Fourth International Conference

INFORMATION RESEARCH AND APPLICATIONS

20-25 June 2006, Varna

PROCEEDINGS

FOI-COMMERCE

SOFIA, 2006

Kr. Markov, Kr. Ivanova (Ed.)

Proceedings of the Fourth International Conference “Information Research and Applications” i.TECH 2006, Varna, Bulgaria

Sofia, FOI-COMMERCE – 2006

ISBN-10: 954-16-0036-0
ISBN-13: 978-954-16-0036-8
First edition

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

All rights reserved.

© 2006 Krassimir Markov, Krassimira Ivanova - Editors
© 2006 Institute of Information Theories and Applications FOI ITHEA - Preprint processing
© 2006 FOI-COMMERCE - Publisher
© 2006 For all authors in the issue

® i.TECH is a trade mark of Krassimir Markov

ISBN-10: 954-16-0036-0
ISBN-13: 978-954-16-0036-8
C\o Jusautor, Sofia, 2006

PREFACE

The Fourth International Conference "***Information Research and Applications***" (**i.TECH 2006**) is organized as a part of "ITA 2006 - Joint International Scientific Events on Information Theories and Applications".

The main organizer of the ITA 2006 as well as the i.TECH 2006 is

International Journal on Information Theories and Applications (IJ ITA).

The aim of the conference is to be one more possibility for contacts for IJ ITA authors. The usual practice of IJ ITA is to support several conferences at which the IJ ITA papers may be discussed before submitting them for referring and publishing in the journal. 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.

Accent of i.TECH 2006 are the new trends in imaging for security and medical applications (section "Information Technologies in Biomedicine").

We would like to express our thanks to all who support the i.TECH 2006 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. The group is one of the foundational groups of the European Molecular Computing Consortium (http://openit.disco.unimib.it/emcc/welcome.html). NCG is involved in natural computing activities researches including: Molecular Computing (DNA Computing and Membrane Computing), Artificial Neural Networks (new architectures and learning strategies, Chaos controlling by Artificial Neural Networks, etc), Artificial Intelligence, Evolutionary algorithms, etc. NCG is constituted by members of some departments of the Technical University of Madrid, having its site at the Department of Artificial Intelligence of the Faculty of Computer Science. NCG has participated in different national and international projects including INTAS and Framework Programs of European Union (MolCoNet).

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 2006 Proceedings has been edited in the *Institute of Information Theories and Applications FOI ITHEA* in collaboration with the leading researchers from *Institute of Cybernetics "V.M.Glushkov", NASU (Ukraine), Kiev University "T.Shevchenko" (Ukraine), Institute of Mathematics and Informatics, BAS (Bulgaria), University of Calgary (Canada); VLSI Systems Centre, Ben-Gurion University (Israel).*

The i.TECH 2006 Conference found the best support in the work of organizing secretary, IJ ITA editor, Ilia Mitov.

To all participants of i.TECH 2006 we wish fruitful contacts during the conference days and efficient work for preparing the high quality papers to be published in the International Journal on Information Theories and Applications.

Varna, June 2006 Kr Markov, Kr. Ivanova

i.TECH 2006 has been organized by:

International Journal "Information Theories and Applications"
Association of Developers and Users of Intelligent Systems(Ukraine)
Institute of Information Theories and Applications FOI ITHEA (Bulgaria)
V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine
Natural Computing Group of the Technical University of Madrid (Spain)
Institute of Mathematics and Informatics, BAS (Bulgaria)
National Academy of Sciences of Ukraine
Varna Free University "Chernorizets Hrabar" (Bulgaria)
New Technik Publishing Ltd. (Bulgaria)

Program Committee:

Victor Gladun (Ukraine) Krassimir Markov (Bulgaria)

Alexey Voloshin (Ukraine)	Avram Eskenazi (Bulgaria)	Luis Fernando de Mingo (Spain)
Adil Timofeev (Russia)	Konstantin Gaindrik (Moldova)	Orly Yadid-Pecht (Israel)
Alexander Gerov (Bulgaria)	Krassimir Manev (Bulgaria)	Peter Stanchev (USA)
Alexander Kuzemin (Ukraine)	Krassimira Ivanova (Bulgaria)	Petia Asenova (Bulgaria)
Alexander Palagin (Ukraine)	Laura Ciocoiu (Romania)	Plamen Mateev (Bulgaria)
Anna Kantcheva (Bulgaria)	Levon Aslanyan (Armenia)	Radoslav Pavlov (Bulgaria)
Arkady Zakrevskij (Belarus)	Maria Kasheva (Bulgaria)	Rumiana Kirkova (Bulgaria)
Ilia Mitov (Bulgaria)	Martin Mintchev (Canada)	Stefan Dodunekov (Bulgaria)
Ivan Popchev (Bulgaria)	Natalia Ivanova (Rusia)	Stoyan Poryazov (Bulgaria)
Juan Castellanos (Spain)	Nelly Maneva (Bulgaria)	Tsvetanka Kovacheva (Bulgaria)
Jury Zaichenko (Ukraine)	Nikolay Lutov (Bulgaria)	Vitaliy Lozovskiy (Ukraine)
Koen Vanhoof (Belgium)	Nikolay Zagoruiko (Russia)	Vladimir Ryazanov (Russia)

Organizing secretary: Ilia Mitov

Papres of i.TECH 2006 are collated in six sections:

- **Information Technologies in Biomedicine**
- **Knowledge Engieneering**
- **Information Models**
- **Software Engineering**
- **Information Systems**
- **Networks and Telecommunications**

Official languages of the conference are English and Russian.

General sponsor of the i.TECH 2006 is FOI BULGARIA (www.foibg.com).

TABLE OF CONTENTS

Information Technologies in Biomedicine

MANOMETRY-BASED COUGH IDENTIFICATION ALGORITHM

Jennifer A. Hogan, Martin P. Mintchev

***Abstract**: Gastroesophageal reflux disease (GERD) is a common cause of chronic cough. For the diagnosis and treatment of GERD, it is desirable to quantify the temporal correlation between cough and reflux events. Cough episodes can be identified on esophageal manometric recordings as short-duration, rapid pressure rises. The present study aims at facilitating the detection of coughs by proposing an algorithm for the classification of cough events using manometric recordings. The algorithm detects cough episodes based on digital filtering, slope and amplitude analysis, and duration of the event. The algorithm has been tested on in vivo data acquired using a single-channel intra-esophageal manometric probe that comprises a miniature white-light interferometric fiber optic pressure sensor. Experimental results demonstrate the feasibility of using the proposed algorithm for identifying cough episodes based on real-time recordings using a single channel pressure catheter. The presented work can be integrated with commercial reflux pH/impedance probes to facilitate simultaneous 24-hour ambulatory monitoring of cough and reflux events, with the ultimate goal of quantifying the temporal correlation between the two types of events.*

***Keywords**: Biomedical signal processing, cough detection, gastroesophageal reflux disease.*

***ACM Classification Keywords**: I.5.4 Pattern Recognition: Applications – Signal processing; J.3 Life and Medical Sciences*

1. Introduction

The latest comprehensive statistics provided by the National Institutes of Health suggests that gastroesophageal reflux (GER) and related symptoms affect approximately 20% of the US population, resulting in over 700,000 annual hospitalizations [1]. GER, which occurs in healthy individuals without causing discomfort, is characterized by the movement of gastric content from the stomach into the esophagus. Reflux episodes are considered a disorder, known as gastroesophageal reflux disease (GERD), when the episodes become frequent, prolonged, and/or are ineffectively cleared back into the stomach. An individual suffering from GERD shows problematic symptoms, such as heartburn, discomfort, and chest pain, which result due to the prolonged exposure of the refluxed material with the distal esophageal mucosa. Excessive exposure can lead to further consequences, such as esophagitis and esophageal ulceration.

GERD is commonly cited as a significant cause of chronic cough [2-5]: In multiple studies, GERD has been documented to be a cause of chronic persistent cough in 38 to 82% of patients [2]. Irwin et al. report that GER can cause cough either by aspiration of gastric content, or by stimulation of the distal esophagus due to repetitive or prolonged GER events [3]. To determine if a patient suffering from chronic cough should be treated for GERD, it is important to quantify whether there is a causal relationship between cough and reflux. In addition to facilitating the diagnosis, more accurate identification of the temporal correlation between cough and reflux can lead to a more appropriate treatment [6].

Numerous studies have focused on determining this temporal correlation [5-8]. However, they rely on manual identification of cough, either through the interpretation of user diaries and user-triggered events [5, 7, 8], or through the manual analysis of manometric recordings, identifying cough as simultaneous, short duration, rapid pressure rises (time to peak less than 1 second) across multiple manometric recording sites [6]. This results in a time-consuming process, prone to user error.

In this paper, we present an algorithm for accurately identifying coughs based on a single-channel manometric recording using a custom pressure probe optimized for cough detection [9]. The algorithm can be implemented in real-time, allowing a simultaneous indication of cough events with the recording of reflux episodes. The two-stage algorithm, which consists of a filtering stage followed by a decision stage, has some common elements with the real-time QRS detection algorithm proposed by Pan and Tompkins [10, 11]. In particular, the cough detection algorithm has a similar filtering stage to its QRS counterpart, with both algorithms sharing the goal of reducing high-frequency noise components and removing low-frequency baseline drifts and pressure changes, thus amplifying the pressure response of interest. The filtering stage of the proposed algorithm differs mainly in the bandpass frequency, which is specific to coughs, and in the utilization of integration in a time window corresponding to the typical duration of a cough. The decision stage of the cough detection algorithm has also been adapted specifically to isolate the pressure response characteristics of a cough.

2. Methodology

2.1 Algorithm Description

2.1.1 Overview

Initially, the presented cough detection algorithm employs a combination of bandpass filtering, differentiation, and moving window integration to magnify the short, rapid pressure rises characteristic of cough events and to reduce the baseline drift and the low-frequency pressure variations characteristic of esophageal peristalsis. The second stage of the algorithm consists of decision logic which performs the recognition process, determining the occurrence of a cough based on a combination of dual-threshold and width detection. To prove feasibility of the algorithm on acquired manometry tracings, development was initially performed in Matlab.

In the following sections, each stage of the algorithm is described in more detail.

2.1.2 Filtering Stage

The analysis of manometry tracings acquired using a single-channel fiberoptic pressure catheter showed that the frequency components of the pressure rise associated with a cough typically range from 2.5 Hz to 5 Hz. An initial bandpass filter stage is used to attenuate signal components and noise artifact outside this frequency range. The bandpass filter is achieved using cascaded low-pass and high-pass filters. Both filters are zero-phase, least-squares, 17th order FIR filters with a maximum passband ripple of 0.05 dB. The low-pass filter, which has a 3-dB frequency of 5 Hz, reduces high-frequency noise. The high-pass filter, which has a 3-dB frequency of 2.5 Hz, reduces lower-frequency peristaltic contractions and baseline drifts. Following the bandpass filtering stage, the signal is differentiated. The differentiation stage focuses on identifying cough by its characteristically steep slope. The resulting signal is then rectified, making all data points positive facilitating the decision process. Finally, the signal is integrated to provide an average of the output, thus incorporating the width of the cough episode into the output signal.

2.1.3 Decision Stage

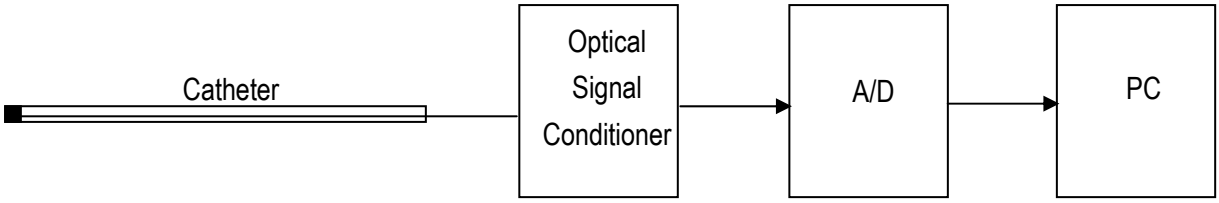
The output of the filtering stage results in the amplification of cough events and the reduction of high-frequency noise and low-frequency pressure rises due to peristalsis. Because this stage dramatically improves the cough signal-to-noise ratio, the decision stage of the cough detection process is simplified. Cough detection on the filtered data relies on first determining the maximum peak within a given sampling interval. This can easily be determined recursively in real-time. The threshold is then adjusted automatically based on the current maximum

peak. The threshold has been empirically set to 0.33 of the maximum peak of the current interval. By periodically adjusting the threshold based on the maximum peak of the current sampling interval, the algorithm adapts to changing characteristics in the signal. After determining the threshold, the algorithm searches for a dual crossing of the threshold, with the first crossing registering an increasing slope, the second crossing registering a negative slope, and the time interval between the crossings falling within the cough duration parameter. The cough duration has been set to 400 ms, based on measuring cough durations on manometry tracings.

2.2 Evaluating the Algorithm

Evaluation of the discussed algorithm was performed using esophageal manometry recordings acquired with a single-channel pressure catheter optimized for cough detection. The catheter [9], comprises a miniature white-light interferometric fiber optic pressure sensor encapsulated in such a manner as to optimize sensitivity to pressure at the catheter tip, such as that experienced during a cough event, while reducing sensitivity to circumferential force, such as that experienced during esophageal peristalsis. The esophageal probe interfaces to an optical signal conditioner which converts the optical signal to an analog output. The analog output is sampled at 50 Hz using a National Instruments PCMCIA data acquisition board and a custom-designed real-time software application on a laptop computer. A block diagram of the system is shown in Figure 1.

Figure 1 - Block diagram of experimental setup.



In vivo esophageal recordings were acquired from a healthy volunteer. Three separate trials were performed to ensure repeatability. In each trial, the volunteer performed a series of respiratory and gastrointestinal events, such as swallowing, belching, heavy breathing, coughing, throat clearing, and laughing. The volunteer also initiated movement (i.e. head rotation, bending, and standing) to attempt to introduce unwanted artifacts. All events were marked on the recording at the exact time of their occurrence.

3. Results

Preliminary testing of the algorithm on in vivo data samples demonstrated a consistent detection of cough events. Figure 2 shows the original signal and the output of each filtering stage for one in vivo interval recorded from the healthy volunteer. In Figure 2a, the raw signal is shown, with gastrointestinal and respiratory events that occurred during the sampling interval indicated. The output of the low pass filter can be observed in Figure 2b, which reduces the high-frequency noise ripples. Figure 2c, which shows the output of the high pass filter, demonstrates the amplification of the cough events, while attenuating the slower pressure variations resulting from talking and swallowing. The differentiation process amplifies slope information, shown in Figure 2d, which is then rectified resulting in a signal comprising positive data points, shown in Figure 2e. The final output of the moving window integrator is shown in Figure 2f, where it can be observed that the cough signals have been amplified, while other pressure variations have been reduced.

The final result of the cough detection algorithm is depicted on Figure 3, where the cough identification results (Figure 3c) are compared with the raw signal (Figure 3a) and the output of the filtering stage (Figure 3b), which supplies the input signal to the decision process. Despite the presence of noise artifacts and two high pressure variations resulting from peristalsis, the algorithm has successfully identified all cough events by targeting a dual crossing of the established threshold within the set duration. This can be observed in Figure 3c.

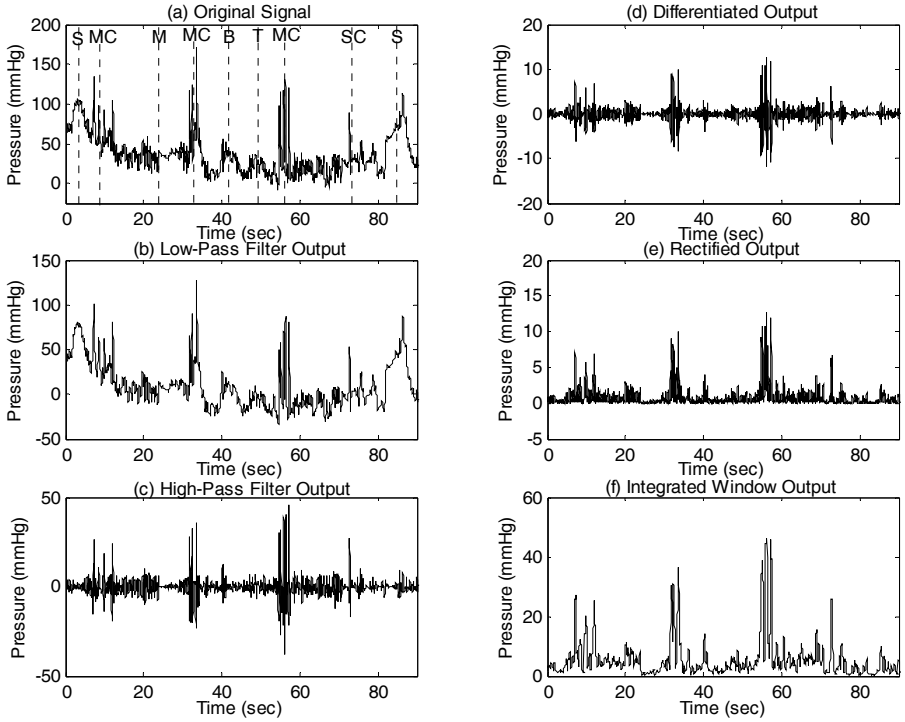


Figure 2 - Sample data of (a) original signal, (b) low-pass filter output, (c) high-pass filter output, (d) differentiated output, (e) rectified output, and (f) integrated window output. *S* indicates a peristaltic contraction resulting from a swallow, *MC* indicates the occurrence of multiple coughs, or a cough burst, *M* indicates patient movement, *B* indicates a belch, *T* indicates patient talking, and *SC* indicates the occurrence of a single cough.

