges to a more stable solution thereby ending the learning process. The Backpropagation algorithm on the other hand, may not converge at all throughout the learning process.

## Bibliography

1 J. M. Zurada, “Introduction to artificial neural systems,” *M. G. Road, Mumbai: Jaico,* (2002).

2 P. Mehra and B. W. Wah, “Artificial neural networks: concepts and theory,” *IEEE Comput. Society Press*, (1992).

3 Bob Waterbury, “Artificial intelligence expands frontiers in asset management, condition monitoring and predictive maintenance systems make AI pay off with lower maintenance costs, improved forecasting, and fewer unplanned shutdowns,” *Putman Media*, (November 16, 2000).

4 D.B Parker, “Learning logic, technical report TR-47,” *Center for Computational Research in Economics and Management Sciences*, MIT, (1985).

5 D.E. Rumelhart, G.E Hinton, and R.J. Williams, “Learning internal representations by error propagation,” *Parallel Distribution Processing*, vol. 1, pp. 318-362. MIT Press, Cambridge, MA (1986).

6 P.J. Werbos, “The roots of backpropagation,” *John Wiley and Sons, Inc.*, New York (1994).

7 T. Nitta, “An analysis on decision boundaries in the complex-backpropagation network,” *IEEE World Congress on Computational Intelligence*, vol. 2, pp. 934-939, Orlando, FL, June 1994, IEEE Computer Society Press.

8 Y. Zhao, “On-line neural network learning algorithm with exponential convergence rate,” *Electron. Lett.*, vol. 32, no. 15, pp. 1381–1382 (July 1996).

9 G. Zhou and J. Si, “Advanced neural network training algorithm with reduced complexity based on Jacobian deficiency,” *IEEE Trans. Neural Networks*, vol. 9, pp. 448–453 (May 1998).

10 D.B. Parker, “Optimal algorithms for adaptive networks: second order backpropagation, second order direct propagation and second order Hebbian learning,” *First IEEE International Conference on Neural Networks*, San Diego, pp. 593-600 (1987).

11 E.M. Johansson, F.U. Dowla, and D.M. Goodman, “Backpropagation learning for multi-layer feed-forward neural networks using the conjugate gradient method,” *Intl. J. Neural Systems*, vol. 2: pp. 291-301 (1992).

12 D.J. Mortana and L. Davis, “Training feed-forward networks using genetic algorithms,” *In Proceedings of 11th Intl. Joint Conf. in Artificial Intelligence (IJCAI)*, Detroit, MI, pp. 762-767, Morgan Kaufmann, San Mateo, CA (1989).

13 D. Whitley and T. Hanson, “Optimizing neural networks using faster, more accurate genetic search,” *Proceedings of 3rd Intl. Conf. Genetic Algorithms*, pp 391-396, Morgan Kaufmann, San Mateo, CA (1989).

14 R. Parisi, E. D. Di Claudio, G. Orlandi, and B. D. Rao, “A generalized learning paradigm exploiting the structure of feed-forward neural networks,” *IEEE Trans. Neural Networks*, vol. 7, pp. 1450–1459, Nov. (1996).

15 M. T. Hagan and M. B. Menhaj, “Training feed-forward neural networks with the Marquardt algorithm,” *IEEE Trans. Neural Networks*, vol. 5, pp. 989–993, Nov. (1994).

16 X. Yu, M. O. Efee, and O. Kaynak, “A general backpropagation algorithm for feed-forward neural networks learning,” *IEEE Trans. Neural Networks*, vol. 13, no. 1, pp. 251–254 (January 2002).

17 S. Saarinen, R. Brambley, and G. Cybenko, “Ill-conditioning in neural network training problems,” *SAIM J. Sci. Comput.*, 14(3):693-714 (May 1993).

18 S. E. Fahlman, "Faster-Learning variations on backpropagation: an empirical study in D. Touretzky, G. Hinton, and T. eds. Sejnowski," *Proceedings of the 1988 Connectionist Models Summer School*, Morgan Kaufmann, 38-51(1998).

19 J.S. Judd, "Neural Network Design and the complexity of Learning," *MIT Press*, Cambridge, MA (1990).

20 D. R. Hush and B. G. Horne, "Progress in supervised neural networks," *IEEE Signal Processing Magzine*, vol. 10, no. 1, pp. 8-39 (1993).

21 R.A. Jacobs, "Increased rate of convergence through learning rate adaptation," *Neural Networks*, pp 295-307 (1988).

22 P.R. Adby and M.A.H. Dempster, "Introduction to Optimization Methods," *Haisted Press*, New York (1974).

23 L. Cooper and D. Steinberg, "Introduction to Methods of Optimization," *Saunders, Philadelphia* (1970).

24 R.W. Daniels, "An introduction to numerical methods and optimization techniques," *North-Holland*, New York, 1978.

25 Karl-Heinz Elster, editor, "Modern Mathematical Methods of Optimization," *VCH*, New York (1993).

26 R. Fletcher, "Practical Methods of Optimization," *2nd ed. Wiley*, New York (1987).

## Authors' Information

***Anil Ahlawat, Sujata Pandey*** *– Department of Computer Science and Engineering, Amity School Of Engineering and Technology, 580, Delhi Palam Vihar Road, Bijwasan, New Delhi, 110061, India; e-mail:* ***a_anil2000@yahoo.com***

# STUDY WITH NEURAL NETWORKS OF RELATIONSHIPS BETWEEN DIFFERENT BERRY COMPONENTS IN GRAPES

**Angel Castellanos, Marita Esteban, Ana Martinez, Valentin Palencia**

***Abstract****: The impact of water availability on Vitis vinifera L. cv. Tempranillo grape yields and juice composition over a three-year period was studied. Grape juice composition during the different stages of berry growth was compared. The analytical data collected were used to investigate the relationships between some of the different components studied in these berries during the ripening period. Our goal is to study new neural networks models with analysis of sensibility in order to predict more accurately the relationship existing between them.*

***Keywords****: Neural Networks, clustering, Vitis vinifera, grapes, sugars, organic acids, mineral elements.*

***ACM Classification Keywords****: C.1.3, I.5.2*

## Introduction

The purpose of irrigation is to offset crop water deficits and thereby maximize yields and must quality, to increase profits [Rühl EH and Alleweldt G, 1985]. There are many regions with dry summers in Spain in which irrigation is an effective mean of regulating water availability to grape vines.

As has previously been noted by other workers [Williams LE and Matthews MA, Grapevine, 1990], irrigation of grape vines affects vine physiology, which may directly or indirectly affect yield and grape composition (°Brix, pH, total acidity, etc.) two aspects that also influence wine quality. There is considerable controversy in the literature concerning the positive and negative effects of vine irrigation on must and wine quality [Van Zyl JL, 1984]. Response to irrigation will depend upon such factors as harvest time, crop load, soil water availability and primarily summer rainfall.

Sugar concentration is used as an indicator of fruit maturity, being glucose and fructose the principal sugars in grape juices [Ough CS and Amerine MA, 1988]. Irrigation has a variable effect on sugar accumulation in the berries, and an increase, a decrease, or no change in sugar concentration have all been observed [Hardy

P.J.,1968]. The sugar to acidity ratio parameter is ordinarily useful in evaluating the ripening period [Ribéreau-Gayon J, Peynaud E, Ribéreau-Gayon P and Sudraud P, 1975]. Both titratable acidity and pH are of great importance for grape juice stability and are parameters commonly used as an indicator of quality. This is because the concentration of organic acids does not only contributes to the acid taste of the must but also influences subsequent wine color and microbiological stability [Boulton RB, 1980]. According to Hrazdina et al. [Hrazdina G, Parsons GF and Mattick LR, 1984] changes in the pH of grape berries are caused by the metabolism of the major acids and the accumulation of cations, which transform free acids into their corresponding salts. Some authors [McCarthy MG, Cirami RM and McCloud P, 1983 ], [Romero EG, Muñoz GS and Ibañez MDC, 1993] have stated that decreases in titratable acidity are primarily due to losses in malic acid concentration and to the formation of potassium salts. Potassium is the main mineral cation in grapes [Peynaud E and Ribéreau-Gayon J, 1971], and is predominantly involved in neutralization of tartaric acid and malic acid in the berries, thereby affecting the acid characteristics of the grapes [Hale CR, 1977].

The object of the present study was to ascertain whether irrigation, which has a quantitative effect on the values of the different components analysed in berries of the Tempranillo (Vitis vinífera L.) grape variety, though that effect is not always significant [Esteban MA, Villanueva MJ and Lissarrague JR, 1999], [Esteban MA, Villanueva MJ and Lissarrague JR, 2001] affects the relationships between the different components considered.

We use neural networks models with analysis of sensibility. This model predict more accurately the relationship existing.

Neural networks can predict any continuous relationship between inputs and the target. Similar to linear or non-linear regression, artificial neural networks develop a gain term that allows prediction of target variables for a given set of input variables. Physical–chemical relationships between input variables and target variables may or may not built into the association of target and input variables.

Neural networks [Anderson, James A. and Edward Rosenfield., 1988] are non-linear systems whose structure is based on principles observed in biological neuronal systems [Hanson, Stephen J. and David J. Burr. 1990]. A neural network could be seen as a system that can be able to answer a query or give an output as answer to a specific input. The in/out combination, i.e. the transfer function of the network is not programmed, but obtained through a training process on empiric datasets. In practice the network learns the function that links input together with output by processing correct input/output couples. Actually, for each given input, within the learning process, the network gives a certain output that is not exactly the desired output, so the training algorithm modifies some parameters of the network in the desired direction. Hence, every time an example is input, the algorithm adjusts its network parameters to the optimal values for the given solution: in this way the algorithm tries to reach the best solution for all the examples. These parameters we are speaking about are essentially the weights or linking factors between each neuron that forms our network.

Neural Networks' application fields are typically those where classic algorithms fail because of their inflexibility (they need precise input datasets). Usually problems with imprecise input datasets are those whose number of possible input datasets is so big that they cannot be classified. A field where classic algorithms are in troubles is the analysis of those phenomena whose mathematical rules are unknown. There are indeed rather complex algorithms which can analyses these phenomena but, from comparisons on the results, it comes out that neural networks result far more efficient: these algorithms use Fourier's transform to decompose phenomena in frequential components and for this reason they result highly complex and they can only extract a limited number of harmonics generating a big number of approximations. A neural network trained with complex phenomena's data is able to estimate also frequential components, this means that it realizes in its inside a Fourier's transform even if it was not trained for that.

One of the most important neural networks' applications is undoubtfully the estimation of complex phenomena such as meteorological, financial, socio-economical or urban events. Thanks to a neural network it is possible to predict, analyzing historical series of datasets just as with these systems but there is no need to restrict the problem or use Fourier's transform. A defect common to all those methods it is to restrict the problem setting certain hypothesis that can turn out to be wrong. We just have to train the neural network with historical series of data given by the phenomenon we are studying [Anderson, James A. and Edward Rosenfield., 1988.].

Calibrating a neural network means to determinate the parameters of the connections (synapses) through the training process. Once calibrated there is needed to test the network efficiency with known datasets, which has not been used in the learning process. There is a great number of Neural Networks [Anderson, James A. 1995] which are substantially distinguished by: type of use, learning model (supervised/non-supervised),earning algorithm, architecture, etc. Multilayer perceptrons (MLPs) are layered feed forward networks typically trained with static backpropagation. These networks have found their way into countless applications requiring static pattern classification. Their main advantage is that they are easy to use, and that they can approximate any input-output map. In principle, backpropagation provides a way to train networks with any number of hidden units arranged in any number of layers. In fact, the network does not have to be organized in layers any pattern of connectivity that permits a partial ordering of the nodes from input to output is allowed. In other words, there must be a way to order the units such that all connections go from earlier (closer to the input) to later ones (closer to the output). This is equivalent to stating that their connection pattern must not contain any cycles. Networks that respect this constraint are called feed forward networks; their connection pattern forms a directed acyclic graph or dag.

## Materials And Methods

### Plant material

This experiment was conducted during three consecutive years in a vineyard planted with Richter 110 rootstock and grafted to Vitis vinifera L., cv. Tempranillo. The vineyard was located at the experimental fields of the Polytechnic University of Madrid. Vine spacing was 2m between rows and 1.35m within the row (3700 vines per hectare). Row orientation was North-South. All vines were head trained and cane-pruned (Guyot), and shoots were positioned with a vertical shoot positioning trellis system.