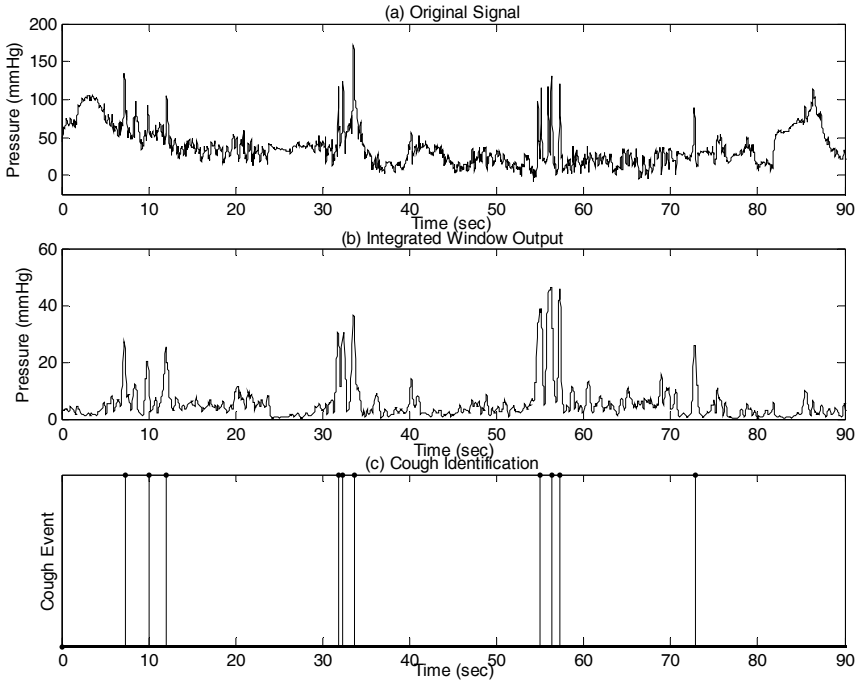


Figure 3 - Sample data showing cough identification, including (a) original signal, (b) integrated window output, and (c) final algorithm results.

4. Discussion

In recent years, numerous studies have attempted to define the temporal correlation between cough events and reflux events in a pursuit to fully understand the relationship between chronic cough and GERD [5-8]. In these investigations, a tool to facilitate the identification of cough episodes without manual analysis of sound recordings, diary annotations, or manometry tracings is still lacking. In the present study, we developed an algorithm for conveniently identifying coughs based on a single manometric channel recording.

Using an algorithm consisting of a filter stage to amplify frequency components associated with the characteristics of cough while attenuating other pressure artifacts, followed by decision logic to identify the cough events from the filtered signal, cough episodes have been correctly identified in samples acquired via manometry. The successful demonstration of the feasibility of the algorithm paves the way for more robust testing using a larger database of manometric readings. Using the algorithm structure presented, specific parameters, including the filter cut-off frequencies, the duration of cough events, and the adaptive threshold setting, may be adjusted to more accurately satisfy the characteristics of cough based on a larger sampling pool of test recordings. Also, a further qualifier can be added to the real-time implementation, in which the definition of a cough can be modified to include bursts of two, three, or more cough events.

5. Conclusion

An innovative algorithm for identifying cough events based on manometric recordings was developed. Initial testing of the algorithm was performed on in vivo data samples acquired using a single-channel pressure catheter. Preliminary studies indicate that the algorithm is suitable for cough detection. By integrating the technique with a single-channel pressure and pH/impedance catheter for 24-hour ambulatory monitoring of cough and reflux, a suitable diagnostic tool for GERD could be achieved.

Acknowledgement

This study was supported in part by the Natural Sciences and Engineering Research Council of Canada, the Alberta Ingenuity Fund, and the Informatics Centre of Research Excellence.

Bibliography

[1] United States National Institute of Health, “Digestive disease statistics”, [Online]. Available: http://digestive.niddk.nih.gov/statistics/statistics.htm. Accessed on January 26, 2006.

[2] K. F. Chung, J. G. Widdicombe, and H. A. Boushey, *Cough: Mechanisms and Therapy*, Oxford, Great Britain: Blackwell Publishing, pp. 97-106, 2003.

[3] R. S. Irwin, C. T. French, F. J. Curley, J. K. Zawacki, and F. M. Benett, “Chronic cough due to gastroesophageal reflux: clinical, diagnostic, and pathogenetic aspects,” *Chest*, vol. 105, pp. 1511-1517, 1993.

[4] Y. W. Novitsky, J. K. Zawacki, R. S. Irwin, C. T. French, V. M. Hussey, and M. P. Callery, “Chronic cough due to gastroesophageal disease: efficacy of antireflux surgery,” *Surgical Endoscopy*, vol. 16, no. 4, pp. 567-571, 2002.

[5] A. W. Wunderlich and J. A. Murray, “Temporal correlation between chronic cough and gastroesophageal reflux disease”, *Digestive Diseases and Sciences*, vol. 48, no. 6, pp. 1050-1056, 2003.

[6] D. Sifrim, L. Dupont, K. Blondeau, X. Zhang, J. Tack, and J. Janssens, “Weakly acidic reflux in patients with chronic unexplained cough during 24 hour pressure, pH, and impedance monitoring,” *Gut*, vol. 54, pp. 449-454, 2005.

[7] S. M. Harding, M. R. Guzzo, and J. E. Richter, “24-h esophageal pH testing in asthmatics,” *Chest*, vol. 115, no. 3, pp. 654-659, 1999.

[8] B. Avidan, A. Sonnenberg, T. G. Schnell, and S. J. Sontag, “Temporal associations between coughing or wheezing and acid reflux in asthmatics,” *Gut*, vol. 49, pp. 767-772, 2001.

[9] J. A. Hogan and M. P. Mintchev, "Method and apparatus for intra-esophageal cough detection," Technical Report submitted to University Technologies International, Calgary, Alberta, March, 2006.

[10] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions in Biomedical Engineering*, vol. 32, no. 3, pp. 230-236, 1985.

[11] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmiac database," *IEEE Transactions in Biomedical Engineering*, vol. 33, pp. 1157-1165, 1986.

Authors' Information

Jennifer A. Hogan – M.Sc. graduate student, Department of Electrical and Computer Engineering; University of Calgary, Calgary, Alberta, Canada, T2N 1N4

Martin P. Mintchev – Prof., Dr., Department of Electrical and Computer Engineering; University of Calgary; Calgary, Alberta, Canada, T2N 1N4; Department of Surgery, University of Alberta; Edmonton, Alberta T6G 2B7

Phone: (403) 220-5309; Fax (403) 282-6855; e-mail: mmintche@ucalgary.ca

LOW-POWER TRACKING IMAGE SENSOR BASED ON BIOLOGICAL MODELS OF ATTENTION

Alexander Fish, Liby Sudakov-Boreysha, Orly Yadid-Pecht

Abstract*: This paper presents implementation of a low-power tracking CMOS image sensor based on biological models of attention. The presented imager allows tracking of up to N salient targets in the field of view. Employing "smart" image sensor architecture, where all image processing is implemented on the sensor focal plane, the proposed imager allows reduction of the amount of data transmitted from the sensor array to external processing units and thus provides real time operation. The imager operation and architecture are based on the models taken from biological systems, where data sensed by many millions of receptors should be transmitted and processed in real time. The imager architecture is optimized to achieve low-power dissipation both in acquisition and tracking modes of operation. The tracking concept is presented, the system architecture is shown and the circuits description is discussed.*

Keywords: *Low-power image sensors, image processing, tracking imager, models of attention, CMOS sensors*

ACM Classification Keywords*: B.7.0 Integrated circuits: General, I.4.8 Image processing and computer vision: scene analysis: tracking*

1. Introduction

Real time visual tracking of salient targets in the field of view (FOV) is a very important operation in machine vision, star tracking and navigation applications. To accomplish real time operation a large amount of information is to be processed in parallel. This parallel processing is a very complicated task that demands huge computation resources. The same problem exists in biological vision systems. Compared to the state-of-the-art artificial imaging systems, having about twenty millions sensors, the human eye has more than one hundred million receptors (rods and cones). Thus, the question is how biological vision systems succeed to transmit and to process such a large amount of information in real time? The answer is that to cope with potential overload, the brain is equipped with a variety of attentional mechanisms [1]. These mechanisms have two important functions:

(a) attention can be used to select relevant information and/or to ignore the irrelevant or interfering information; (b) attention can modulate or enhance the selected information according to the state and goals of the perceiver. Most models of attention mechanisms are based on the fact that a serial selection of regions of interest and their subsequent processing can greatly facilitate the computation complexity. Numerous research efforts in physiology were triggered during the last five decades to understand the attention mechanism [2]-[10]. Generally, works related to physiological analysis of the human attention system can be divided into two main groups: those that present a spatial (spotlight) model for visual attention [2]-[4] and those following object-based attention [5]-[10]. The main difference between these models is that the object-based theory is based on the assumption that attention is referenced to a target or perceptual groups in the visual field, while the spotlight theory indicates that attention selects a place at which to enhance the efficiency of information processing.

The design of efficient real time tracking systems mostly depends on deep understanding of the model of visual attention. Thus, a discipline, named neuromorphic VLSI that imitates the processing architectures found in biological systems as closely as possible was introduced [11]. Both spotlight and object-based models have been recently implemented in analog neuromorphic VLSI design [12]-[23]. Most of them are based on the theory of selective shifts of attention which arises from a saliency map, as was first introduced by Koch and Ullman [12]. Object-based selective attention systems VLSI implementations in 1-D and lately 2-D were presented by Morris et al [13]-[16]. An additional work on an analog VLSI based attentional search/tracking was presented by Horiuchi and Niebur in 1999 [17].

Many works on neuromorphic VLSI implementations of selective attention systems have been presented by Indiveri [19]-[21] and others [22]-[23]. In 1998 Brajovic and Kanade presented a computational sensor for visual tracking with attention. These works often use winner-take-all (WTA) [24] networks that are responsible for selection and tracking inputs with the strongest amplitude. This sequential search method is equivalent to the spotlight attention found in biological systems.

Most previously presented neuromorphic imagers utilize image processing implemented on the focal plane level and employ photodiode or phototransistor current-mode pixels. Typically, each pixel consists of a photo detector and local circuitry, performing spatio-temporal computations on the analog signal. These computations are fully parallel and distributed, since the information is processed according to the locally sensed signals and data from pixel neighbors. This concept allows reduction in the computational cost of the next processing stages placed in the interface. Unfortunately, when image quality and high spatial resolution are important, image processing should be performed in the periphery. This way a high fill factor (FF) can be achieved even in small pixels.

This paper presents implementation of a low-power tracking CMOS image sensor based on a spotlight model of attention. The presented imager allows tracking of up to N salient targets in the field of view. Employing image processing at the sensor focal plane, the proposed sensor allows parallel computations and is distributed, but on the other hand most of the image processing is performed in the array periphery, allowing image quality and high spatial resolution. The imager architecture is optimized to achieve low-power dissipation both in acquisition and tracking modes of operation. This paper is a continuation of the work presented in [25], where we proposed to employ a spotlight model of attention for the bottleneck problem reduction in high resolution "smart" CMOS image sensors and of the work presented in [26], where the basic concept for an efficient VLSI tracking sensor was presented.

Section 2 briefly describes spotlight and object-based models of attention and presents system architecture of the proposed sensor. Low-power considerations, as well imager circuits description are shown in Section 3. Section 4 discusses advantages and limitations of the proposed system. Conclusions and future work are presented in Section 5.

2. Tracking Sensor Architecture

The proposed tracking sensor operation is based on the imitation of the spotlight model of visual attention. Because this paper presents concepts taken from different research disciplines, first, a brief description of existing models of attention is presented for the readers that are not familiar with this field. Then, the proposed sensors architecture is shown.

2.1 Existing Attention Models

Much research was done in attention during the last decades and numerous models have been proposed over the years. However, there is still much confusion as to the nature and role of attention. All presented models of attention can be divided to two main groups: spatial (spotlight) or early attention and object-based, or late attention. While the object-based theory suggests that the visual world is parsed into objects or perceptual groups, the spatial (spotlight) model purports that attention is directed to unparsed regions of space. Experimental research provides some degree of support to both models of attention. While both models are useful in understanding the processing of visual information, the spotlight model suffers from more drawbacks than the object-based model. However, the spotlight model is simpler and can be more useful for tracking imager implementations, as will be shown below.

2.1.1 The Spatial (Spotlight) Model

The model of spotlight visual attention mainly grew out of the application of information theory developed by Shannon. In electronic systems, similar to physiological, the amount of the incoming information is limited by the system resources. There are two main models of spotlight attention. The simplest model can be looked upon as a spatial filter, where what falls outside the attentional spotlight is assumed not to be processed. In the second model, the spotlight serves to concentrate attentional resources to a particular region in space, thus enhancing processing at that location and almost eliminating processing of the unattended regions. The main difference between these models is that in the first one the spotlight only passively blocks the irrelevant information, while in the second model it actively directs the "processing efforts" to the chosen region.

Figure 1(a) and Figure 1(b) visually clarify the difference between the spatial filtering and spotlight attention.

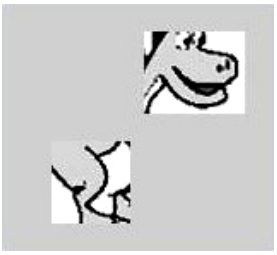


Figure 1 (a). An example of spatial filtering

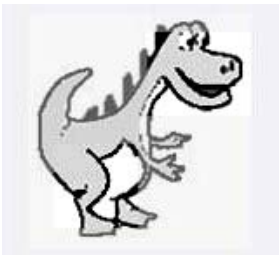


Figure 1 (b). An example of spotlight model of attention

A conventional view of the spotlight model assumes that only a single region of interest is processed at a certain time point and supposes smooth movement to other regions of interest. Later versions of the spotlight model assume that the attentional spotlight can be divided between several regions in space. In addition, the latter support the theory that the spotlight moves discretely from one region to the other.

2.1.2 Object-based Model

As reviewed above, the spotlight metaphor is useful for understanding how attention is deployed across space. However, this metaphor has serious limitations. A detailed analysis of the spotlight model drawbacks can be found in [1]. An object-based attention model suits more practical experiments in physiology and is based on the assumption that attention is referred to discrete objects in the visual field. However being more practical, in

contrast to the spotlight model, where one would predict that two nearby or overlapping objects are attended as a single object, in the object-based model this divided attention between objects results in less efficient processing than attending to a single object. It should be noted that spotlight and object-based attention theories are not contradictory but rather complementary. Nevertheless, in many cases the object-based theory explains many phenomena better than the spotlight model does.

The object-based model is more complicated for implementation, since it requires objects' recognition, while the spotlight model only requires identifying the regions of interest, where the attentional resources will be concentrated for further processing.

2.2 System Architecture

The proposed sensor has two modes of operation: target acquisition and target tracking. In the acquisition mode *N* most salient targets of interest in the FOV are found. Then, *N* windows of interest with programmable size around the targets are defined. These windows define the active regions, where the subsequent processing will occur, similar to the flexible spotlight size in the biological systems. In the tracking mode, the system sequentially attends only to the previously chosen regions, while completely inhibiting the dataflow from the other regions.

The proposed concept permits choosing the attended regions in the desired order, independent on the targets saliency. In addition it allows shifting the attention from one active region to the other, independent of the distance between the targets. The proposed sensor aims to output the coordinates of all tracking targets in real time. Similar to biological systems, which are limited in their computational resources, the engineering applications are constrained with low-power dissipation. Thus, maximum efforts have been done to reduce power consumption in the proposed sensor. This power reduction is based on the general idea of "no movement – no action", meaning that minimum power should be dissipated if no change in the targets position occurred.

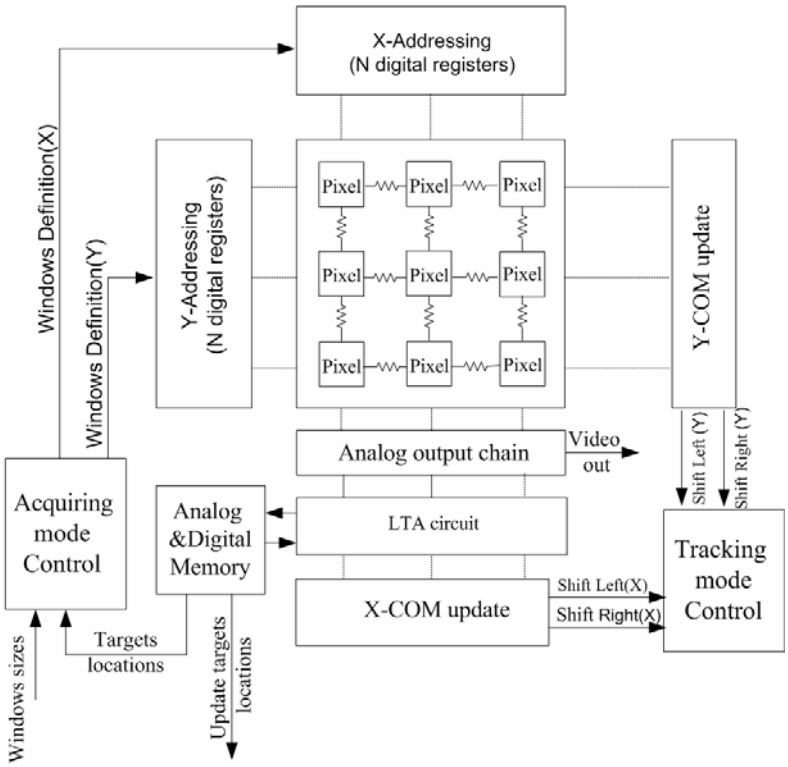


Figure 2. Architecture of the proposed tracking sensor.

Figure 2 shows the architecture of the proposed tracking sensor. The sensor includes (a) A Pixel array with a two dimensional resistive network , (b) Y-Addressing and X-Addressing circuitry consisting of N digital registers for target windows definition and for image readout control, (c) an analog front end (AFE) for image readout, (d) A current Looser-take-all (LTA) circuit for target detection during the acquisition mode, (e) two analog 1-D center of mass (COM) computation circuits for X and Y COM coordinates update (each consists of 1-D analog current mode Winner-take-all (WTA) circuit), (f) Analog memory for temporal storage of loser values of all rows during the acquisition mode and digital memory for targets coordinates storage, (g) acquisition mode control and target mode control blocks.

In the acquisition mode the sensor finds the *N* most salient targets in the FOV and calculates their centroid coordinates. This is achieved in the following way: all neighboring pixels are connected by resistors (implemented by transistors), creating the resistive network. The pixel voltage becomes smaller if it is more exposed, and the local minimum of the voltage distribution can be regarded as the centroid of the target, corresponding to the exposed area. This minimum is found using an analog looser-take-all circuit (LTA). At the first stage of the acquisition mode, all pixels of the whole image are activated. The global minimum, corresponding to the brightest target is located using a one dimensional LTA circuit. To achieve this purpose, the whole image is scanned row by row (using one of the digital shift registers in the Y-addressing circuitry), finding the local minimum in each row. Then, the row local minima are input to the same LTA circuit again and the global minimum is computed. A more detailed description of this concept can be found in [27], where the two dimensional WTA computation was performed using two 1-D WTA circuits. Once the first brightest target is found, the system defines a small size programmable window, with the center located at the target centroid coordinates. The size of this window is predefined by the user before the acquisition mode starts and depends on the target size. While finding the second bright target in the FOV, all pixels of the first window, consisting of the brightest target found during the first search, are deactivated. This way, the bright pixels of the first target do not influence the result of the second search. The remains *N-1* targets are found in the same way. As a result, at the end of the acquisition mode all centroid coordinates of the *N* most salient targets in the FOV are stored in the memory and *N* small windows around these coordinates are defined. The window definition is performed using two digital shift registers. Thus, 2N shift registers are required to define N different windows. The acquisition mode control block is responsible for defining and positioning these active windows. Note, that the acquisition mode is very inefficient in terms of power dissipation because the whole sensor array is activated and the LTA operation and windows definition are power inefficient operations. On the other hand, the acquisition is a very rare operation and its time can be neglected in comparison with the tracking period.

Once the sensor has acquired *N* salient targets, the tracking mode is initiated. The predefined windows serve as a spotlight in biological systems, such that only the regions inside the windows are processed. Opposite to biological systems, these "spotlights" attend only to the regions predefined in the acquisition mode. Thus, even if new more salient objects appear during the tracking, the attention to the chosen regions is not influenced.

Because the sensor is in the tracking mode most of the time, it is very important to achieve very low-power dissipation in this mode. In the proposed system this is achieved in the following ways:

1. Only pixels of active windows and the circuitry responsible for proper centroid detection and pixels readout are active. The remaining circuits (including most pixels of the array) are disconnected from the power supply.
2. All shift registers in Y-addressing and X-addressing circuitries are optimized for low frequencies operation by leakage currents reduction.
3. During the tracking mode the sensor doesn't calculate new centroid coordinates. A simple analog circuit (COM update block in Figure 2) checks if the new centroid location differs from the centroid location of the previous frame. In the case that no difference was found, the circuit does not need to perform any action, significantly reducing system power dissipation. This principle suits the general idea of "no movement – no action". If the target changes its position, the "shift left" or "shift right" (both for x and y) signals are produced

by the COM update blocks. These signals are input to the tracking mode control block and the appropriate shift register performs movement to the right direction, correcting the location of the window.

4. Each active window definition is performed using two shift registers. This windows definition method allows switching from one target of interest to another without any need in accessing the memory and loading the new target coordinates. This way the switching time between different objects does not depend on the distance between the targets and sensor power dissipation is reduced.

3. Circuits Description

In this Section we present some of the most important circuits, utilized by the sensor. This includes current mode LTA circuit, current mode WTA circuit, X-COM and Y-COM update circuits and ultra low-power shift registers. The pixel is implemented as a standard global-shutter active pixel sensor (APS) [28] with current mode readout instead of a conventional source follower amplifier inside the pixel. This current mode readout allows parallel read out and summarization of currents from all pixels in the active window at the same time. These summarized currents then are used for further processing by the X-COM and Y-COM update circuits, as described below. In addition, a conventional video data readout is available.

3.1 Current Mode Loser-Take-All (LTA) circuit

As previously mentioned, LTA circuit is responsible for targets detection and their COM computation during the acquisition mode. Since the COM computation is done in a serial manner and should be performed accurately, high speed and high precision LTA circuit is required. In addition, current mode LTA is required (pixels outputs are currents). Most of the previously presented LTA solutions are not suitable for the proposed sensor design. As a result, we use a LTA circuit that we have recently developed to achieve high speed and high precision [29].

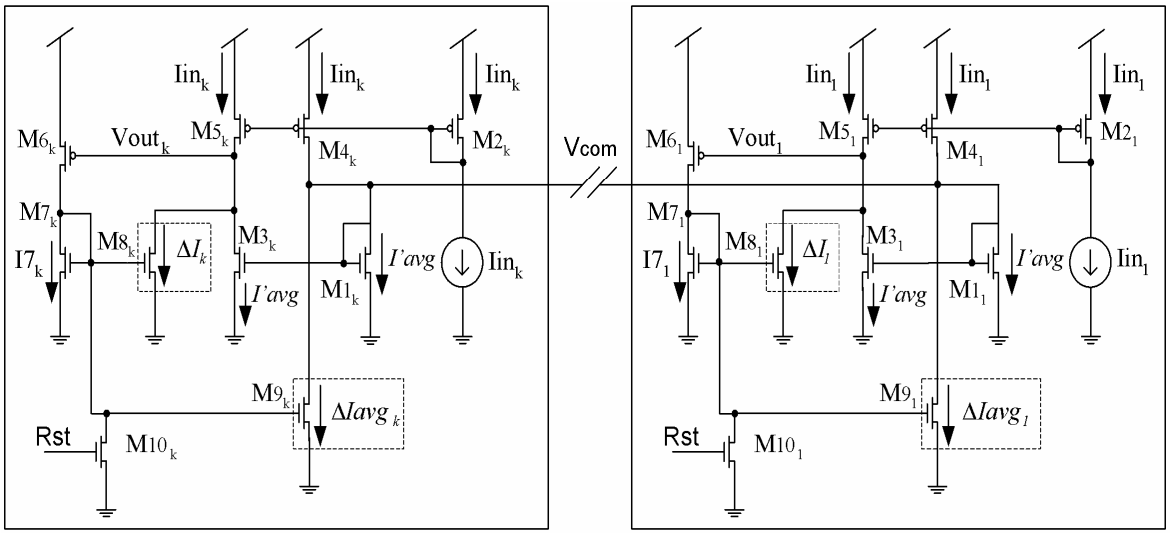


Figure 3: Cells 1 and k (out of N) of the LTA circuit.

Figure 3 shows cells 1 and k (out of the N interacting cells) of the LTA circuit. Cell k receives a unidirectional input current, I_{ink}, and produces an output voltage V_{outk}. This output has a low digital value if the input current I_k is identified as loser, and high, otherwise. The circuit applies two feedbacks to enhance the speed and precision: the excitatory feedback ΔI_k and inhibitory feedback ΔI_{avgk}. The basic operation of the LTA is based on input currents average computation and comparison of the input current of each cell to that average. The local excitatory feedback works to increase the compared average value in each cell, allowing that cell to be a loser. Oppositely, the inhibitory feedback works globally by reduction of the input currents average value and thus allowing inhibition of non-losing cells. A more detailed description of the circuit operation is provided below. The

LTA circuit operates as follows: the drains of M_1 transistors of all N cells of the array are connected to the drains of M_4 transistors by a single common wire with voltage Vcom. The circuit starts the competition by applying Rst='1' for a short period of time. This way the excitatory feedback ΔIin and the inhibitory feedback $\Delta Iavg_k$ are cancelled. Assuming that all N cells in the array are identical and Rst='1' is applied, $\Delta Iavg_k$ = 0A and the current I'avg, through $M1_k$, is equal to the average of all input currents of the array, neglecting small deviations in the referenced input currents. I'avg is copied to $M3_k$ by the NMOS current mirror ($M1_k$ and $M3_k$) and is compared with the input current Iin_k copied by the PMOS current mirror ($M2_k$ and $M5_k$). If Iin_k=I'avg then $Vout_k$=VDD/2, assuming the same drivability factor K of $M3_k$ and $M5_k$ transistors.

An increase in input current Iin_k relatively to I'avg causes an increase in $Vout_k$ due to the Early effect. This way, during the reset phase, input currents of all cells are compared to the average of all input currents of the array, producing a unique output $Vout_k$ for every cell. The cell having the smallest input current value produces the smallest $Vout_k$ voltage. With the completion of the reset phase, i.e. Rst='0', the excitatory feedback ΔI_k and the inhibitory feedback $\Delta Iavg_k$ are produced. The $Vout_k$ node inputs to the gate of $M6_k$ PMOS transistor, thus the cell with the smaller Vx_k (smaller input current) produces a higher current $I7_k$ through $M6_k$ and $M7_k$. This current is copied by the NMOS current mirror ($M7_k$ and $M8_k$), creating the excitatory feedback ΔI_k. On the other hand, $I7_k$ is copied by the NMOS current mirror ($M7_k$ and $M9_k$), resulting in inhibitory feedback $\Delta Iavg_k$. ΔIk is added to the I'avg flowing through $M3_k$ and $\Delta Iavg_k$ is subtracted from the average of all input current by connection $M9_k$ transistor to the COM node, decreasing the I'avg value. This way, every cell produces a new Voutk voltage value, according to the comparison between the input current Iin_k and a sum of a current produced by the excitatory feedback ΔI_k and a new value of current I'avg, that is now given by:

$$I'avg = \frac{\sum_{k=1}^{N} Iin_k - \sum_{k=1}^{N} \Delta Iavg_k}{N} = Iavg - \frac{\sum_{k=1}^{N} \Delta Iavg_k}{N} \quad (1)$$

where Iavg is the average of all input currents of the array and N is the number of array cells. For the cell, having the smallest input current, the difference between Iin_k and a sum of ΔI_k and I'avg grows, thus decreasing $Vout_k$ value.

The computation phase is finished after one cell only is identified as a loser, producing $Vout_k$='0'. All other cells are identified as winners with $Vout_k$='1'. In this steady-state the excitatory and inhibitory feedbacks of the all winner cells and I'avg are approximately equal to zero, while $\Delta Iavg_k$ of the loser cell is approximately equal to the sum of all input currents. This way the circuit states stable preventing the selection of other potential losers unless the next reset is applied and a new computation starts. A more detailed description on the circuit operation can be found in [29].

To examine the presented LTA circuit it was designed, simulated and fabricated in 0.35um, 3.3V, n-well, 4-metal, CMOS, TSMC technology process supported by MOSIS. Table 1 summarizes the main characteristics of the circuit. As can be seen, the circuit achieves both high precision and high speed.

Parameter	Typical value	Worst case value (if exists)
Range of input currents	4 – 25 [μA]	------
Voltage supply	1.8V	------
Power Dissipation	58uW per cell	75 μW per cell
Delay	5nsec	95nsec
Precision	0.1 μA	0.5 μA
Occupied area (per cell)	26μm*22μm	------

Table 1: The main characteristics of the designed circuit.

3.2 X-COM and Y-COM update circuits

As mentioned, during the tracking mode the sensor doesn't calculate the new centroid coordinates. Instead, very simple X-COM and Y-COM update circuits (see Figure 4) check if the new X or Y centroid locations (respectively) differ from the centroid locations of the previous frame. In the case that no difference was found, the sensor does not perform any action, significantly reducing system power dissipation. The X-COM and Y-COM circuits have the same implementation, consisting of (K+1) controlled resistors (implemented by transistors), (K+2) digital switches, controlled by window location and (K+2) size analog current mode WTA circuit for the array, having K columns/row, respectively. Each COM update circuit receives K unidirectional input currents from the sensor array. The input current I_{ink} to the X-COM update circuit is the sum of all output pixel currents of the column K, while from the input current I_{ink} to the Y-COM update circuit is the sum of all output pixel currents of the row K. During the tracking mode (the COM update circuits are activated only in this mode), only pixels inside the windows of interest are activated. Therefore, in this mode, the input current to the COM circuit I_{ink} represents the sum of all row/column K active window output currents, for the case where the row/column K falls into the window of interest. In case, where the row/column K falls out the window of interest, the value of I_{ink} is zero. All input currents input to the resistive network, consisting of (K+1) controlled resistors (can either have a normal resistance R or very high resistance R_{high}) and then routed to the WTA circuit by switches. Both switches and the resistors are controlled by the windows of interest locations. For the active window, spread between column *i* and column (*i*+c), the resistors 1 to (*i*-1) and (*i*+2) to (K+1) have very high resistance R_{high}, while the resistors *i* to (*i*+1) are set to have a normal resistance R. This way resistive network, consisting of R value resistors is created around the active window of interest. Only two switches (*i*-1) and (*i*+1) are set to be on, while the others are set to be off. As a result, the WTA circuit receive only two non-zero input currents, representing the X/Y COM coordinates of the target located inside the window of interest. In case these currents are equal (the WTA circuit does not succeed to find the winner), the COM coordinates are exactly in the center of the window and therefore window location update is not required. On the other hand, if (i-1) current is larger than (i+1), it means that the target has moved right relatively to its location in the previous frame and Shift Right output is activated. In case if (i-1) current is smaller than (i+1), the Shift Left output is activated.

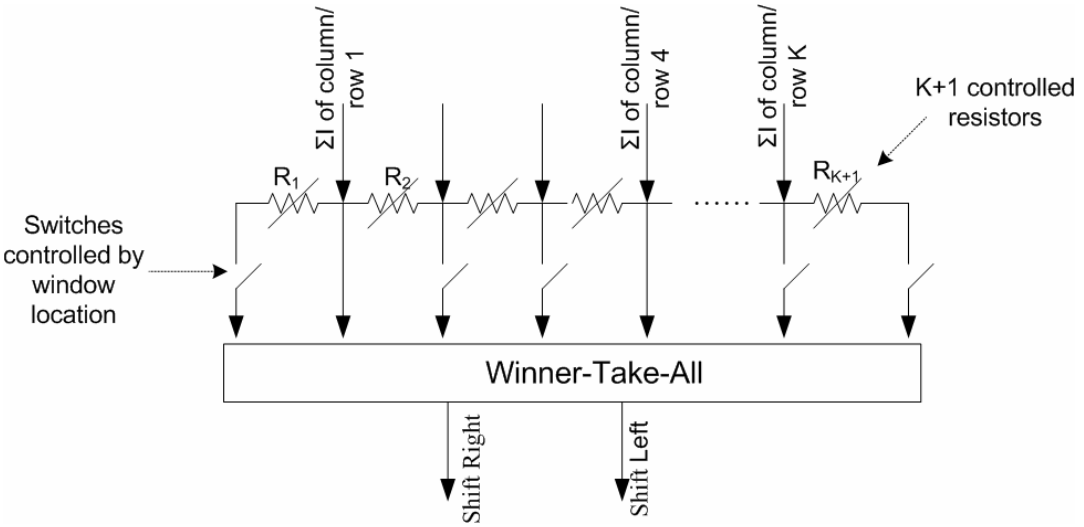


Figure 4 The X-COM and Y-COM update circuit implementation

3.3 Current Mode Winner-take-all (WTA) circuit

As mentioned, the X-COM and Y-COM update circuits utilize current mode WTA circuit. The implementation of the WTA circuit (see Figure 5) is very similar to the implementation of the LTA circuit, presented in sub-section 3.1 and detailed description of its operation can be found in [30]. Similarly to the LTA, the WTA circuit employs inhibitory and local excitatory feedbacks based on input currents average computation, enhancing precision and speed performance of the circuit. Local excitatory feedback provides a hysteretic mechanism that prevents the selection of other potential winners unless they are stronger than the selected one by a set hysteretic current. The WTA circuit can be useful for integration with circuits operating in the strong inversion region and supplying input currents of 3μA-50μA, as well as for subthreshold applications with inputs of 0nA-50nA. It achieves very high speed (32nsec for high currents of 3μA-50μA (measured) and 34nsec for subthreshold currents – (simulated)) in case when a very small difference between two input currents is applied (30nA for high currents and 1.8nA for subthreshold applications). These circuit performances are the direct result from very strong feedbacks applied in the circuit. The circuit ability to cope with wide range of input currents is very important for the COM update circuit implementation since the inputs to this circuit can have very wide range.