### Irrigation treatments

Two irrigation regimes were established: irrigated (I) and non-irrigated (NI) vines. The object was to replace weekly vineyard evapotranspiration (ETc) in the soil from the earliest stages of plant growth, which has depended of when the precipitation took place, as it has been described previously [Esteban MA, Villanueva MJ and Lissarrague JR, 1999]. The potential evapotranspiration (ET0) was calculated from a class A pan evaporation [Doorenbos J and Pruitt WO, 1977]. Daily trickle irrigation was applied at 0.6 x ET0 in the irrigated treatment (I), and no water was applied in the non-irrigated treatment (NI) over the entire growing season. Precipitation amounts less than 5 mm were ignored, and the irrigation application efficiency was considered to be 90%. The soil at this site had a water availability of 131 mm/m. Data on seasonal and annual rainfall, effective rainfall, total water applied, irrigation period, and accumulated growing degree days (10ºC basis) from budbreak to harvest have been described in an earlier paper [Esteban MA, Villanueva MJ and Lissarrague JR, 1999].

Four replications of each of the two treatments were randomly distributed in the vineyard, each replication consisting of three rows with nine vine plots.  Measurements were made on the central seven vines of the middle row.

### Analytical determinations

***General variables***: Total soluble solids (ºBrix) was measured using an Abbé type refractometer (Zeiss, mod.B) equipped with a temperature control system (20ºC). Must pH was measured with a pH meter (Crison mod. MicropH 2001), using a glass electrode. Finally, titratable acidity was measured by titration with a base to an end point of pH=8.2 (20ºC), and the results were expressed in g/L tartaric acid.

***Glucose and fructose***: Analysis of these two sugars was performed by HPLC according to the procedure described by Esteban et al. [Esteban MA, Villanueva MJ and Lissarrague JR, 1999].

The chromatograph employed was equipped with a refractive index detector (Waters 410 differential refractometer), and the sample and reference cells were held at 40 ºC. An Aminex HPX-87P column (300 mm x 7.8 mm i.d., 9-μm particle size) with a guard column cartridge (Bio-Rad Laboratories, Richmond, CA, U.S.A.) was used. Data were processed using the Waters Millennium 2.0 chromatographic data system.

***Tartaric acid and malic acid***: Individual acids were determined by HPLC as previously described by Esteban et al. [Esteban MA, Villanueva MJ and Lissarrague JR, 1999] The chromatograph employed was a Waters liquid chromatograph equipped with a Waters model 996 PDA detector. An Aminex HPX-87C cation exchange column (300 mm x 7.8 mm i.d., 9-μm particle size) was used, with a guard column cartridge (Bio-Rad Laboratories, Richmond, CA, U.S.A.). Data were processed using the Waters Millennium 2.0 chromatographic data system.

## Study with neural networks of Relationships between different berry components.

Multilayer feedforward networks are often used for modeling complex relationships between the data sets. Deleting unimportant data components in the training sets could lead to smaller networks and reduced-size data vectors. The process of finding relevant data components is based on the concept of sensitivity analysis applied to a trained neural network. ANN models predict changes for certain combinations of input variables, detecting the most important influence in the output variables.

We have studied different analysis for detecting relathionships between berry weight or ºBrix and other grape components in the two irrigation treatments ( T1=Irrigated and T2=non irrigated) during the ripening period.

In order to study the relationships between different variables it has been used neural networks models with a single hidden layer with 6 axons and a Tanhaxon transfer function and based on the momentum learning rule.

### Study of the relationships between different variables and ºBrix in I y NI treatments

Analysis of the results

| I | *Berry weight* | *Ph* | *Total acidity* | *ºBrix* | NI | *Berry weight* | *Ph* | *Total acidity* | *ºBrix* |
|---|---|---|---|---|---|---|---|---|---|
| | 11.342 | 17.709 | 70.949 | 100.000 | | 33.765 | 25.123 | 41.111 | 100.000 |
| | | 22.712 | 77.288 | 100.000 | | | 41.958 | 58.042 | 100.000 |
| | 17.930 | 82.070 | | 100.000 | | 55.994 | 44.006 | | 100.000 |
| | 14.877 | | 85.123 | 100.000 | | 37.859 | | 62.141 | 100.000 |

Active performance of the analysis

| I | *MSE* | *NMSE* | *r* | *%Error* | NI | *MSE* | *NMSE* | *r* | *%Error* |
|---|---|---|---|---|---|---|---|---|---|
| | 0.03 | 0.01 | 0.99 | 5.76 | | 0.004 | 0.01 | 0.99 | 6.66 |
| | 0.004 | 0.01 | 0.99 | 6.43 | | 0.008 | 0.02 | 0.98 | 9.61 |
| | 0.008 | 0.03 | 0.98 | 8.67 | | 0.005 | 0.01 | 0.99 | 6.72 |
| | 0.003 | 0.01 | 0.99 | 5.86 | | 0.004 | 0.01 | 0.99 | 6.64 |

During the ripening period in berries of the cv. Tempranillo grape variety along a period of three years, it has been studied that the values of ºBrix in the two irrigated treatments are different. It has been analysed the importance that has the impact of some components (total acidity, pH and berry weight) on the ºBrix. Thus we observed that total acidity is the variable that influences most in the irrigated treatment with 70.9%, followed of pH with 17,7% and finally the berry weight with 11.3%. In the non-irrigated treatment occurs the same, reaching total acidity a value of 41,1%, however, the berry weight (33.7%) influences more than pH (25.1%). Analyzing variables two to two we verified that those models are the same in both treatments. Thus, the impact of the berry weight in the ºBrix in the irrigated treatment is less than in the non- irrigated treatment, this could be because of the concentration effect that takes place since the absolute values in both treatments are the same.

We have also analyzed Tartaric and Malic

| I | Berry weight | Tartaric | Malic | ºBrix | NI | Berry weight | Tartaric | Malic | ºBrix |
|---|---|---|---|---|---|---|---|---|---|
| | 8.556 | 31.008 | 60.436 | 100.000 | | 40.110 | 14.112 | 45.778 | 100.000 |

Active performance of the analysis

| I | MSE | NMSE | r | %Error | NI | MSE | NMSE | r | %Error |
|---|---|---|---|---|---|---|---|---|---|
| | 0.003 | 0.001 | 0.99 | 5.22 | | 0.003 | 0.01 | 0.99 | 6.04 |

We have analyzed the two most important acids in the grape because they determine the value of the total acidity. As it happens with the total acidity, both acids influence in the ºBrix value more than the berry weight in the irrigated treatment, whereas in the non- irrigated treatment this only happens with the malic acid and nor with the tartaric acid.