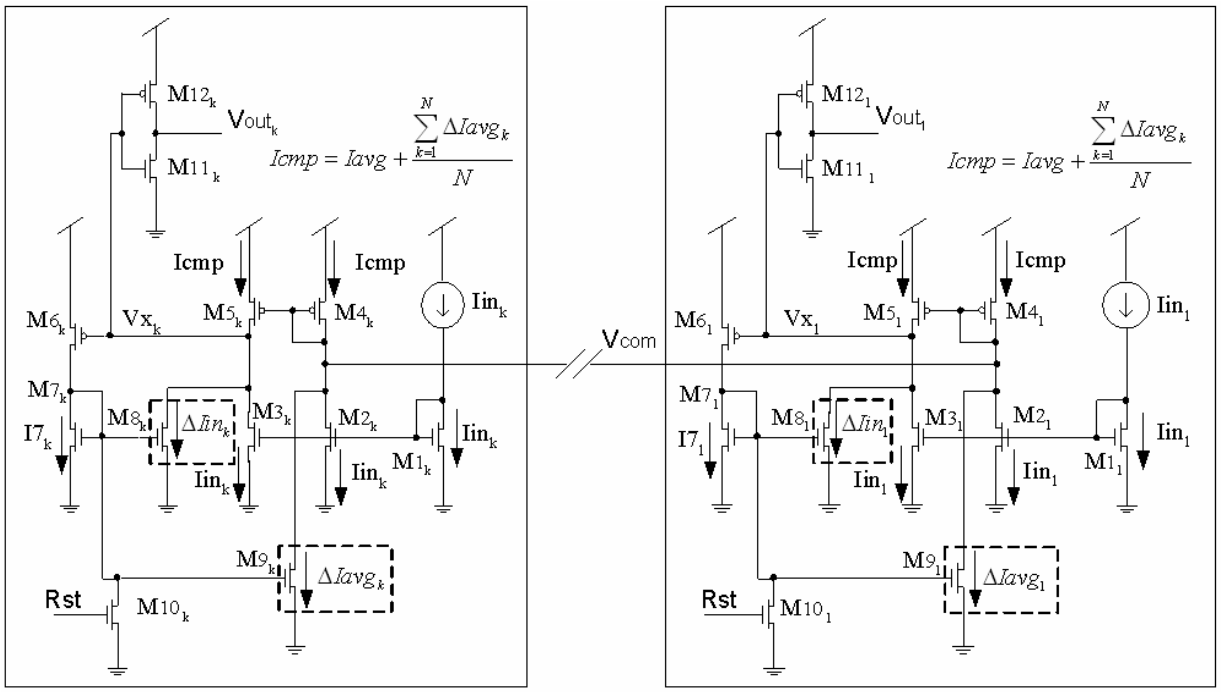


Figure 5. Cells 1 and k (out of n) of the WTA circuit

3.4 Y-addressing and X-Addressing circuitry implementation

Generally, the Y-Addressing and X-Addressing circuitry can be implemented using digital decoders. However, using shift registers for read out control reduces power dissipation and the number of global buses. In addition, windows of interest can be easily defined using shift-register. As previously mentioned, 2N shift registers are required to define N different windows. The architecture of shift register, used in the discussed tracking sensor (see Figure 6) , is based on the conventional simple shift register structure and utilizes low-power D-Flip-Flops (DFFs), described in details in [31].

This register allows shifting of the vector of bits right or left – very important function in windows definition. The power dissipation of the shift register can be reduced by examining the nature of the inputs to the register. When the register is used for signal readout control, its input vector consists of a single '1' and of (N-1) digital zeros ('0'), assuming an N size register. Thus, in steady state, only one (out of N) DFF has '1' in its output. For the case, when the register is used for window definition, its input vector consists of K high digital bits and (N-K) low digital bits, assuming an N size register defines K*K size window. Usually, K<<N, resulting in the same solution for both cases. Figure 7 shows the DFF, optimized for these kinds of input vectors. This master-slave FF is constructed by cascading two different transmission gate latches: the first one is the dynamic latch and the second one is pseudo-static latch. Having only 15 transistors, this circuit is optimized for leakage current reduction for the case of '0' at the FF output. For this case all possible leakage currents in the circuit (signed by arrows in Figure 7) are reduced due to connection of two series connected "off" transistors. Note, that for the case of the shift register, having an input vector with almost all "high" bits, the presented FF can be optimized in a similar way, taking in account the high digital value in the FF output.

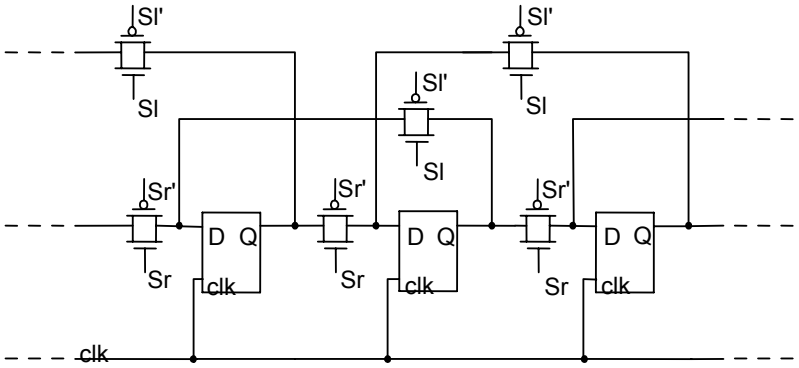


Figure 6 Architecture of the conventional shift-right and shift-left register

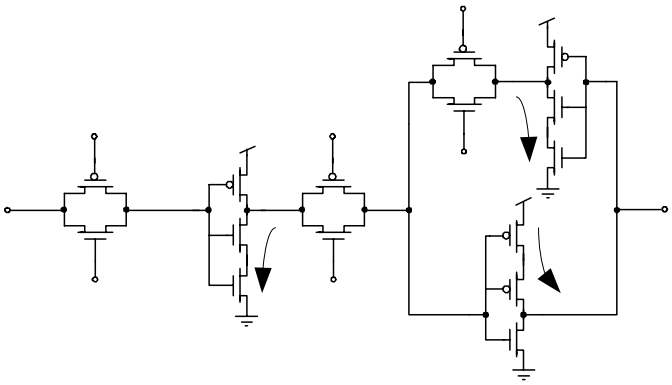


Figure 7. DFF, optimized for an input vector consisting of a large number of zeros

To check the suitability of the mentioned DFF circuit for the proposed tracking system, it has been designed and implemented in 0.18μm technology to compare the FF design with a set of representative flip-flops, commonly used for high performance design. In addition, the low leakage shift register is compared to the register, based on conventional FFs. Simulation results show up-to 10% power reduction in case of 10KHZ operation and up-to 60%

reduction in case of 5 pixels size window definition at 30Hz frequency. More detailed description of the FF and SR performances can be found in [31].

4. Discussion

In this section we present the expected performance of the proposed tracking imager and briefly discuss advantages and limitations of the current architecture. Table 2 summarizes the expected characteristics of the proposed system. The sensor will be fabricated in a standard 0.18μm CMOS technology and will be operated using 1.8V supply voltage. The pixel size is expected to be 7 x 7μm and to achieve fill factor of at least 60%. At this stage the test chip will include a relatively small array of 64 x 64. The reason for the small array is failure probability reduction and limited budget. On the other hand, this array size still allows showing the proof of concept. At the first fabrication phase the system will able to track up to 3 salient targets of interest at 30 frames per second. In future designs we are working to increase the number of the tracked targets and to improve real time operation, allowing tracking at up to 100 frames per second. As mentioned, the proposed imager employs spatial filtering version of the spotlight models of attention, where what falls outside the attentional spotlight is assumed not to be processed. The drawback of this method is that during the tracking mode the sensor filters all information outside windows of interest, including potential targets that appear in the FOV during the tracking. In our future implementations we plan to upgrade the tracking sensor with the spotlight attention model, where the spotlight serves to concentrate attentional resources to a particular region in space, thus enhancing processing at that location and almost eliminating processing of the unattended regions (but still checking theses regions). An additional limitation is that the proposed system does not utilize Correlated Double Sampling (CDS) circuit to reduce Fixed Pattern Noise (FPN). CDS implementation in such kind of tracker is not trivial and thus it will be implemented only in the next version of the system to reduce failure probability at this stage. Finally, the system is expected to achieve very low-power dissipation of less than 2mW. The next generation of this tracking system will achieve power dissipation of less than 1mW.

Parameter	Expected value
Technology	0.18μm standard CMOS technology
Array size	64 x 64
Voltage supply	1.8V
Pixel Size	7 x 7 μm
Fill Factor	> 60%
No. of tracked targets	3
Real time operation	60 frames/second
Sensor read out method	Global Shutter
Utilized Attention model	Spatial Filtering
Bad Pixel Elimination	Yes
FPN Reduction	No
Power Dissipation	< 2mW

Table 2: The expected characteristics of the tracking system.

5. Conclusions

Implementation of low-power tracking CMOS image sensor based on biological models of attention was presented. Imager architecture and principle of operation were discussed, as well designs of the most important circuits, like Winner-Take-All, Looser-Takes-All, COM update and X/Y-Addressing circuits, utilized by the tracking system were shown. A brief description of the spatial and object-based models of attention was also presented. The expected system performance was discussed, showing advantages and drawbacks of the proposed sensor. The presented imager allows tracking of up to N salient targets in the field of view. The imager architecture is optimized to achieve low-power dissipation both in acquisition and tracking modes of operation. Further research includes improvement of the current sensor architecture and its realization in an advanced CMOS technology.

Acknowledgements

We would like to thank the Israeli Ministry of Science and Technology for funding this project.

Bibliography

[1] Chun, M. M., & Wolfe, J. M. Visual Attention. In E. B. Goldstein (Ed.), Blackwell's Handbook of Perception, Vol. Ch 9, pp. 272-310). Oxford, UK: Blackwell, 2001.

[2] R. W. Remington, L. Pierce. Moving attention: Evidence for Time-invariant shifts of visual selection attention. Perception and Psychophysics, vol. 35, pp. 393-399, 1984.

[3] J. Moran, R. Desimone. Selective attention gates visual processing in the Extrastriate Cortex. Science, vol. 229, pp. 784-787, 1985.

[4] G. Sperling, E. Weichselgartner. Episodic theory of the dynamics of spatial attention. Psychological review, vol. 102, no. 3, pp. 503-532, 1995.

[5] A. Treisman. Features and targets: The fourteenth Barlett memorial lecture. Quarterly Journal of Experimental Psychology, 40A, pp. 201-237, 1988.

[6] A. Treisman. Feature binding, attention and target perception. Philosophical Transactions of the Royal Society of London, vol. 353, pp. 1295-1306, 1998.

[7] A. Treisman, G. Gelade. A feature-integration theory of attention. Cognitive Psychology, vol. 12, pp. 97-136, 1980.

[8] A. Treisman, D. Kahneman, J. Burkell. Perceptual targets and the cost of filtering. Perception & Psychophysics, vol. 33, pp. 527-532, 1983.

[9] S. Yantis. Targets, attention, and perceptual experience in “Visual Attention". R. D. Wright (Eds.), pp. 187-214. Oxford, NY: Oxford University Press, 1998.

[10] M. I. Posner. Orienting of attention. Quarterly Journal of Experimental Psychology, vol. 32,pp. 3-25, 1980.

[11] C. Koch, B. Mathur. Neuromorphic vision chips. IEEE Spectrum, vol. 33, pp. 38–46, May 1996

[12] C. Koch and S. Ullman, “Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry,” Human Neurobiology, vol. 4, pp. 219–227, 1985.

[13] T. G. Morris, T. K. Horiuchi, and P. DeWeerth, “Target-Based Selection Within an Analog Visual Attention System”, IEEE Transactions on circuits and systems-11, Analog and Digital Signal processing, vol.45, no.12, December 1998.

[14] T. G. Moris and S. P. DeWeerth, “A Smart-Scanning Analog VLSI Visual-Attention System”, Analog Integrated Circuits and Signal Processing, vol.21, p.67-78, 1999.

[15] T. G. Morris and S. P. DeWeerth, “Analog VLSI excitatory feedback circuits for attentional shifts and tracking.” Analog Integrated Circuits and Signal Processing, 13(1/2), pp. 79–92, May/June 1997.

[16] C. S. Wilson, T. G. Morris, and P. DeWeerth, “A two-dimensional, target-based analog VLSI visual attention system", Twentieth Anniversary Conference on Advanced Research in VLSI, IEEE Computer Society Press: Los Alamitos, CA. Vol. 20. pp. 291-308. March 1999.

[17] T. Horiuchi and E. Niebur, “Conjunction search using a 1-D analog VLSI based attentional search/tracking chip,” In Wills, S. (Ed.), Proceedings of the 1999 Conference in Advanced Research in VLSI., pp. 276-90 Georgia Tech, Atlanta, 1999.

[18] V. Brajovic and T. Kanade, “Computational sensor for visual tracking with attention”, IEEE journal of Solid-State Circuits, Vol.33, No.8, August 1998.

[19] G. Indiveri, “Neuromorphic analog VLSI sensor for visual tracking: Circuits and application examples,” IEEE Trans. on Circuits and Systems II, vol. 46, no. 11, pp. 1337–1347, November 1999.

[20] G. Indiveri, A. M. Whatley, and J. Kramer, “A reconfigurable neuromorphic VLSI multi-chip system applied to visual motion computation,” in Proc. 7th Int. Conf. Microelectronics Neural, Fuzzy and Bio-Inspired Systems; Microneuro’99. Los Alamitos, CA: IEEE Computer Society Press, Apr. 1999, pp. 37–44.

[21] G. Indiveri, "A 2D neuromorphic VLSI architecture for modeling selective attention", Proc. IJCNN2000, 2000.

[22] R. Etienne-Cummings, J. Van der Spiegel, and P. Mueller,“A visual smoot pursuit tracking chip," Advances in Neural Information Processing Systems, D. S. Touretzky, M. C. Mozer, and Hasselmo M. E., Eds. 1996, vol. 8, MIT Press.

[23] C. Higgins and C. Koch, “A modular multi-chip neuromorphic architecture for real-time visual processing,” J. Analog Integrated Circuits Signal Process, vol. 26, no. 3, pp. 195–211, Sept. 2000.

[24] J. Lazzaro., S. Ryckebusch, M. A. Mahowald, and C.A. Mead, “Winner-tale-all networks of O(n) complexity”, ed. D.S.Touretzky, Morgan Kaumann, San Mateo, CA, 1:703-711, 1989.

[25] A. Fish and O. Yadid-Pecht, "Bottleneck Problem Solution using Biological Models of Attention in high resolution tracking sensors", to be appear International Journal on Information Theories and Applications.

[26] A. Fish, A. Spivakovsky, A. Golberg and O. Yadid-Pecht, “VLSI Sensor for multiple targets detection and tracking”, IEEE ICECS, December 2004, CD ROM.

[27] A. Fish, D. Turchin, O. Yadid-Pecht, " An APS with 2-Dimensional winner-take-all selection employing adaptive spatial filtering and false alarm reduction", IEEE Trans. on Electron Devices, Special Issue on Image Sensors, Vol. 50, No. 1, pp. 159-165, January, 2003.

[28] O. Yadid-Pecht, R. Etienne-Cummings. CMOS imagers: from phototransduction to image processing. Kluwer Academic Publishers, 2004.

[29] A. Fish, V. Milrud and O. Yadid-Pecht, “High speed and high resolution current Loser-take-all circuit of o(N) complexity”, IEEE ICECS, December 2004, CD ROM.

[30] A. Fish, V. Milirud and O. Yadid-Pecht, “High speed and high resolution current winner-take-all circuit in conjunction with adaptive thresholding”, IEEE Transactions on Circuits and Systems II, vol. 52, no. 3, pp. 131-135, March 2005.

[31] A. Fish, V. Mosheyev, V. Linkovsky and O. Yadid-Pecht, "Ultra Low-Power DFF based Shift Registers Design for CMOS Image Sensors Applications", IEEE ICECS, Tel-Aviv, Israel, December 2004, CD ROM.

Authors' Information

Alexander Fish – The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel; e-mail: afish@ee.bgu.ac.il

Liby Sudakov-Boreysha – The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel; e-mail: libys@bgu.ac.il

Orly Yadid-Pecht – The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel, and Dept. of Electrical and Computer Engineering, University of Calgary, Alberta, Canada; e-mail: oyp@ee.bgu.ac.il

IMAGE SENSORS IN SECURITY AND MEDICAL APPLICATIONS

Evgeny Artyomov, Alexander Fish, Orly Yadid-Pecht

***Abstract**: This paper briefly reviews CMOS image sensor technology and its utilization in security and medical applications. The role and future trends of image sensors in each of the applications are discussed. To provide the reader deeper understanding of the technology aspects the paper concentrates on the selected applications such as surveillance, biometrics, capsule endoscopy and artificial retina. The reasons for concentrating on these applications are due to their importance in our daily life and because they present leading-edge applications for imaging systems research and development. In addition, review of image sensors implementation in these applications allows the reader to investigate image sensor technology from the technical and from other views as well.*

***Keywords:** Image sensors, security applications, medical applications, low-power CMOS image sensors.*

1. Introduction

Fast development of low-power miniature CMOS image sensors triggers their penetration to various fields of our daily life. Today we are commonly used to meet them in digital still and video cameras, cellular phones, web and security cameras, toys, vehicles, factory inspection systems, medical equipment and many other applications (see Figure 1). The advantages of current state-of-the-art CMOS imagers over conventional CCD sensors are the possibility in integration of all functions required for timing, exposure control, color processing, image enhancement, image compression and analog-to-digital (ADC) conversion on the same chip. In addition, CMOS imagers offer significant advantages in terms of low power, low voltage, flexibility, cost and miniaturization. These features make them very suitable especially for security and medical applications. This paper presents a review of image sensors utilization in part of the security and the medical applications.

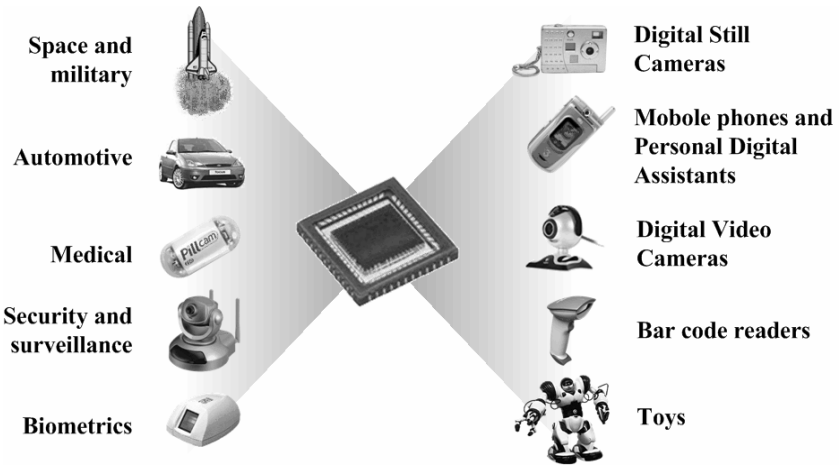


Figure 1. Image sensors applications

During the last few years imaging systems for security applications have been significantly revolutionising. Large, high cost and inefficient cameras mostly used for specific military and government applications have been replaced with compact, low-cost, low-power smart camera systems, becoming available not only for military and government, but for wide spreading in civilian applications. In this paper we will concentrate on two major categories: (a) surveillance systems – usually used for observation, anomaly detection and alarming, employing one or multiple cameras, (b) biometrics systems – used for access control and person identification. Each of the

presented categories requires sensors having different specifications: for example, while low-power and compactness are the most important features for some surveillance systems, robustness and high image quality are the most important requirement in biometric systems.

Medical applications also benefit from the fast image sensors technology development. Introduction of miniature, ultra-low power CMOS image sensors have opened new perspectives to minimally-invasive medical devices, like wireless capsules for gastrointestinal tract observation [32]. Here we will review two very important medical applications:

(a) artificial retina – used as an artificial replacement or aid to the damaged human vision system,

(b) wirelss capsule endoscopy – used in minimally invasive gastrointestinal tract diagnostics.

The remainder of the paper is organized as follows: Section II briefly presents CMOS image sensor technology with reference to "smart" CMOS image sensor architecture. The role of image sensors in security applications is described in Section III. Section IV reviews medical applications employing state-of-the-art CMOS imagers. Section V concludes the paper.

2. CMOS Image Sensor Technology in a Glance

The continuous advances in CMOS technology for processors and DRAMs have made CMOS sensor arrays a viable alternative to the popular charge-coupled devices (CCD) sensor technology. Standard CMOS mixed-signal technology allows the manufacture of monolithically integrated imaging devices: all the functions for timing, exposure control and ADC can be implemented on one piece of silicon, enabling the production of the so-called “camera-on-a-chip” [33]. Figure 2 is a diagram of a typical digital camera system, showing the difference between the building blocks of commonly used CCD cameras and the CMOS camera-on-a-chip [34]. The traditional imaging pipeline functions—such as color processing, image enhancement and image compression—can also be integrated into the camera. This enables quick processing and exchanging of images. The unique features of CMOS digital cameras allow many new applications, including network teleconferencing, videophones, guidance and navigation, automotive imaging systems, robotic and machine vision and of course, security and bio-medical image systems.

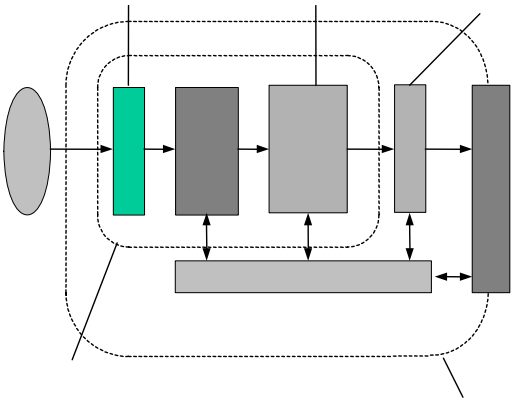


Figure 2. Block diagram of a typical digital camera system.

Most digital cameras still use CCDs to implement the image sensor. State-of-the-art CCD imagers are based on a mature technology and present excellent performance and image quality. They are still unsurpassed for high sensitivity and long exposure time, thanks to extremely low noise, high quantum efficiency and very high fill factors. Unfortunately, CCDs need specialized clock drivers that must provide clocking signals with relatively

large amplitudes (up to 10 V) and well-defined shapes. Multiple supply and bias voltages at non-standard values (up to 15 V) are often necessary, resulting in very complex systems.

Figure 3 is a block diagram of a widely used interline transfer CCD image sensor. In such sensors, incident photons are converted to charge, which is accumulated by the photodetectors during exposure time. In the subsequent readout time, the accumulated charge is sequentially transferred into the vertical and horizontal CCDs and then shifted to the chip-level output amplifier. However, the sequential readout of pixel charge limits the readout speed. Furthermore, CCDs are high-capacitance devices and during readout, all the capacitors are switched at the same time with high voltages; as a result, CCD image sensors usually consume a great deal of power. CCDs also cannot easily be integrated with CMOS circuits due to additional fabrication complexity and increased cost. Because it is very difficult to integrate all camera functions onto a single CCD chip, multiple chips must be used. A regular digital camera based on CCD image sensors is therefore burdened with high power consumption, large size and a relatively complex design; consequently, it is not well suited for portable imaging applications.

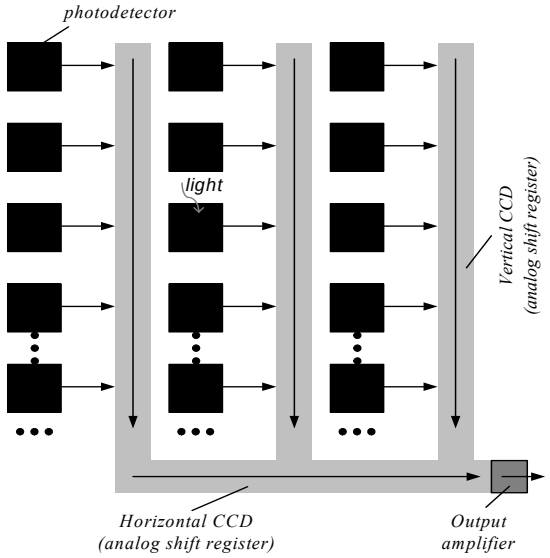


Figure 3. Block diagram of a typical interline transfer CCD image sensor.

Unlike CCD image sensors, CMOS imagers use digital memory style readout, using row decoders and column amplifiers. This readout overcomes many of the problems found with CCD image sensors: readout can be very fast, it can consume very little power, and random access of pixel values is possible so that selective readout of windows of interest is allowed. The power consumption of the overall system can be reduced because many of the supporting external electronic components required by a CCD sensor can be fabricated directly inside a CMOS sensor. Low power consumption helps to reduce the temperature (or the temperature gradient) of both the sensor and the camera head, leading to improved performance.

An additional advantage of CMOS imagers is that analog signal and digital processing can be integrated onto the same substrate, allowing fabrication of so called "smart" image sensors. Many "smart" image sensors have already been demonstrated in the literature. They performed functions of real time object tracking [35]-[42], motion detection [43]-[44], image compression [45]-[46], widening the dynamic range of the sensor [47]-[51] and others. These functions are usually performed by digital or nonlinear analog circuits and can be implemented inside the pixels and in the periphery of the array. Offloading signal processing functions makes more memory and DSP processing time available for higher-level tasks, such as image segmentation or tasks unrelated to imaging.

CMOS pixels can be divided into two main groups, passive pixel sensors (PPS and active pixel sensors (APS). Each individual pixel of a PPS array has only a photosensing element (usually a photodiode) and a switching MOSFET transistor. The signal is detected either by an output amplifier implemented in each column or by a single output for the entire imaging device. These conventional MOS-array sensors operate like an analog DRAM, offering the advantage of random access to the individual pixels. They suffer from relatively poor noise performance and reduced sensitivity compared to state-of-the-art CCD sensors. APS arrays are relatively novel image sensors that have amplifiers implemented in every pixel; this significantly improves the noise parameter.

Figure 4 shows the general architecture of the "smart" CMOS APS based image sensor. The core of this architecture is a camera-on-a-chip, consisting of a pixel array, a Y-addressing circuitry with a row driver, an X-addressing circuitry with a column driver, an analog front end (AFE), an analog-to-digital converter (ADC), a digital timing and control block, a bandgap reference and a clock generator. Optional analog and digital processing blocks "upgrade" the camera-on-a-chip core to a "smart" imager, and they are used to perform additional functions, that can vary from design to design, depending on the application and system requirements.

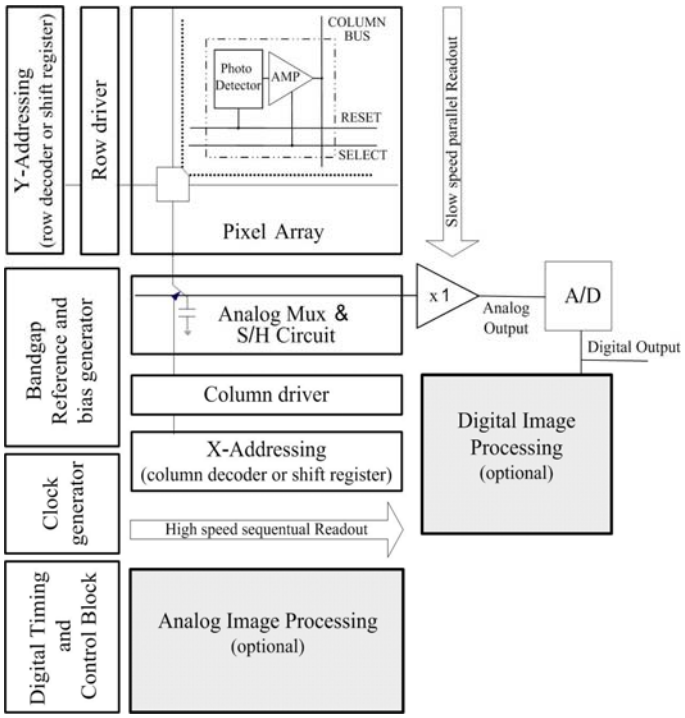


Figure 4. General architecture of the "smart" CMOS APS based image sensor.

The basic imager operation, depends on the chosen photodetector and pixel types, readout mode, Y-addressing and X-addressing circuitries, ADC type and of course, the analog and/or digital image processing. A brief description of the main imager building blocks is presented herein.

A. *APS Pixel Array* – the imager pixel array consists of N by M active pixels, while the most popular is the basic photodiode APS pixel, employing a photodiode and a readout circuit of three transistors. Generally, many types of photodetectors and pixels can be found in the literature. This includes a p-i-n photodiode, photogate and pinned photodiode based pixels, operating either in rolling shutter or in global shutter (snapshot) readout modes. The difference between these modes is that in the rolling shutter approach, the start and end of the light collection for each row is slightly delayed from the previous row, leading to image distortion when there is relative motion

between the imager and the scene. On the other hand, the global shutter technique uses a memory element inside each pixel and provides capabilities similar to a mechanical shutter: it allows simultaneous integration of the entire pixel array and then stops the exposure while the image data is read out. A detailed description of both rolling shutter and global shutter pixels can be found in [34].

Note, although most of today's "cameras-on-a-chip utilize" very simple pixels, many "smart" imagers employ more complicated pixels. Some of them perform analog image processing tasks at the pixel level. Very good examples for these imagers are neuromorphic sensors, where each pixel consists of a photo detector and local circuitry, performing spatio-temporal computations on the analog brightness signal. Another example is an imager, where the A/D conversion is performed in the pixel level.

B. Scanning Circuitry – Unlike CCD image sensors, CMOS imagers use digital memory style readout, usually employing Y-Addressing and X-Addressing to control the readout of output signals through the analog amplifiers and allow access to the required pixel. The array of pixels is accessed in a row-wise fashion using the Y-Addressing circuitry. All pixels in the row are read out into column analog readout circuits in parallel and then are sequentially read out using the X-Addressing circuitry (see Figure 4).

C. Analog Front End (AFE) - all pixels in a selected row are processed simultaneously and sampled onto sample-and-hold (S/H) circuits at the bottom of their respective rows. Due to this column parallel process, for an array having M columns, the AFE circuitry usually consists of 2*M S/H circuits, M size analog multiplexer, controlled by the X-Addressing circuitry, and one or M amplifiers to perform correlated double sampler (CDS). The CDS improves the signal-to-noise ratio (SNR) by eliminating the fixed pattern noise (FPN). A programmable- (or variable-) gain amplifier (PGA or VGA) follows the CDS to amplify the signal and better utilize the full dynamic range of the A/D converter (ADC).The number of amplifiers, required to perform the CDS functionality depends on the chosen CDS architecture and is equal to 2*N in case the subtraction is done separately for each column. The choice of an AFE configuration depends on many factors, including: the type of sensor being used, dynamic range, resolution, speed, noise, and power requirements. The considerations regarding making appropriate AFE choices for imaging applications can be found in [52].

D. Analog-to-digital conversion (ADC) – ADC is the inherent part of state-of-the-art "smart" image sensors. There are three general approaches to implementing sensor array ADC:

1. Pixel-level ADC, where every pixel has its own converter [53]-[54]. This approach allows parallel operation of all ADCs in the APS array, so a very low speed ADC is suitable. Using one ADC per pixel has additional advantages, such as higher SNR and simpler design.

2. Column-level ADC, where an array of ADCs is placed at the bottom of the APS array and each ADC is dedicated to one or more columns of the APS array [55]-[56]. All these ADCs are operated in parallel, so a low-to-medium-speed ADC design can be used, depending on the sensor array size. The disadvantages of this approach are the necessity of fitting each ADC within the pixel pitch (i.e., the column width) and the possible problems of mismatch among the converters at different columns.

3. Chip-level ADC, where a single ADC circuit serves the whole APS array [57]-[58]. This method requires a very high-speed ADC, especially if a very large array is used. The architecture shown in Figure 4, utilizes this approach for ADC implementation.

E. Bandgap reference and current generators – these building blocks are used to produce on-chip analog voltage and current references for other building blocks like amplifiers, ADC, digital clock generator and others. It is very important to design high precision and temperature independent references, especially in high resolution state-of-the-art image sensors, where the temperature of the die can vary by many tens of degrees.

F. Digital timing and control block, clock generator - aim to control the whole system operation. Their implementation in the chip level decreases the number of required I/O pads and thus reduces system power dissipation. Synchronized by the generated clock, the digital timing and control block produces the proper

sequencing of the row address, column address, ADC timing and the synchronization pulses creation for the pixel data going offchip. In addition, it controls the synchronization between the imager and the analog and digital processing.

G. Analog and Digital Image Processing – although these blocks are optional, they play a very important role in today's "smart" image sensors. Conventional vision systems are put at a disadvantage by the separation between a camera for “seeing” the world, and a computer or DSP for “figuring out” what is seen. In these systems all information from the camera is transferred to the computer for further processing. The amount of processing circuitry and wiring necessary to process this information completely in parallel is prohibitive. In all engineered systems, such computational resources are rarely available and are costly in terms of power, space, and reliability. Opposite to a conventional camera-on-a-chip, which only captures the image and transfer it for the further processing, "smart" image sensors reduce the computational cost of the processing stages interfaced to it by carrying out an extensive amount of computation at the focal plane itself (analog and digital image processing blocks in Figure 4), and transmitting only the result of this computation (see Figure 5).

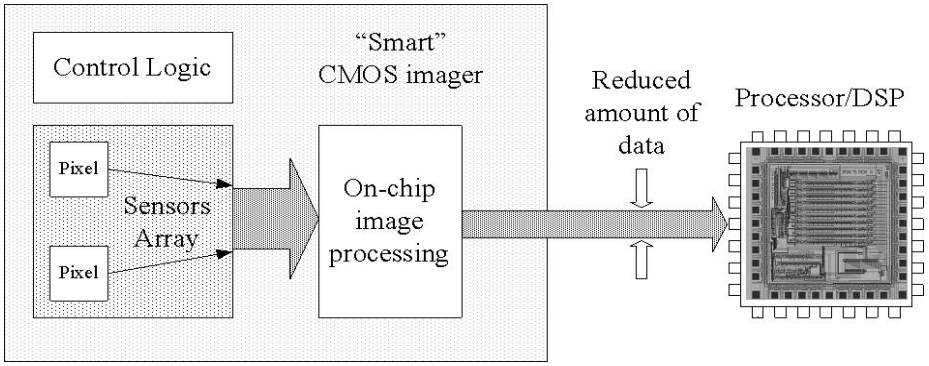


Figure 5. An example of an imaging system, employing a "smart" CMOS image sensor with on-chip processing and processors/DSPs for image processing

Both analog and digital processing can be performed either in the pixel or in the array periphery. There are advantages and disadvantages for both methods. In-pixel digital image processing is very rare because it requires pixel-level ADC implementation and results in very poor fill factor and large pixel size. In-pixel analog image processing is very popular, especially in the field of neuromorphic vision chips. In these chips in-pixel computations are fully parallel and distributed, since the information is processed according to the locally sensed signals and data from pixel neighbours. Usually, neuromorphic visual sensors have very low-power dissipations due to their operation in the subthreshold region, but suffer from low resolution, small fill-factor and very low image quality. Other applications employing in-pixel analog processing are tracking chips, wide dynamic range sensors, motion and edge detection chips, compression chips and others. The periphery analog processing approach assumes that analog processing is performed in the array periphery without penalty on the imager spatial resolution and it is usually done in a column parallel manner. While this approach has computational limitations compared to in-pixel analog processing, it allows better image quality. Periphery digital processing is the most standard and usually simpler. It is performed following the A/D conversion, utilizes standard existing techniques for digital processing and is usually done on the chip level. The main disadvantage of this approach is its inefficiency by means of area occupied and power dissipation. Note, all mentioned techniques can be mixed and applied together on one chip to achieve better results.