**Study of the relationships between different variables and berry weight in I y NI treatments**

Analysis of the results

| I | ºBrix | Ph | Total acidity | Berry weight | NI | ºBrix | Ph | Total acidity | Berry weight |
|---|---|---|---|---|---|---|---|---|---|
| | 29.960 | 15.411 | 54.629 | 100.000 | | 60.653 | 28.628 | 10.719 | 100.000 |
| | 29.211 | 70.789 | | 100.000 | | 67.386 | 32.614 | | 100.000 |
| | 37.080 | | 62.920 | 100.000 | | 76.970 | | 23.030 | 100.000 |
| | | 38.506 | 61.494 | 100.000 | | | 47.754 | 52.246 | 100.000 |

Active performance of the analysis

| I | MSE | NMSE | r | %Error | NI | MSE | NMSE | r | %Error |
|---|---|---|---|---|---|---|---|---|---|
| | 0.01 | 0.07 | 0.96 | 8.99 | | 0.005 | 0.02 | 0.98 | 5.77 |
| | 0.02 | 0.08 | 0.94 | 10.4 | | 0.005 | 0.02 | 0.98 | 5.88 |
| | 0.02 | 0.09 | 0.95 | 10.3 | | 0.005 | 0.02 | 0.98 | 5.83 |
| | 0.03 | 0.12 | 0.93 | 10.9 | | 0.02 | 0.11 | 0.93 | 11.08 |

It has been studied the importance of the impact of some variables (total acidity, pH and ºBrix) on the berry weight. Thus we observed that total acidity is the variable that influences most in the irrigated treatment with 54.6%, followed by ºBrix with 29.9% and finally pH with 15,4%. In the non-irrigated treatment ºBrix is the variable that influence the most in the berry weight with a value of 60.6%, then the pH (28.6%) and the total acidity (10.7%).

We have also analyzed Glucose, Fructose, Tartaric and Malic.

| I | Ph | Glucose | Fructose | Tartaric | Malic | Berry weight |
|---|---|---|---|---|---|---|
| | 22.645 | 18.624 | 33.515 | 10.830 | 14.387 | 100.000 |
| | | 20.461 | 34.891 | 11.707 | 32.942 | 100.000 |

| NI | Ph | Glucosa | Fructosa | Tartárico | Málico | Berry weight |
|---|---|---|---|---|---|---|
| | 18.590 | 36.107 | 25.640 | 8.388 | 11.275 | 100.000 |
| | | 41.808 | 24.284 | 14.115 | 19.793 | 100.000 |

Active performance of the analysis

| I | MSE | NMSE | r | %Error |
|---|---|---|---|---|
| | 0.01 | 0.05 | 0.97 | 7.93 |

| NI | MSE | NMSE | r | %Error |
|---|---|---|---|---|
| | 0.004 | 0.02 | 0.98 | 5.3 |

Glucose and fructose are the most important sugars in the grapes and they are the ones that determine mainly the ºBrix value. In the irrigated treatment pH is the variable that influences most in berry weight followed by fructose and glucose, although the amounts of the two sugars influence more than any other variable. However, in the non- irrigated treatment the impact of these two sugars is the highest.

## Conclusion

These results provides that in both treatments, irrigated and non-irrigated vines, and during the different stages of the berry growth it is possible to stablish significative relationships between the parameters studied. The results with neural networks show that total acidity is the variable that influence most in ºBrix value in both treatments, when the analysis has been total or with tartaric and malic acids except in the case of tartaric acid in the non irrigated treatment. It is also shown that ºBrix value is the variable that influences most in the berry weight non irrigated treatment while total acidity is in the irrigated one. However, in both treatments glucose and fructose influence more in the berry weight than tartaric and malic acids.

## Bibliography

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

[Anderson, James A. and Edward Rosenfield., 1988] Anderson, James A. and Edward Rosenfield. Neurocomputing: Foundations of Research

Cambridge, MA: MIT Press (1988).

[Boulton RB, 1980] Boulton RB, The relationship between total acidity, titratable acidity and pH in grape tissue. Vitis 19:113-120 (1980).

[Doorenbos J and Pruitt WO, 1977] Doorenbos J and Pruitt WO, Guidelines for predicting crop water requirements. FAO Irrig. Drain. Paper 24 (1977).

[Esteban MA, Villanueva MJ and Lissarrague JR, 1999] Esteban MA, Villanueva MJ and Lissarrague JR, Effect of irrigation on changes in berry composition of Tempranillo during maturation. Sugars, organic acids and mineral elements. Amer J Enol Viticult 50:418-434 (1999).

[Esteban MA, Villanueva MJ and Lissarrague JR, 2001] Esteban MA, Villanueva MJ and Lissarrague JR, Effect of irrigation on changes in the anthocyanin composition of the skin of cv Tempranillo (Vitis vinifera L.) grape berries during ripening. J Sci Food Agric 81:409-420 (2001).

[Hale CR, 1977] Hale CR, Relation between potassium and the malate and tartrate contents of grape berries. Vitis 16:9-19 (1977).

[Hardy P.J.,1968] Hardy P.J. Metabolism of sugars and organic acids in immature grape berries. Plant Physiol. 43:224-228 (1968).

[Hanson, Stephen J. and David J. Burr. 1990] Hanson, Stephen J. and David J. Burr. "What connectionist models learn: Learning and representation in connectionist networks." Behavioral and Brain Sciences, vol. 13, no. 3, pp. 471-518 (Sept. 1990).

[Hrazdina G, Parsons GF and Mattick LR, 1984] Hrazdina G, Parsons GF and Mattick LR, Physiological and biochemical events during development and maturation of grape berries. Am J Enol Vitic 35:220-227 (1984).

[Kliewer WM, 1968] Kliewer WM, Changes in concentration of free amino acids in grape berries during maturation. Amer J Enol Viticult 19:166-174 (1968).

[McCarthy MG, Cirami RM and McCloud P, 1983 ] McCarthy MG, Cirami RM and McCloud P, Vine and fruit responses to supplementary irrigation and canopy management. S Afr J Enol Vitic 4:67-76 (1983).

[Ough CS, 1968] Ough CS, Proline content of grapes and wine. Vitis 7:321-31 (1968).

[Ough CS and Amerine MA, 1988] Ough CS and Amerine MA, Methods for analysis of musts and wines. John Wiley and Sons, New York (1988).

[Peynaud E and Ribéreau-Gayon J, 1971] Peynaud E and Ribéreau-Gayon J, The grape. In: The biochemistry of fruits and their products. Vol. II. Ed by Hulme AC, Academic Press, London and New York, pp. 171-205 (1971).

[Ribéreau-Gayon J, Peynaud E, Ribéreau-Gayon P and Sudraud P, 1975] Ribéreau-Gayon J, Peynaud E, Ribéreau-Gayon P and Sudraud P, Traité d´oenologie Sciences et techniques du vin. Vol. 2. Ed. Dunod, Paris (1975).