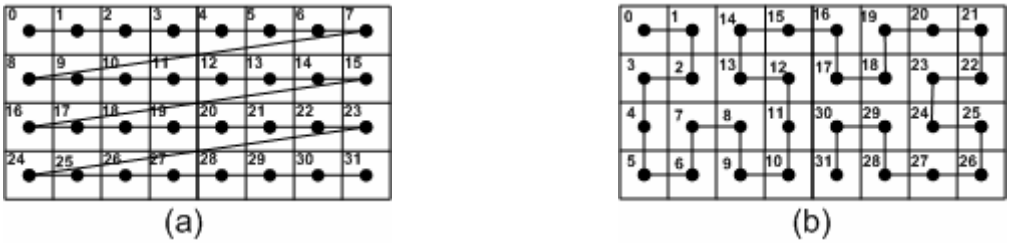
3. Image Sensors in Security Applications

The importance of security applications has significantly increased due to numerous terrorists' attacks worldwide. This area also greatly benefits from the achievements in the image sensors field. Today we can meet the cameras not only in military applications, but also in commercial and civilian applications. They are present in the shops and on the streets, in the vehicles and on the robots. The applications are numerous and can not be covered in this short paper. We have decided to concentrate on two important applications that represent a large fraction of the total security market. These applications are surveillance and biometrics. Both of the applications are extensively utilized in military, commercial and civilian fields.

3.1 Surveillance

Surveillance systems enable a human operator [59] to remotely monitor activity over large areas. Such systems are usually equipped with a number of video cameras, communication devices and computer software or some kind of DSP for real-time video analysis. Such analysis can include scene understanding, attention based alarming, colour analysis, tracking, motion detection, windows of interest extraction etc. With recent progress in CMOS image sensor technology and embedded processing, some of the mentioned functions and many others can be implemented in dedicated hardware, minimizing system cost and power consumption. Of course, such integration affects system configurability, but not all applications require configurable systems: some of them benefit from low cost and low power dedicated hardware solutions.

For example, in [60] we have presented an image sensor that can be used for such applications. Due to a specific scanning approach this sensor can be used efficiently for motion detection, tracking, windowing and digital zoom. Figure 6 shows the standard approach for sensor data scan - raster and the alternative – Morton or Z scan.



(a) raster scan – conventional approach in image sensors,

(b) Morton (Z) scan – newly proposed and implemented

Figure 6. Two approaches for data scan

The Morton (Z) scan poses a very valuable feature, neighbour pixels that are concentrated in blocks appear at the output sequentially, one after other. With this scanning approach the image blocks can be easily extracted and processed with simple on-chip hardware. For example, for constructing video camera with $\times 4$ digital zoom, the blocks of 4×4 pixels need to be extracted and averaged. Similarly, cameras with digital zoom $\times 8$ and $\times 16$ can be easily constructed. Figure 7shows measurements from our test chip.

Figure 7. Morton scan chip test results

Another example is a wide dynamic range (WDR) imager. Dynamic range (DR) quantifies the ability of a sensor to image highlights and shadows. If we define the dynamic range of the sensor as 20log(S/N), where S is the maximal signal value and N is the sensor noise, the typical image sensors will have a very limited dynamic range, about 65-75 dB. Wide dynamic range imaging is very important in many surveillance systems. The dynamic range can be increased in two ways: the first one is noise reduction and thus enabling expansion of the dynamic range toward darker scenes; the second method is incident light saturation level expansion, thus improving the dynamic range toward brighter scenes.

Herein we present one of the possible solutions for dynamic range extension in CMOS Active Pixel Sensors (APS) [2]. As in a traditional CMOS APS, this imager is constructed of a two-dimensional pixel array, with random pixel access capability and row-by-row readout rolling shutter method. Each pixel contains an optical sensor to receive light, a reset input and an electrical output representing the illumination received. This imager implements a simple function for saturation detection, and is able to control the light exposure time on a pixel-by-pixel basis, resulting in no saturation. The pixel value can then be determined as a floating-point representation. To do so, the outputs of a selected row are read out through the column-parallel signal chain, and at certain points in time are also compared with an appropriate threshold value, as shown in Figure 8. If a pixel value exceeds the threshold, i.e. the pixel is expected to be saturated at the end of the exposure time; the reset is given at that time to that pixel. The binary information concerning the reset (i.e., if it is applied or not) is saved in a digital storage for later calculation of the scaling factor. Thus, we can represent the pixel output in the following floating-point format: $M \cdot 2^{EXP}$, where the mantissa (M) represents the digitized pixel value and the exponent (EXP) represents the scaling factor. This way a customized, linear, large increase in the dynamic range is achieved.

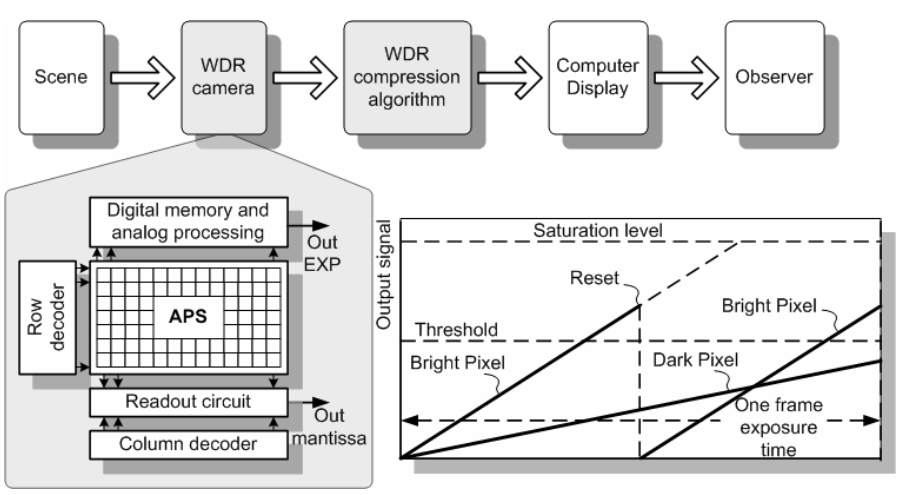


Figure 8. Imaging pipeline, image sensor architecture and work principle.

Figure 9 (a) and Figure 9 (b) show a comparison between an image captured by a traditional CMOS imager and by the autoexposure system described here. In Figure 9 (a), a scene is imaged with a strong light hitting the object; hence, some of the pixels are saturated. At the bottom of Figure 9 (b), the capability of the autoexposure sensor for imaging the details of the saturated area in real time may be observed. Since the display device is limited to eight bits, only the most relevant eight-bit part (i.e., the mantissa) of the thirteen-bit range of each pixel is displayed here. The exponent value, which is different for different areas, is not displayed.

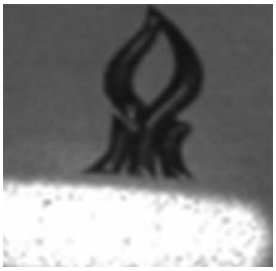


Figure 9. (a) Scene observed with a traditional CMOS APS sensor,

Figure 9. (b) Scene observed with our in-pixel autoexposure CMOS APS sensor.

3.2 Biometric personal identification

Biometric personal identification is strongly related to security and it refers to "identifying an individual based on his or her distinguishing physiological and/or behavioural characteristics (biometric identifiers)" [61]. Figure 10 shows the most frequently used biometric characteristics.

Figure 10. Biometric characteristics

Almost all biometric characteristics, shown in Figure 10, require some kind of sensing. Usually, conventional image sensors with external hardware or software image processing are used. The difficulty for on-chip integration is caused by the complexity of the required image processing algorithms. However, there are some developments that successfully achieve the required goals by parallel processing utilization.

To give some more detailed examples in the field, we concentrate on fingerprint sensors. Generally these sensors can be classified by the physical phenomena used for sensing: optical, capacitance, pressure and temperature. The first two classes are the most popular and both mainly employ CMOS technology.

In Figure 11 various technologies for fingerprint sensing are shown [62]. The most popular approach (see Figure 11 (a)) is based on optical sensing and light reflection from the finger surface. Also, this type provides high robustness to finger condition (dry or wet), but the system itself is tend to be bulky and costly. Alternative solutions that can provide compact and lower cost solutions, are based mostly on solid state sensors where the finger is directly placed on the sensor. However, in these solutions the sensor size needs to be at least equal to the size of the finger part used for sensing. Two sensors of this type are shown in Figure 11 (b) and (c). The first one is based on light transmitted through the finger and then sensed by the image sensor, while the second one is the non-optical sensor that can be implemented either as pressure, capacitance or temperature sensor. The fingerprint sensor, known as a "sweep" sensor and shown in Figure 11 (d), can be implemented using either the optical or other previously mentioned techniques. A "sweep" sensor employs only a few rows of pixels, thus in order to get a complete fingerprint stamp the finger needs to be moved over the sensing part. Such technology greatly reduces the cost of the sensor due to reduced sensor area and solves the problem of fingerprint stamp that needs to be left on the surface in the first two methods.

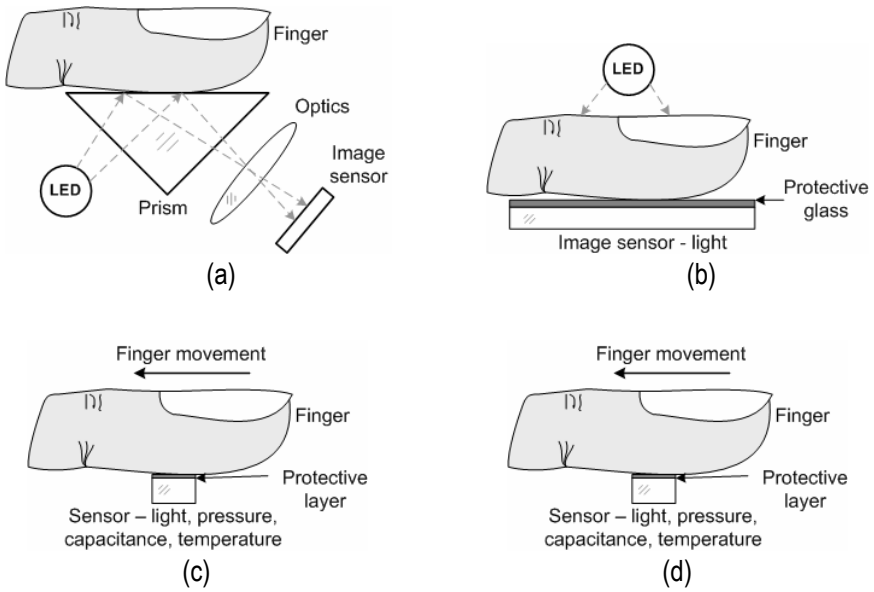


Figure 11. Fingerprint sensors
(a) optical - reflection based, (b) optical – transmission based
(c) non-optical – based on pressure, capacitance or temperature (d) “sweep” sensor

In all presented methods, the output signal is usually an image and the sensors are composed of pixels that sense either temperature, pressure, photons or change in capacitance. The overall architectures of these sensors are similar to the architecture described in section II and they integrate various image and signal processing algorithms, implemented the same die. Various research papers have been published in this area and numerous companies are working on such integration. For example, in [63] the authors implement image enhancement and robust sensing for various finger conditions. Capacitive sensing CMOS technology is used and data is processed in a column parallel way. The same technology is used also in [65], but the fingerprint identifier is also integrated and the data is processed massively in parallel for all pixels.

Despite the fact that fingerprint technology is quite mature, there is much work to be done to reduce power consumption, to improve technology and image processing algorithms and to achieve better system miniaturization.

4. Image Sensors in Medical Applications

Almost all medical and near medical areas benefit from image sensors utilization. These sensors are used for patients’ observation and drug production, inside the dentists offices and during surgeries. In most cases the sensor itself represents only a small fraction (in size and cost) of the larger system, but its functionality plays a major role in the whole system. Figure 12 shows examples of medical applications where CMOS image sensors are used. In this section of the paper we mostly concentrate on applications that push current image sensor technology to the edge of the possibilities. These applications are wireless capsule endoscopy and retinal implants. Both of these applications will play an important role in millions of patients’ lives in the near future.

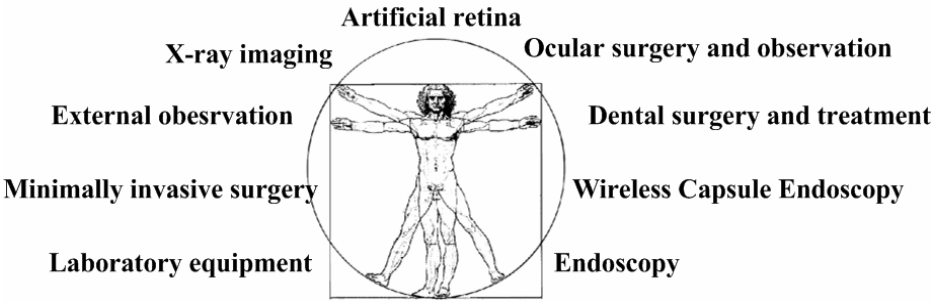


Figure 12. Image sensors applications in medicine

4.1 Wireless Capsule Endoscopy

Conventional medical instrumentation for gastrointestinal tract observation and surgery uses an endoscope that is externally penetrated. These systems are well developed and provide a good solution for inter-body observation and surgery. However, the small intestine (bowel) was almost not reachable using this conventional equipment, leaving it for observation only through surgery or through an inconvenient and sometimes painful push endoscopy procedures. Few years ago the sphere was revolutionized by the invention of the wireless image sensor capsule, which after swallowing, constantly transmits a video signal during its travel inside the body [32]. The capsule movement is insured by the natural peristalsis. According to Gavriel Iddan [32], the founder of Given Imaging™ [66] that commercializes this technology, "The design of the video capsule was made possible by progress in the performance of three technologies: complementary metal oxide silicon (CMOS) image sensors, application-specific integrated circuit (ASIC) devices, and white-light emitting diode (LED) illumination".

The general architecture of the capsule is shown in the Figure 13. It consists of LEDs, optics, camera, digital system processing, transmitter or transceiver and a power source. The dashed blocks represent additional future requirements for such capsules.

All capsule electronic components are required to be low power consumers to enable constant video transmission for a prolonged time (for about 6-8 hours) and/or high capacity batteries. An alternative solution to in-capsule batteries [67] is to use an external wireless power source that supplies energy to the capsule through electromagnetic coils. Such a solution enables to relax power requirements for the capsule electronics. This solution also provides an advantage in freeing space inside the capsule for other useful functions such as biopsy or medication. Also, the capsule position can be controlled externally through a strong magnetic field. But the required strong magnetic field can limit the capsule usage in spite of position control advantages [64].

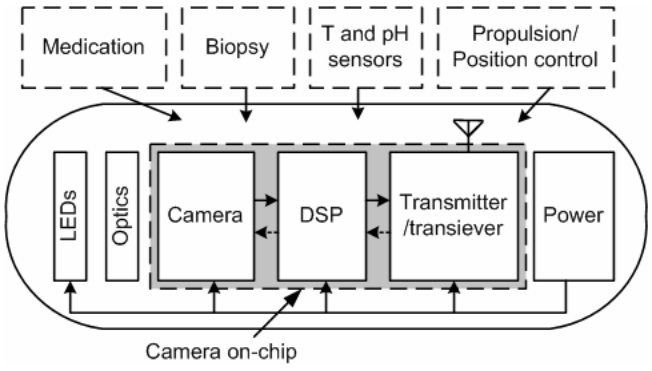


Figure 13. The swalable capsule architecture.
In the dashed boxes additional functionality that will be required in the future is shown

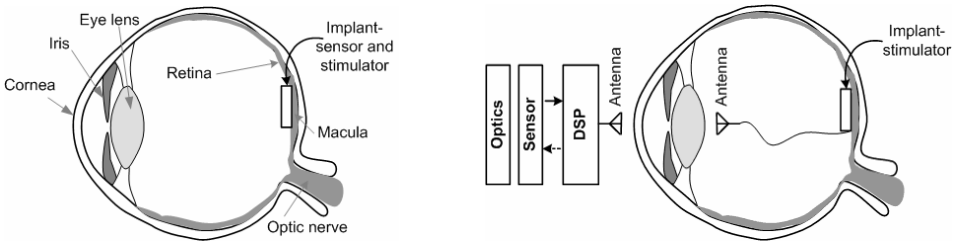
Currently the Given Imaging™ capsule developers have reached very encouraging results enabling two capsules: one intended for the Esophagus part (the upper part) of the gastrointestinal tract and the second for small

intestine observation. The first kind of the capsule is equipped with two CMOS image sensors and can transmit the video signal for about 20 minutes with 14 frames per second for each camera. The second one consists of only one CMOS image sensor and can transmit two frames per second for about eight hours. The company is developing now a new capsule generation that can transmit four frames per second.

Despite these encouraging results, a lot of work should be done to allow further miniaturization, image processing and compression algorithms integration, power reduction by various means (system integration, technology scaling etc.), frame-rate increase, quality improvement and usage of alternative power sources with larger capacity. The ultimate goal that needs to be achieved is full video frame-rate transmission for about 7-8 hours. To achieve these goals, a number of additional research groups work worldwide on wireless capsules development: eStool by Calgary university in Canada [68], MiRO by Intelligent Microsystems Center in Korea [69], EndoPill by Olympus [70].

4.2 Artificial Retina

Artificial vision is another example of CMOS image sensors implementation in medical applications. Today millions of people are suffering from full or partial blindness that was caused by various retinal deceases. In the early eighties it was shown that electrical stimulation of the retinal nerves can simulate visual sensation even in the patients with fully degraded receptors. Recently, researchers in a number of research institutes have developed miniature devices that can be implanted into the eye and stimulate the remaining retinal neural cells, returning partial vision ability for the blind patients. Such implants are called artificial retinas. Usually they are implanted in the macula area that normally is densely populated by the receptors and enables high-resolution vision. This break-through was enabled by the progress in electronics, surgical instrumentation, and biocompatible materials. Currently there are two major approaches for artificial retina development (see Figure 14).



(a) (b)

Figure 14. Artificial retinas (a) implantable sensor (b) external sensor

The first and the most promising one is the integration of sensing and stimulation elements in the same device and the second is separation of sensing and stimulation. In the first approach, an artificial retina device is an autonomous circuitry that does not require external control and the optics that is used for sensing is the natural optics of the eye composed of the cornea and lens. In the second, all the sensing and processing is performed outside of the eye and only stimulating elements are implanted during surgery. The data transfer from the sensing part to the stimulation part is performed through an RF link or through a tiny cable.

Actually there are two groups that have shown very promising results and are now performing clinical trials and commercialization through companies named Optobionics™ [71] and Second Sight [72]. Both groups already have a number of patients with such implants.

The device developed by Optobionics™ group does not require any power source, integrates about 5000 sensing (microphotodiodes) and stimulation (electrodes) elements, features two millimetres in diameter and is implanted under retina. The basic artificial silicon retina unit is shown in Figure 15 [73]. It is composed of a stimulating electrode and three PIN photodiodes connected in series to increase the output voltage.

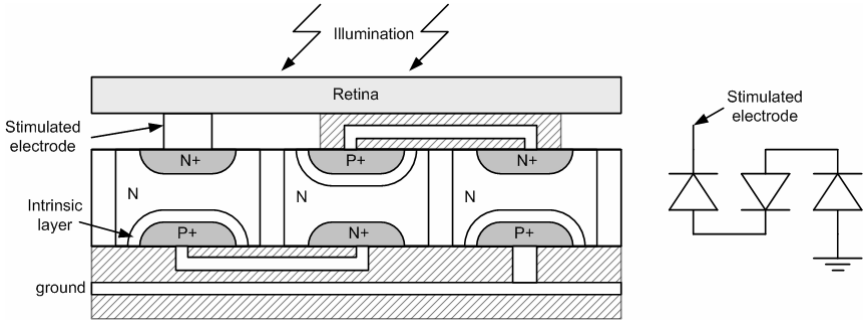


Figure 15. Artificial silicon retina – basic unit

The second group decided to follow the second approach and separate sensing from stimulation. The camera with the processor is situated on the patient glasses and the signal is transmitted through a cable to the eye that has an implanted stimulator. Currently, the implant resolution is not so impressive compared to the first group and features only 16 electrodes, but the developers plan to increase the resolution in the future models to 60 and 1000 electrodes [74].

5. Conclusions

In this paper we have presented a brief review of CMOS image sensors utilization in security and medical applications. In these applications image sensors play a major role and usually define the edge of the imaging technology. Despite the CMOS image sensor technology already exists for more than a decade, it is continuously developing and penetrating into new fields that were unreachable by its predecessor, CCD technology. Although many successes have been achieved during the last decade, a lot of work still needs to be done in this area. It requires extensive collaboration between various fields such as: electrical engineering, materials, computer science, medicine, psychology, chemistry etc. As to the electrical engineering and sensing fields, the work should be concentrated in the directions of power consumption reduction, functionality improvement and system integration. However, like in every multidisciplinary, electrical engineers, developing the electronic devices for medical purposes, are required to understand all above mentioned fields to successfully implement such devices.

References

[32] G. Iddan, G. Meron, A. Glukhovsky, P. Swain, “Wireless capsule endoscopy”, *Nature*, Vol. 405, p. 417, May 2000

[33] E. Fossum, "Low power camera-on-a-chip using CMOS Active Pixel Sensor technology", in *IEEE Symposium on Low Power Electronics*, pp. 74–77, 1995.

[34] O. Yadid-Pecht and R. Etienne-Cummings, "CMOS imagers: from phototransduction to image processing", *Kluwer Academic Publishers*, 2004.

[35] A. Fish, D. Turchin, O. Yadid-Pecht, " An APS with 2-Dimensional winner-take-all selection employing adaptive spatial filtering and false alarm reduction", *IEEE Trans. on Electron Devices*, Special Issue on Image Sensors, January, 2003.

[36] V. Brajovic and T. Kanade, “Computational sensor for visual tracking with attention”, *IEEE journal of solid-state circuits*, Vol.33, No.8, August 1998.

[37] T. Horiuchi and E. Niebur, “Conjunction search using a 1-D, analog VLSI-based attentional search/tracking chip,” *Conference for Advanced Research in VLSI, D. Scott Wills and Stephen P. DeWeerth*, Eds., pp. 276–290. IEEE Computer Society, 1999.

[38] G. Indiveri, “Neuromorphic analog VLSI sensor for visual tracking: Circuits and application examples.” *IEEE Trans. On Circuits and Systems*, II 46(11), pp. 1337–1347, November 1999.

[39] C. S. Wilson, T. G. Morris, and P. DeWeerth, “A two-dimensional, object-based analog VLSI visual attention system", *Twentieth Anniversary Conference on Advanced Research in VLSI, IEEE Computer Society Press*: Los Alamitos, CA. Vol. 20. pp. 291-308. March 1999.

[40] M.Clapp and R..Etienne-Cummings, “A dual pixel-type imager for imaging and motion centroid localozation”, *Proc. ISCAS’01*, Sydney, Australia, May 2001

[41] N. Mei Yu, T. Shibata and T. Ohmi, "A Real-Time Center-of-Mass Tracker Circuit Implemented by Neuron MOS Technology", *IEEE transactions on circuits and systems*—II, vol. 45, no. 4, April 1998.

[42] R.C. Meitzler, K. Strohbehn and A.G. Andreou, "A silicon retina for 2-D position and motion computation", *Proc. ISCAS'95*, New York, USA, 1995.

[43] A. Simoni, G. Torelli, F. Maloberti, A. Sartori, S. E. Plevridis and A. N. Birbas, "A Single-Chip Optical Sensor with Analog Memory for Motion Detection", *IEEE Journal of Solid-State Circuits*, Vol. 30, No. 7, July 1995.

[44] M. Clapp and R. Etienne-Cummings, "Dual Pixel Array for Imaging, Motion Detection and Centroid Tracking," *IEEE Sensors Journa*l, Vol, 2, No. 6, pp. 529-548, December 2002.

[45] S. Kawahito, M. Yoshida, M. Sasaki, K. Umehara, D. Miyazaki, Y. Tadokoro, K. Murata, S. Doushou, and A. Matsuzawa, "A CMOS Image Sensor with Analog Two-Dimensional DCT-Based Compression Circuits for One-Chip Cameras", *IEEE Journal of Solid-State Circuits*, Vol. 32, No. 12, 1997.

[46] K. Aizawa, H. Ohno, Y. Egi, T.Hamamoto, M. Hatory, H. Maruyama and J. Yamazaki "On sensor Image Compression," *IEEE Transaction On Circuits And Systems For Video Technology*, vol 7, no. 3, pp. 543-548, June 1997.

[47] O. Yadid-Pecht, A. Belenky " In-Pixel Autoexposure CMOS APS " *IEEE Journal of Solid-State Circuits*, Vol. 38, No. 8, pp. 1425-1428, August 2003.

[48] A. Fish, A. Belenky and O. Yadid-Pecht, "Wide Dynamic Range Snapshot APS for Ultra Low-Power Applications, *IEEE Transactions on Circuits and Systems II*, vol. 52, no. 11, pp. 729-733, November, 2005.

[49] O. Yadid-Pecht, C. Clark, B. Pain, C. Staller, E. Fossum , Wide dynamic range APS star tracker, in *Proc. SPIE/IS&T Sym. on Electronic Imaging: Science and Technology*, San Jose, California, SPIE Vol. 2654, Jan 29-Feb3, 1996, pp. 82-92.

[50] O. Yadid-Pecht, E. Fossum, "Wide Intrascene Dynamic Range CMOS APS Using Dual Sampling" , *IEEE Trans. Elec. Dev., special issue on solid state image sensors*, Vol. 44, No. 10, pp. 1721-1724, October 1997.

[51] O. Yadid-Pecht, "Wide dynamic range sensors", *Optical Engineering*, Vol. 38, No. 10, pp.1650-1660, October 1999.

[52] K. Buckley, "Selecting an Analog Front-End for Imaging Applications", *Analog Dialogue*, vol. 34-6, pp. 1-5, 2000.

[53] D.X.D. Yang, A. El Gamal, B. Fowler and H. Tian, "A 640x512 CMOS Image Sensor with Ultra Wide Dynamic Range Floating Point Pixel Level ADC," IEEE ISSCC, WA 17.5, 1999.

[54] B. Pain, S. Mendis, R. Scober, R. Nixon, and E. Fossum, "Low-power low-noise analog circuits for on-focal-plane signal processing of infrared sensors," IEEE Workshop on Charge Coupled Devices and Advanced Image Sensors, June, 1995.

[55] A. Dickinson, S. Mendis, D. Inglis, K. Azadet, and E. Fossum, "CMOS Digital Camera with Parallel Analog-to Digital Conversion Architecture," IEEE Workshop on Charge Coupled Devices and Advanced Image Sensors, April, 1995.

[56] A. Krymski and N. Tu, "A 9-V/Lux-s 5000-Frames/s 512x512 CMOS Sensor," IEEE Trans. Electron Devices, vol. 50 pp. 136-143, Jan. 2003.

[57] S. Smith, J.Hurwitz, M. Torrie, D. Baxter, A. Holmes, M.Panaghiston, R. Henderson, A. Murray, S. Anderson, and P. Denyer, "A single-chip 306x244-pixel CMOS NTSC video camera," ISSCC Digest of Technical Papers, pp. 170-171, February 1998.

[58] M. Loinaz, K. Singh, A. Blanksby, D. Inglis, K. Azadet, and B. Acland," A 200mW 3.3V CMOS color camera IC producing 352x288 24b Video at 30frames/s," ISSCC Digest of Technical Papers, pp 186-169, February 1998.

[59] G. L. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, and T. Ellis, "Active video-based surveillance system", *IEEE Signal Processing Magazine*, pp. 25-37, March 2005.

[60] E. Artyomov, Y. Rivenson, G. Levi, Orly Yadid-Pecht, "Morton (Z) Scan Based Real-Time Variable Resolution CMOS Image Sensor", *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 15, Issue 7, July 2005, pp. 947 – 952.

[61] A. Jain, L. Hong, S. Pankanti, "Biometric identification", *Communications of the ACM*, Vol. 43, No. 2, 2000.

[62] K. Uchida, "Fingerprint identification", *NEC Journal of Advanced Technology*, Vol. 2, No. 1, pp. 19-27, 2005.

[63] S.J. Kim, K.H. Lee, S.W. Han, and E. Yoon, "A 200×160 Pixel CMOS Fingerprint Recognition SoC with Adaptable Column-Parallel Processors,", *IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 250-251, February 2005.

[64] "Robotics road map", EURON Technology Roadmaps, April 23, 2004, www.org.id.tue.nl/IFIP-SG16/robotics-roadmap-2004.pdf

[65] S. Shigematsu, H. Morimura, Y. Tanabe, T. Adachi, and K. Machida, "A Single-Chip Fingerprint Sensor and Identifier", *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, Vol. 34, No. 12, pp. 1852-1859, December 1999

[66] www.givenimaging.com

[67] www.rfnorika.com
[68] K.N.C. Hin, O. Yadid-Pecht, M. Mintchev, "e-Stool: Self-Stabilizng Capsule for Colonic Imaging", *Neuro-stimulation Conf*, France, July 2005.
[69] microsystem.re.kr/main_eng/menu04/sub_menu01.asp
[70] "Olympus Launches High-resolution Capsule Endoscope in Europe", *Olympus press release*, October 13, 2005, www.olympus.co.jp/en/news/2005b/nr051013capsle.cfm
[71] www.optobionics.com
[72] www.2-sight.com
[73] U.S. Patent US-007003354-B2, given to Optobionics™
[74] S. Derra, "Bring sight to the blind", *R&D magazine*, 2005, www.rdmag.com

Authors' Information

Evgeny Artyomov – e-mail: artemov@bgu.ac.il

Alexander Fish – e-mail: afish@ee.bgu.ac.il

Orly Yadid-Pecht – e-mail: oyp@ee.bgu.ac.il

The VLSI Systems Center, Ben-Gurion University, Beer Sheva, Israel

MULTIMODAL MAN-MACHINE INTERFACE AND VIRTUAL REALITY FOR ASSISTIVE MEDICAL SYSTEMS

Adil Timofeev, Alexander Nechaev, Igor Gulenko, Vasily Andreev, Svetlana Chernakova, Mikhail Litvinov

Abstract: *The results of research the intelligence multimodal man-machine interface and virtual reality means for assistive medical systems including computers and mechatronic systems (robots) are discussed. The gesture translation for disability peoples, the learning-by-showing technology and virtual operating room with 3D visualization are presented in this report and were announced at International exhibition "Intelligent and Adaptive Robots–2005".*

Keywords: *multimodal man-machine interface, virtual reality, assistive medical systems.*

ACM Classification Keywords*: I.2. Artificial Intelligence*

1. Introduction

The modern medicine is actually required the development of new information technologies and the multimodal man-machine interface (MMI) for control of medical robots and mechatronic systems, for automation of surgery with virtual reality means for creating telemedical diagnostic systems, etc.

The NATO-grant № PST.CLG 975579 "The man-machine interface for assistive systems in neurosurgery", executed in 1999-2000 by partners from the St.-Petersburg Institute for Informatics and Automation of Russian Academy of Science (SPIIRAS), the State University of Aerospace Instrumentation, University of Karlsruhe (Germany) and Harvard Medical School (USA) has been directed on development of the man-machine interface, adaptive robots and multiagent technologies intended for neurosurgery [1]. The results of researches the medical MMI, robotics and mechatronic systems, including the results submitted in this report, have been presented at the international exhibition and a symposium "Intelligent and Adaptive Robots - 2005" [2].

2. Video Capture of Motion for Translating from Sign Language on Natural Language and Back

The development of anthropomorphous robots and their animation are very important for solution of various problems. Innovative results in this area, based on models and means of virtual reality, may be applied at such human activities, as medicine, sports, learning, computer and cognitive graphics. For example, in medicine the motion analysis of person (patient) can be essentially for automatic diagnostics and treatment of orthopedic diseases.

The researches of human's or anthropomorphous robot's motion are very important in the computer graphics for animation of virtual actors with methods and means of video capture was offered in [3]. Currently these methods being adapted for solution the problem of translation from a natural language to a sign language and back and for development of MMI for disable peoples.

The communication for people with restrictions on hearing or speech is taking place in completely different way, than it is for people without such restrictions. Peoples with such restrictions cannot watch TV, listen to radio without additional means and communicate by phone, as it done usually by people. So it is necessary to develop the approach to realization of specialized interfaces, which could be used in any telecommunication devices. The interfaces being discussed include the input-output informative means and data links.

As for data links, it is widely used a highly developed and reliable Information Telecommunication (IT) technologies. The advanced input-output informative systems for disability peoples are the novel MMI technologies.

As for gesture exchange with MMI, it is offered the output of gesture translation to be carried out by means of animated "Avatar", the simplified 3D computer model of human. Avatar can reproduce gestures by two ways [3, 4]:

- Generating gestures from corresponding text expressions in natural language using script sequences from database;
- Reproducing animation of gestures commands transmitted directly under the data link.

The gesture input means also can work by two following ways:

- Text generating in natural language according to analysis of gestures reproduced by operator;
- Transforming gestures of operator to gesture command sequences of animation for avatar.

3. Man-machine Interface for Aassistive Medical Systems Based on Video Capture of Motions

The important task in the development of robotic systems is a design of the man-machine interface. The new approach in creation of such interface is based on technologies of video capture of robot's motion and a virtual reality means [5, 6].

The general structure of man-machine interface and robot control system designed by using means of video capture and virtual reality technology includes two subsystems, see Fig. 1.

Subsystem 1 (subsystem of interaction with operator) includes the following components:

- System of video capture of operator's motions,
- Virtual control devices,
- Visualization procedures.

Subsystem 2 (subsystem of interaction with robot) has a various structures depending on application. However it is necessary to include the following components in this subsystem:

- Means of processing and distribution of robot control signals,
- Means of supervision over a condition of robots.

It is also proposed to develop the following components increasing the control efficiency and reliability of MMI systems [3, 4]:

- Active multi-channel system of video motion captures;
- Means of visual observation.