[Romero EG, Muñoz GS and Ibañez MDC, 1993] Romero EG, Muñoz GS and Ibañez MDC, Determination of organic acids in grape musts, wines and vinegars by high-performance liquid chromatography. J Chromatogr 655:111-117 (1993).

[Ribéreau-Gayon J, Peynaud E, Ribéreau-Gayon P and Sudraud P, 1975] Ribéreau-Gayon J, Peynaud E, Ribéreau-Gayon P and Sudraud P, Traité d´oenologie Sciences et techniques du vin. Vol. 3. Ed. Dunod, Paris (1976).

[Rühl EH and Alleweldt G, 1985] Rühl EH and Alleweldt G, Investigations into the influence of time of irrigation on yield and quality of grapevines. Acta Horticulturae 171:457-462 (1985).

[Van Zyl JL, 1984] Van Zyl JL, Response of Colombar grapevines to irrigation as regards quality aspects and growth. S Afr J Enol Vitic 5:19-28 (1984).

[Williams LE and Matthews MA, Grapevine, 1990] Williams LE and Matthews MA, Grapevine. In: Irrigation of Agricultural Crops, (Agronomy Monograph No. 30), Ed by Stewart, BA and Nielsen DR, ASA-CSSA-SSSA, Madison, WI, pp 1019-1055 (1990).

## Authors' Information

***Castellanos A.*** *– 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; e-mail: angel.castellanos@upm.es*

***Esteban M.A.*** *– Ministerio de Agricultura. P.º Infanta Isabel 1, 28014 Madrid, Spain; e-mail: marita.esteban@ya.com*

***Martinez A.*** *– Natural computing group. Universidad Politécnica de Madrid, Spain; e-mail: a.martinez@upm.es*

***Palencia V.*** *– Natural computing group. Universidad Politécnica de Madrid, Spain; e-mail: vpalencia@fi.upm.es*

# CONSTRUCTION OF THE MODEL OF INDIVIDUAL MULTIFACTOR ASSESSMENT BY MEANS OF GMDH-NEURAL NETWORK

**Eduard Petrov, Konstantin Petrov, Tatyana Chaynikova**

***Abstract:*** *This paper deals with one of the approaches to the solution of the problem of structurally-parametric identification of the model of individual multifactor assessment which is based on the use of artificial GMDH-neural network. The use of artificial neural network (ANN) methodology opens the prospects for paralleling of the model synthesis process and realizing computational procedure of linear complexity.*

***Keywords:*** *comparative identification, GMDH – neural network, model of multifactor assessment, Kolmogorov-Gabor polynomial.*

## Introduction

Decision-making process is defined as a particular kind of mental work, composed of one or several alternate solutions from some admissible set.

General problem of decision making can be divided into the following steps: goal formation and analysis; definition of probable ways set for the goal achievement (the problem of admissible solutions set formation); formation of assessment allowing to compare admissible solutions with each other in quality (assessment problem); admissible set ranking and extremal solution selection (optimization problem).

Without diminishing the importance of the above mentioned steps it should be noted that assessment is one of the most important problems.

The assessments of the alternative solutions formed by experts are subjective and cannot be directly measured in most cases. Therefore classical methods of direct identification of the estimation model are non-applicable. The alternative ones are methods of indirect identification, one of them being the method of comparative identification [1].

General peculiarity of the methods for identification of multifactor assessment model is partitioning of the problem into two parts: the problem of structural identification (generation of the model structure) and parametric identification problem. These problems are solved in series for each version of the model. This makes the procedure of structurally-parametric identification of the model (under high enough dimensionality of tuple of factors characterizing the alternative) very hung us in terms of computing. The prospect to increase the computational efficiency of structurally-parametric identification process presumes:

- algorithmic and computational integration of structural and parametric identification steps into a single process;
- the computational procedure done in such a way that it causes identification of the local models being held in parallel but not in series.

Different approaches to the problem solution are possible, but if we take into account the specifics of the examined then the use of the artificial neural networks appliance (ANN) as a tool turns out to be most prospective.

## Problem statement

Let us have a bounded set of alternate solutions $X = \{ x_1, x_2, ..., x_n \}$. Each alternative $x_i \in X$ , $i = \overline{1,n}$ is described by the tuple of partial characteristics (factors) $K(x_i) = \langle k_1(x_i), k_2(x_i), ..., k_m(x_i) \rangle$, that allow their objective quantitative measuring.

On the basis of the given information analysis an individual is able:

- to choose from X the most preferable decision, for example $x_t$;
- to rank all decisions in descending (ascending) order of their preference, i.e. to establish a strict or no strict order, e.g. $x_1 \succ (\geq) \tilde{o}_2 \succ (\geq) \ldots \succ (\geq) x_n$ or a partial linear order relation, e.g. $\mathrm{x}_1 \succ (\geq) \tilde{\mathrm{o}}_2 \sim \tilde{o}_3 \succ (\geq) \ldots \succ (\geq) x_n$.

Thereto an individual forms subjective assessment for each specific alternative $x_i \in X$ , $i = \overline{1,n}$, that reflects its utility (value). Then according to the utility theory [2], that postulates existence of some scalar quantitative assessment for preference of any alternative $x_i \in X$ , $i = \overline{1,n}$.

The first case can be formalized as the system of the following inequalities:

$$P(x_t) > (\geq) P(x_i),\ \forall i \neq t,\ i = \overline{1,n}, \tag{1}$$

the second accordingly:

$$P(x_1) > (\geq) P(x_2) > (\geq) \ldots > (\geq) P(x_n)$$

or

$$P(x_1) > (\geq) P(x_2) = P(x_3) > (\geq) \ldots > (\geq) P(x_n), \tag{2}$$

where $P(x_j)$ is an unknown scalar assessment to evaluate utility of the alternative. In this connection we find it necessary to construct a mathematical model, model of generalized utility $P(x_j)$ of a decision maker's individual choice.

## The selection of ANN architecture to solve the problem of multifactor assessment model synthesis

Nowadays ANN is viewed as a high-performance special-purpose tool for intellectual problems solving. At that for each class of problems it is necessary to synthesize problem-oriented ANN and to work out an algorithm for specific problem solution.

In most cases for ANN architecture synthesis it is presumed that the model structure is uniquely defined. It means that architecture of ANN that implements the model structure is also fixed and during the process of training only its parametric adaptation takes place. This process is similar to parametric identification of a mathematical model.

Besides there is a big class of indirect-analogy models that approximate a functional model with some polynomials or rows containing considerable nonlinearities. When these models are realized with the help of ANN in general case two problems – realization of nonlinearities and structural adaptation of ANN in the process training.

Solving nonlinear approximation problems with the help of neural nets constructed on classic artificial neurons (AN) of a perceptron type, ADALINE etc. [3] required nonlinearities can be realized with the help of activation functions. It leads to a necessity to use nonlinear methods use training (i.e. to define not only synaptic weights $w_i$, $i=\overline{1,n}$, but also parameters of nonlinear activation functions), that have low rate of convergence and are ponderous in terms of computing. The alternative for classic ANN are special-purpose networks called functionally related [3].

The general theoretical basis for construction of such a type of networks is the hypothesis [4] that dimension increase of variables input space permits to transform linearly inseparable in space $R^n$ set to linearly separable in space $R^m$, where $m >> n$. Functionally it means that input variables $X = \langle x_1, x_2, \ldots, x_n \rangle$ are subjected to necessary linear conversion $\varphi_i(X)$ and the results are examined as new complementary variables. At this expense expansion of variables input space is done and a nonlinear input function is transformed into linear by synaptic weights (parameters). Thus a transformation function of the following type is realized:

$$Y = \sum_{i=1}^{m} w_i \varphi_i(X) \tag{3}$$

This approach allows handling operations of training (parameters identification) nonlinear operators and synaptic weights.

Different types of function $\varphi_i$ can be used, though it was the polynomial extension realizing Kolmogorov-Gabor multinomial [5], i.e. transformation that was most widely spread:

$$y_j = w_{j0} + \sum_{i=1}^{n} w_{ji} x_i + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} w_{ji_1i_2} x_{i_1} x_{i_2} + \ldots$$
$$+ \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \ldots \sum_{i_l=i_{l-1}}^{n} w_{ji_1i_2\ldots i_i} x_{i_1} x_{i_2} \ldots x_{i_l} = w_j^T \varphi(X). \tag{4}$$

This form of ANN has the name of a polynomial.

As it is shown in [6] a model of multifactor assessment is well approximated with Kolmogorov-Gabor polynomial. That is why polynomial ANN is most suitable for solving of the parametric identification problem for a model. But it is also necessary to solve the problem of structural identification of the model. As a result, we have to define the structure of the model having optimal complexity [5] by the minimum error criterion. The specifics of the problem lies in the fact that information about values of alternatives preferences is given not in a quantitative form, but as a system of comparative inequalities of type (1) or (2). Later on, without a loss of generality, let us consider that an individual indicated just the most preferable alternative, i.e. the situation is examined that can be formalized as a system of inequalities of type (1). In this case the model quality can be assessed just by quantity of allowed inequalities (1). Any model that meets all the inequalities is the solution to the problem. There can be quite a great number of such models varying in complexity and parameters values of polynomial approximant (4). Let us call all models of multifactor assessment that meet the whole set of inequalities (1) equivalent models. Besides, we should choose a unique model from the set of equivalent ones. The selection criterion is a minimum of model complexity, i.e. the quantity of tenants and the rank of polynomial approximant (4). To solve this problem different methods can be used. For example, a method of step-by-step complication that can be implemented with the help of functionally related ANN [3].

The main disadvantage of this method is the necessity to use ANN with a large amount of adjustable parameters that causes serious problems connected to the pace of ANN training even while using algorithms with the optimal operation speed.

The alternative method to the method of step-by-step complication of the structure for solving problems demanding not only parametric but also structural ANN training is a group method of data handling (GMDH) [7]. This method allows decreasing the dimensionality of a parametric training problem fundamentally.

GMDH implementation with the help of ANN is based on the use of N-ADALINE (N-A) formal neuron [8]. A scheme of ANN realizing GMDH is given in figure 1.

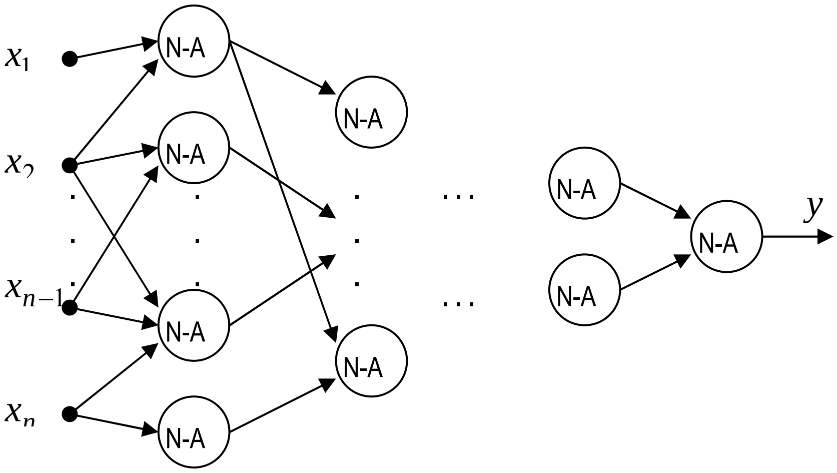


Figure 1. GMDH-neural network

When solving the problem of structurally-parametric synthesis of the multifactor assessment problem with the help of GMDH it can be found that several equivalent models are obtained as output. A unique model is chosen according to the minimal complexity criterion.

## ANN training for synthesis of the multifactor assessment model

On solving the structurally-parametric identification problem for the multifactor assessment model, the initial information is given as a system of comparative inequalities (1) that allows to realize the training procedure with a teacher.

The use of the methodology of variables set increase allows us to obtain the ANN structure that is linear by parameters and thereby excludes the necessity of the nonlinear training.

The criterion of the model quality representing the quantity of allowed comparative inequalities (1) is discrete and non-differentiable, i.e. for training only it is possible to use methods of the zero order.

The problem of ANN linear training (defining the value of the synaptic weights) is formally reduced to a problem of parametric identification at a stipulated model structure. This problem is analyzed in detail in [9,10] where it is shown that taking into account the peculiarities of the problem of multifactor assessment model synthesis, it is appropriate to use a method of Chebyshev point estimation, midpoint technique and genetic algorithms for parametric identification and correspondingly for ANN training. The advantage of the first two methods consists in the fact that their realization is based on the standard algorithms of linear programming. But they define only one solution (point solution) from some accessible region. In contrast to this a genetic algorithm allows to determine the set of equivalent by criterion of inequalities satisfaction models (1).

As an alternative to genetic algorithms for equivalent models determination we suggest to use random search technique [11], a coordinate wise descent method or Gauss-Seidel method [12]. At that to increase computing efficiency Chebyshev point or midpoint can be used as a starting one.

## An algorithm for solving the problem of structurally parametric identification of the multifactor assessment model with the help of ANN

With a view to describe the algorithm obviously and simply we will examine the solution of the problem of multifactor assessment synthesis of the alternatives $x_i$, $i=\overline{1,n}$ in the situation when each of them is characterized by four normalized values of the partial criteria $K(x_i)=\{k_1(x_i),k_2(x_i),k_3(x_i),k_4(x_i)\}$. In case of any other quantity of characteristics procedure of solution will be similar.

1. Selection of artificial neuron model generating ANN.

In contrast to traditional neural networks with fixed architecture a GMDH-network has a structure that can vary ("grow") during training.

As a formal neuron in this network N-ADALINE (N-A) is used. It represents an adaptive linear associator with two inputs and a nonlinear preprocessor. By means of a preprocessor nonlinear dependences of any complexity can be realized.

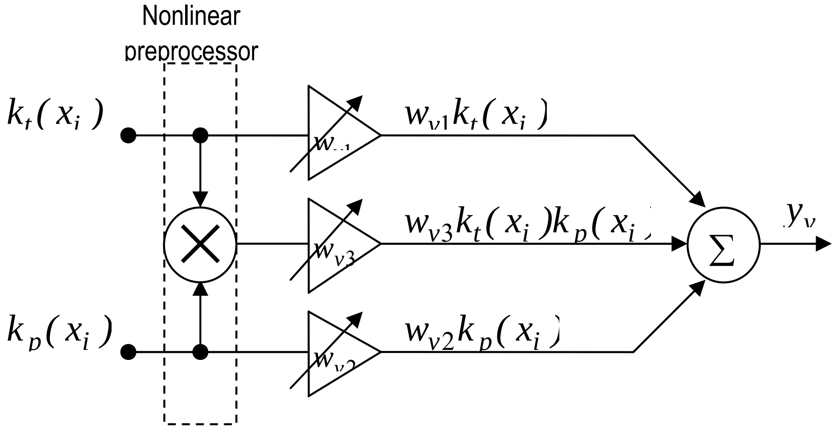


Figure 2. Scheme of N-ADALINE (N-A).

As analysis made in [10] showed for the solution of the structurally-parametric identification of the multifactor assessment problem it is advisable to apply a formal neuron N-A as a basis of ANN , realizing a local polynomial of such a type: $y_v = w_{v1}k_t(x_i)+w_{v2}k_p(x_i)+w_{v3}k_t(x_i)k_p(x_i)$.

Schematic sketch of the neuron model is given in figure 2.

But this does not exclude the possibility to use formal neurons that realize more complicated structures of polynomials.

2. Defining the first layer structure of ANN.

The process of GMDH-network training lies in its configuring starting from the input layer, synaptic weights adjustment for each neuron and increase of layers quantity to achieve a desired value of the model quality criterion.

Let us represent all the set of input signals $K(x_i)=\{k_1(x_i),k_2(x_i),k_3(x_i),k_4(x_i)\}$ as various paired combinations. Their quantity will be equal to a quantity of the first layer neurons that can be defined as $C_m^2$.

For our case ($m=4$) there will be six combinations of such a type $(k_1(x_i),k_2(x_i))$; $(k_1(x_i),k_3(x_i))$, $(k_1(x_i),k_4(x_i))$, $(k_2(x_i),k_3(x_i))$, $(k_2(x_i),k_4(x_i))$,

$(k_3(x_i), k_4(x_i))$. It follows from this that the first layer of ANN will consist of six formal N-A neurons each of them realizing local polynomials of the form

$$\begin{array}{l} y_1(x_i) = w_{11}k_1(x_i) + w_{12}k_2(x_i) + w_{13}k_1(x_i)k_2(x_i), \\ y_2(x_i) = w_{21}k_1(x_i) + w_{22}k_3(x_i) + w_{23}k_1(x_i)k_3(x_i), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_6(x_i) = w_{61}k_3(x_i) + w_{62}k_4(x_i) + w_{63}k_3(x_i)k_4(x_i),. \end{array} \quad i = \overline{1,n} \qquad (5)$$

3. Adjustment of synaptic weights of the first layer neurons.

As it was mentioned above, information which is received as a result of the comparative experiment can be formalized as a set of constraints (1). Each neuron is adjusted with the help of training with a teacher.

The whole the set of constraints (1) is preliminary divided into two subsets: a learning and checking set. This nontrivial procedure is based on the analysis of the quality and volume of obtained experimental data.

The definition of synaptic weights values of the ANN first layer neurons (coefficients) $w_{vu}$, $v = \overline{1,6}$, $u = \overline{1,3}$ is practically a problem of the polynomials (models) parametric identification (5) on the learning subset of constrains from (1).

All possible approaches to this problem solution, namely a method of Chebyshev point estimation, midpoint technique, and genetic algorithms are examined in [9, 10].

4. Quality evaluation of obtained models.

After presenting to a network all the learning population of constrains from (1) and synaptic weights adjustment, ”accuracy” of each neuron is evaluated and a group of neurons is formed, that gives an error below some a priori given threshold.

Because of the specific peculiarities of the simulation object we can use a magnitude equal to a quantity of satisfied inequalities of a checking subset from (1) as a model estimate (outputs of formal first layer neurons). Formally an evaluation stage can be realized in the following way.

Let us assume that a checking set contains $G$ constrains (1). Then for each local polynomial (model) $y_v$, $v = \overline{1,6}$ we can define its quality assessment $O_v \in [0;1]$ by expression:

$$O_v = \frac{G_v}{G}, \qquad (6)$$

where $G_v$ is a quantity of the satisfied constrains of the checking set from (1) for $v$-th polynomial.

On the base of the obtained estimates $O_v$, $v = \overline{1,6}$ ”the best” polynomials (satisfying the largest quantity of checking set constrains) are chosen from (5).

If one or several outputs from the first layer neurons have estimate (6) equal to 1, than the process of ANN construction is stopped and as an output of network a polynomial of minimal complexity is taken.

If there are no such estimates then outputs with maximal estimate values (6) are chosen. At that [13] their quantity should not outnumber input signals of the network. In a given example there will be four of them (from six outputs). Let us denote the corresponding outputs of these neurons (polynomials), for example with $y_1(x_i)$, $y_2(x_i)$, $y_3(x_i)$, $y_4(x_i)$. There are outputs of this group of neurons which are inputs of the second hidden layer.

5. Generating the second and consequent hidden layers of ANN.

Polynomials $y_1(x_i)$, $y_2(x_i)$, $y_3(x_i)$, $y_4(x_i)$ are examined as input signals of the second layer neurons. It means that paired combinations are again made of them, i.e. for each combination polynomials are synthesized:

$$\begin{aligned}
z_1(x_i) &= q_{11}y_1(x_i) + q_{12}y_2(x_i) + q_{13}y_1(x_i)y_2(x_i),\\
z_2(x_i) &= q_{21}y_1(x_i) + q_{22}y_3(x_i) + q_{23}y_1(x_i)y_3(x_i), \qquad i=\overline{1,n} \qquad (7)\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\\
z_6(x_i) &= q_{61}y_3(x_i) + q_{62}y_4(x_i) + q_{63}y_3(x_i)y_4(x_i),
\end{aligned}$$

Then with the help of the same learning subset the second layer neurons are adjusted (realizing (7)) under ”frozen” synaptic weights of the first layer and again a group is formed having maximal values of the estimates calculated by (6).

6. Rule of stopping and reconstruction of the required model structure.

The process of network layers building-up and synaptic weights adjustment lasts until at least one polynomial with the estimate equal to 1 will be obtained (on the basis of the checking set) or values of estimates will not stop to increase.

The best neuron of the last layer is stated as an output neuron of the network as a whole. Then on the base of the polynomial obtained in the output of the network by means of successive substitution of intermediate local descriptions the structure of the required model of alternatives multifactor assessment is rebuilt in terms of initial variables (input network signals).

Schematically the process of GMDH-network building for the examined example can be represented as is shown on figure 3.

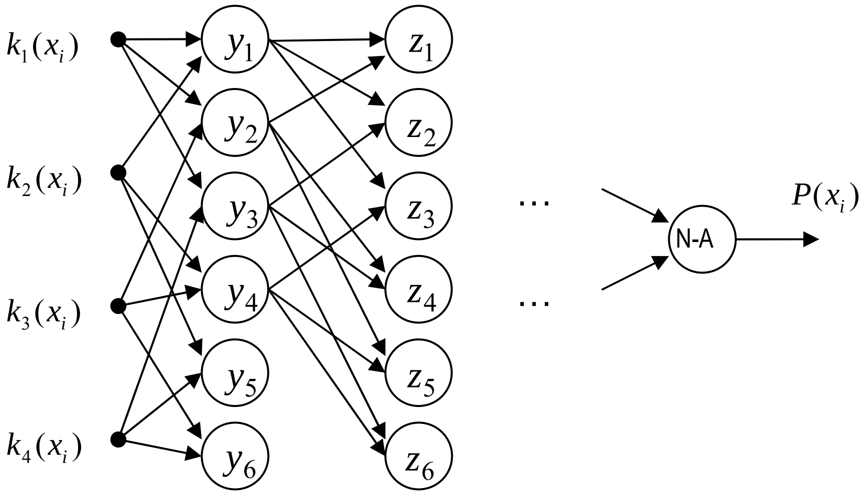


Figure 3. Scheme of the process of GMDH-network building.

## Conclusions

The development of methods for mental work synthesis in general and multifactor assessment models in particular that are not based on subjective expert assessment is of great theoretical and practical value for the theory of multicriteria optimization, multidimensional classification, quality assessment, etc. From this point of view the development of the ”objective” identification methods is perspective. One of these methods presented in this paper is the method of comparative structurally-parametric identification of the multifactor assessment model.

The use of GMDH-neural network as a tool offers challenges for parallel solving of generation problems (simulated emission) of model structures variants and their parameters identification and at the same time allows using procedures of linear complexity for network training. It allows practically eliminating restrictions from structural complexity and criteria dimensionality of the multifactor assessment model.

## Bibliography

Ovezgeldyev A., Petrov K. Comparative identification of the mental work models. Journal Cybernetics and system analysis, №5, 1996, pp. 48-58.

Neiman J, Morgenshtern O. Game theory and economic behaviour. Nauka, Moscow, 1970. P. 124.

Bodyanskij E., Rudenko O. Artificial neural networks: architecture, training, implementation. Teletex, Kharkov, 2004. P. 370.

Cover T.M. Geometrical and statistical of systems of linear inequalities with applications in pattern recognition. In IEEE Trans. On Electronic Computers, 1965, pp. 326 – 334.

Ivakhnenko A. Self-organizing systems of recognition and automated management. Tehnika, Kiev, 1969. P. 392.

Ovezgeldyev A., Petrov K. Assessment and ranking of alternatives under the interval conditions. Journal Cybernetics and system analysis, №4, 2005, pp. 148 – 153.

Ivakhnenko A., Zajchenko Y., Dimitrov V. Decision making on the basis of self-organization.Sovetskoe Radio, Moskow, 1976. P. 280.

Pham D., Liu X. Modeling and Prediction using GMDH Networks of Adalines with Nonlinear Preprocessors. Journal System Science, 1994, pp. 1743 – 1759.

Ovezgeldyev A., Petrov E., Petrov K. Synthesis and identification of multifactor assessment and optimization models. Naukova dumka, Kiev, 2002. P. 162.

Petrov E., Bulavin D., Petrov K. The use of genetic algorithms for the solution of the structurally-identification problem of model of individual multifactor assessment. Journal Bionics problems. №60, 2004, pp. 17-27.

Rastrigin L. Random search in adaptation process. Zinatne, Riga, 1973. P. 132.

Reklejtis G., Rejvindran A., Regsdel K. Optimization in technology. Mir, Moscow, 1986. P. 360.

Ivakhnenko A., Yurachkovskij Y. Complex systems modelling according to experimental data. Radio i svyaz, Moscow, 1987. P. 116.

## Authors' information

***Eduard G. Petrov*** *– Professor. Department of Systems Science, Kharkov National University of Radio Electronics, Lenin Avenue, 14, 61166, Kharkov, Ukraine* kpetrov@kharkov.ukrtel.net

***Konstantin E. Petrov*** *- Senior lecturer. Department of Applied Mathematics, Kharkov National University of the Internal Affairs, 50-letiya SSSR Avenue, 27, 61080, Kharkov, Ukraine;* kept@mail.ru

***Tatyana S. Chaynikova*** *- PhD Student. Department of Systems Science, Kharkov National University of Radio Electronics, Lenin Avenue, 14, 61166, Kharkov, Ukraine* st@kture.kharkov.ua