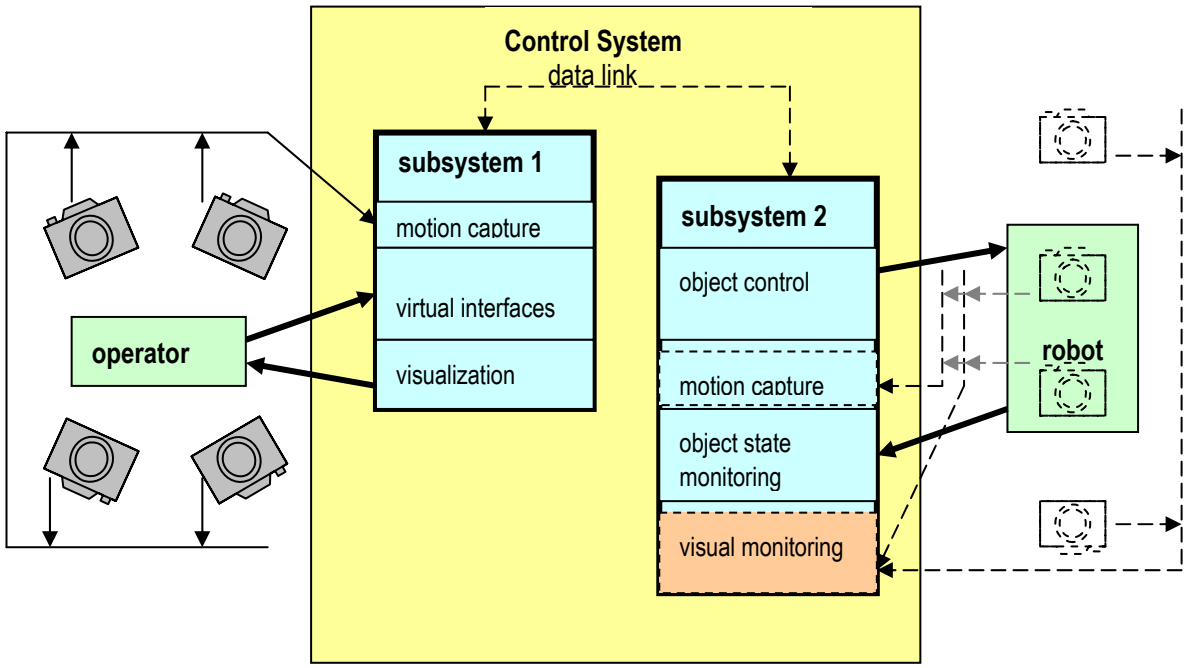


Fig. 1

4. The Learning Technology Using Multimodal Man-machine Interface

The intelligence learning technology with showing the natural movements (gestures) of operator [7, 8] is developed to creating natural and accessible means for people's communication with technical and information systems (fig. 2).

Fig.2

The offered learning technology for medical mechatronic systems (robots) is based on the intelligence MMI [9], which provide the following features for medical applications:

- Simple and correct understanding of gesture and speech commands of doctor or patient [10],
- The intuitive teaching the medical mechatronic system by means of natural operator's hand movements (without traditional movement's programming) [8],
- The visualization of real and virtual 3D medical images (X-ray, endoscopic, thermography, etc),
- The doctor's automatic workplace (AWP) with 3D viewing of X-ray images in real time mode [13];
- The novel stereo-projective systems for observation X-ray, endoscopic and other medical images;

- The human interaction with computer's models and real medical equipment by means tactile and force-torque sensors during diagnostic or planning of medical operation [6, 12],
- The information and telecommunication support of medical equipment control systems using the Virtual Operating Room (VOR).

The development of the Multimodal Man-machine Interface (MMI) for telemedicine and medical robotics assumes the creation of highly realistic effect of doctor's presence in remote environment of patient.

For this purpose virtual models and computer-synthesized 3D-images of virtual objects composed with images of real environment (Augmented Reality technologies, AR) are used. The basic problems of AR technology realization is a problem of exact registration of a computer- synthesized image of geometrical model with real image and a problem of 3D image visualization in a real time mode.

It is suggested to use the following means to 3D visualization of medical images (fig.3):

- Stereo glasses for observation on the computer monitor;
- Stereo-displays glasses for 3D viewing color images without computer monitor;
- Means of augmenting real medical images by virtual images (AR).

Fig. 3

Tactile-force interaction with virtual medical objects is necessary when it is not only required to observe the environment, but it is also necessary to perform any actions in it. These manipulations in remote environment will be more successful, if it will be possible to create realistic and adequate perception of objects in environment surrounding the patient and medical robot, to give an ability of feeling a virtual object with mass, shape, elastic and friction features as a real object.

These technologies have partially been developed within the Partner Project 1992p with EOARD (European Office of Aerospace Research and Development, London, United Kingdom) and based on a long-term experience of medical system development [14].

The Head Tracking System (HTS) and Hand Tracking System (HTS+) prototypes have been developed for accurate measurement of human-operator's head and hands movements. The basic aim of this development is to create the simple (cost-effective), reliable and steady means for human interaction with telecontrol objects.

For medicine the development of so-called "haptic" interface, intended for tactile and force-torque displaying on hand of doctor (surgeon) is especially actual. The developed handle with force-torque feedback sensors reflects the real tactile and force interaction of mechatronic system (robots) instrument with real objects.

The prototype of handle with force-torque sensor for controlling and learning of medical mechatronic systems is shown on Fig. 4. As an example of medical application of novel force-torque sensor is a prototype of robot-masseur with exact control of force contact to patient's muscles and a high level safety of robot-masseur movements near the patient or personnel.

Fig. 4

The offered technologies of 3D visualization, virtual reality means and learning-by-showing technology are especially useful for navigation, programming of movements and control of medical mechatronic systems [4, 11]. Thus information technologies with creating a highly realistic effect of presence the doctor in remote patient's environment are useful for telemedicine application too.

Using of medical robots in surgery or telemedicine may be dangerous to patients. Therefore it is necessary to create virtual models of robot, patient and operating room for testing, planning and safety providing of medical operation with usage of mechatronic systems (medical robots).

5. Virtual Model of Operating Room

Virtual operating room (VOR) should be similar to real one as much as possible, because the doctor's work in usual conditions would be more effective and productive. Therefore the VOR must be "filled" with familiar medical tools, equipment and assistive robots in their computer representation.

In this case there is an opportunity of carrying out of trial educational operations not on the real patient body, but on his virtual model. Carrying out of operation on the virtual "electronic patient" can be considered as a planning stage of real surgical operation.

The second important problem being solved using VOR is an increasing professionalism of medical personnel. The dynamic virtual model of neurosurgical VOR is presented on fig. 5.

The information on real and virtual operation and clinical data of patient are united in a uniform picture of medical operation. It is possible to return to past operation for comparative analysis and finding-out the efficiency of different methods of medical operation. Due to this not only doctors, but also students and post-graduate students can to training with VOR due to improve their skill [3-6, 11].

Fig. 5

During the virtual operation the surgeon can to observe the trajectory of medical tool, for example, under brain tomogram of patient. In a case when the virtual medical tool to reach zone of high risk for virtual patient, the doctor can request the VOR program to find other variant of virtual tool targeting to necessary point of brain on a more safety trajectory or to take a conclusion on impossibility of carrying out of operation.

6. Conclusion

The offered information technologies, the multimodal man-machine interface and a virtual reality technique find more and more wide applications not only in medicine, but also in other areas (service, home assistance, telecommunications, space, etc.). Thus the intelligence man-machine interface, dynamic models of a virtual reality and multi-agent technologies will play especially important role [5, 6].

At the International exhibition «Intelligent and Adaptive Robots – 2005» the authors of this report have been awarded with 4 Gold Medals and Diplomas «for the development and introduction of innovative technology of man-machine interaction for telemedicine and medical robotics».

The presented work is executed at partial support of the grant RFBR № 05-01-08044_ofi and the grant № 17 "Development of Man-Machine Interface and Control Methods for Neurosurgical Robots on the Base of Virtual Reality Models" of Saint-Petersburg Scientific Center of Russian Academy of Sciences.

Bibliography

[1] C. Burghart, O. Schorr, S. Yigit, N. Hata, K. Chinzei, A. Timofeev, R. Kikinis, H. Wörn, U. Rembold A Multi-Agent-System Architecture for Man-Machine Interaction in Computer Aided Surgery. – Proceedings of the 16th IAR Annual Meeting (Strasburg, November 22-23, 2001), pp. 117-123.

[2] Intelligent and Adaptive Robots – 2005. Official Exhibition Catalogue – M.: Open Company "ZETP", 2005, p. 45 (In Russian).

[3] Gulenko I.E., Shugina V.S. Application of technologies of video capture of movement in medicine and sports. – Works of conference "New information technologies" (Sudak, Crimea, May, 22-29, 2005), pp. 318-320 (In Russian).

[4] Timofeev A.V., Gulenko I.E. and Litvinov M.V. Analysis, Processing and Transfer of Dynamic Images in Virtual Reality Models. – Pattern Recognition and Image Analysis, 2004, vol. 16, No.1, pp.97-99.

[5] Timofeev A.V., Andreev V.A., Gulenko I.E., Derin O.A., Litvinov M.V. Design and Implementation of Multi-Agent Man-Machine Interface on the Base of Virtual Reality Models. – Proceedings of 9-th International Conference "Speech and Computer" (20-22 September, 2004), Saint-Petersburg, Russia, pp. 670-676.

[6] Timofeev A.V. Intellectualization for Man-Machine Interface and Network Control in Multi-Agent Infotelecommunication Systems of New Generation. – Proceedings of 9-th International Conference "Speech and Computer" (20-22 September, 2004), Saint-Petersburg, Russia, pp. 694-700.

[7] Kulakov F.M., Nechaev A.I., Chernakova S.E. Modelling of Environment for the Teaching by Showing Process. In Proc. of SPIIRAS, Russia, St. Petersburg, 2002, Issue № 2, pp. 105–113.

[8] Chernakova S.E., Kulakov F.M., Nechaev A.I. Learning of the robot by method of showing with use «sensing» gloves. – Works of the First international conference on mechatronics and a robotics (Saint Petersburg, May, 29 – June, 2, 2000), pp. 155-164 (In Russian).

A. Karpov, A. Ronzhin, A. Nechaev, S. Chernakova. Assistive multimodal system based on speech recognition and head tracking. In Proceedings of 9-th International Conference SPECOM'2004, St. Petersburg, 2004, pp. 521-530.

[9] A. Karpov, A. Ronzhin, A. Nechaev, S. Chernakova. Multimodal system for hands-free PC control. In Proc. of 13-th European Signal Processing Conference EUSIPCO–2005, Antalya, Turkey, 2005.

[10] Litvinov M.V., Timofeev A.V., Popov A.B. Virtual model of the adaptive robot for neurosurgery. – Proceedings of the International conference "Adaptive robots and the general logic theory of systems - AR&GSLT, (Saint Petersburg, on May, 8-11 2004), pp. 119–122 (In Russian).

[11] Timofeev A.V., Shibzuhov Z.M., Sheojev A.M. Designing and learning multi-agent diagnostic systems. – Proceedings of the First international conference on mechatronics and a robotics "M&R.

[12] Nechaev A.I., Nazaruk V.P., Chernakova S.E. Method of registration and visualization of 3D X-ray images in real time for technical nondestructive checking and medical diagnostics. – "Information technologies", 11, 2005, Moscow, pp. 11-21, in Russian.

[13] Chernakova S.E., Kulakov F.M., Nechaev A.I. Advanced man-machine interface for telerobotic based on man-operator's motions and gaze direction tracking – Proceedings of Sixth ISTC Scientific Advisory Committee Seminar "Science and Computing", Moscow. 15–17 September 2003, Volume 1, part II, pp. 303-308.

Authors' Information

Timofeev Adil Vasilievich – Dr. Sc., Professor, Honoured Scientist of Russian Federation, Saint-Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 199178, Russia, Saint-Petersburg, 14-th Line, 39, phone: +7-812-328-0421; fax: +7-812-328-4450, e-mail: tav@iias.spb.su

Nechaev Alexander Ivanovich – Scientific Researcher, Saint-Petersburg Institute for Informatics and Automation of RAS, 199178, Russia, Saint-Petersburg, 14-th Line, 39, e-mail: nechaev@iias.spb.su

Gulenko Igor Evgenievich – Post Graduate Student, Saint-Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, 199178, Russia, Saint-Petersburg, 14-th Line, 39, e-mail: gig@yandex.ru

Andreev Vasily Alexandrovich – Post Graduate Student, Saint-Petersburg Institute for Informatics and Automation of RAS, 199178, Russia, Saint-Petersburg, 14-th Line, 39, e-mail: vasiliy-a@yandex.ru

Chernakova Svetlana Eduardovna – Minor Scientist, Saint-Petersburg Institute for Informatics and Automation of RAS, 199178, Russia, Saint-Petersburg, 14-th Line, 39, e-mail: chernakova@iias.spb.su

Litvinov Mikhail Vladimirovich – Post Graduate Student, Baltic State Technical University "Voenmech", 190005,Saint-Petersburg, Russia, 1-st Krasnoarmeyskaya, 1, e-mail: sid-4d@inbox.ru

MEDICAL DATA-ADVISORY WEB-RESOURCE "MED-HEALTH"

Anatoly I. Bykh, Elena V. Visotska, Tatjana V. Zhemchuzhkina, Andrey P. Porvan, Alexander V. Zhuk

Abstract: *In this article the medical data-advisory web-resource developed by authors is considered. This resource allows carrying out information interchange between consumers of medical services and the medical establishments giving these services, as well as firms-manufacturers of medical equipment and medicaments. Main sections of this web-site, their purposes and capabilities are considered in this article.*

Keywords: *web-resource, information interchange, medical establishment, medical equipment, medical advice.*

Introduction

Since electronic global networks appeared it is possible to watch the tendency of the Internet-technologies applications in different fields of human activity. However, application of such technologies in medicine remained aloof until recently. It is bound with following: not any information can be transferred to a network, and the transmitted information not always corresponds to the general requirements to the data transfer.

One of the guidelines in modern Internet-technology applications is development of medical data-advisory resources which allow getting a different medical information, as well as remote advice of the expert in certain field of medicine. We can refer to these progressive guidelines basic medical sites, different municipal information portals, specialized sites of the medical services, sites of firms-manufacturers of medical equipment or their trade representatives, etc.

Urgency of the Problem

Among known web-resources we can mark sites of private clinics "Dr. Alex", "Firmament", and specialized sites "Mednet", "Doctor-home", giving the information about departments of these medical establishments, medical personnel, the lists of services and medical equipment which is available there. However, these sites contain information concerning only certain medical establishment, and guest of these sites often can't get complete information about necessary question in one place, and this takes expense of time and means. Besides, many institutions save on information about other medical establishments and their services with aim to reduce costs for site making. That's why we can't say about working efficiency of these sites.

In that way, making of the medical data-advisory resource which takes into account all noted above lacks, gives interest and is actual problem.

Working up of the Medical Data-advisory Resource

If to consider medical web-resources (sites) as integrated data-advisory system it is necessary to distinguish key components of this system. It is offered to consider this system as integration of representatives of purposeful groups. Regarding these groups the medical institution has or can have communicative aims. Traditionally distinguish external and internal web-resource environments (fig. 1).

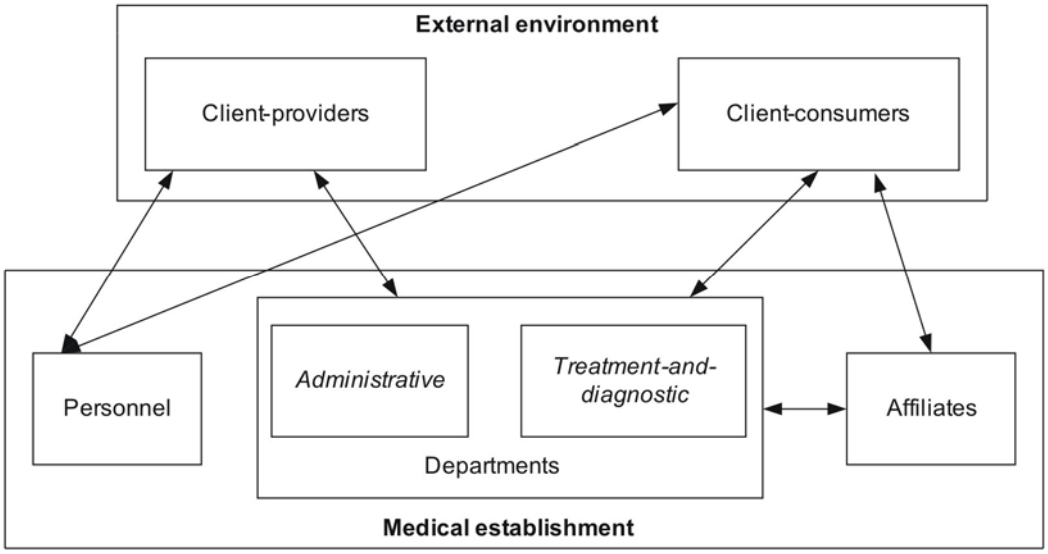


Figure 1 - External and internal web-resource environments

In case of medical data-advisory web-resource external environment contains:

- clients-consumers of services, proposed by medical institutions (population, physicians, pharmacologists, medical establishments of different types);
- clients-providers, giving the information for allocation in web-resource (medical establishments, drugstores, firms-manufacturers of medical equipment and pharmaceutical preparations, etc.).

The internal environment of medical data-advisory web-resource is essentially information resource, containing information of the following contents:

- the index of medical establishments; parts selected in structure of these establishments; existing medical cabinets; medical private offices;
- the personnel of medical establishment (personal data of treating and advising doctors, attendants, etc.);
- the medical equipment (diagnostic and therapeutic equipment, which is available there, capabilities of this equipment, the list of the procedures spent with this equipment; the advertising information of firms-manufacturers of the medical equipment);
- services given by medical establishments (consultations of doctors, diagnostic and therapeutic procedures, etc.).

It is possible to allocate three base functions of an offered Web-resource:

1. Information function - realization of information and advertising activity of the Web-resource (giving of the information about the medical institutions, given services, consultations, the diagnostic and therapeutic equipment, medical preparations, etc.),

2. Communicative function - function of information interchange between clients and a resource (this resource is the information intermediary between consumers of medical services and establishments, which these services give);
3. Service function - giving of advisory services both in on-line and off-line mode, giving of the information about capabilities of diagnostics and treatment in different medical institutions, function of electronic payments and an electronic drugstore.

The site consists of six basic sections: medical establishments, doctors, medical equipment, medical specialization, drugstores and contacts (fig. 2).

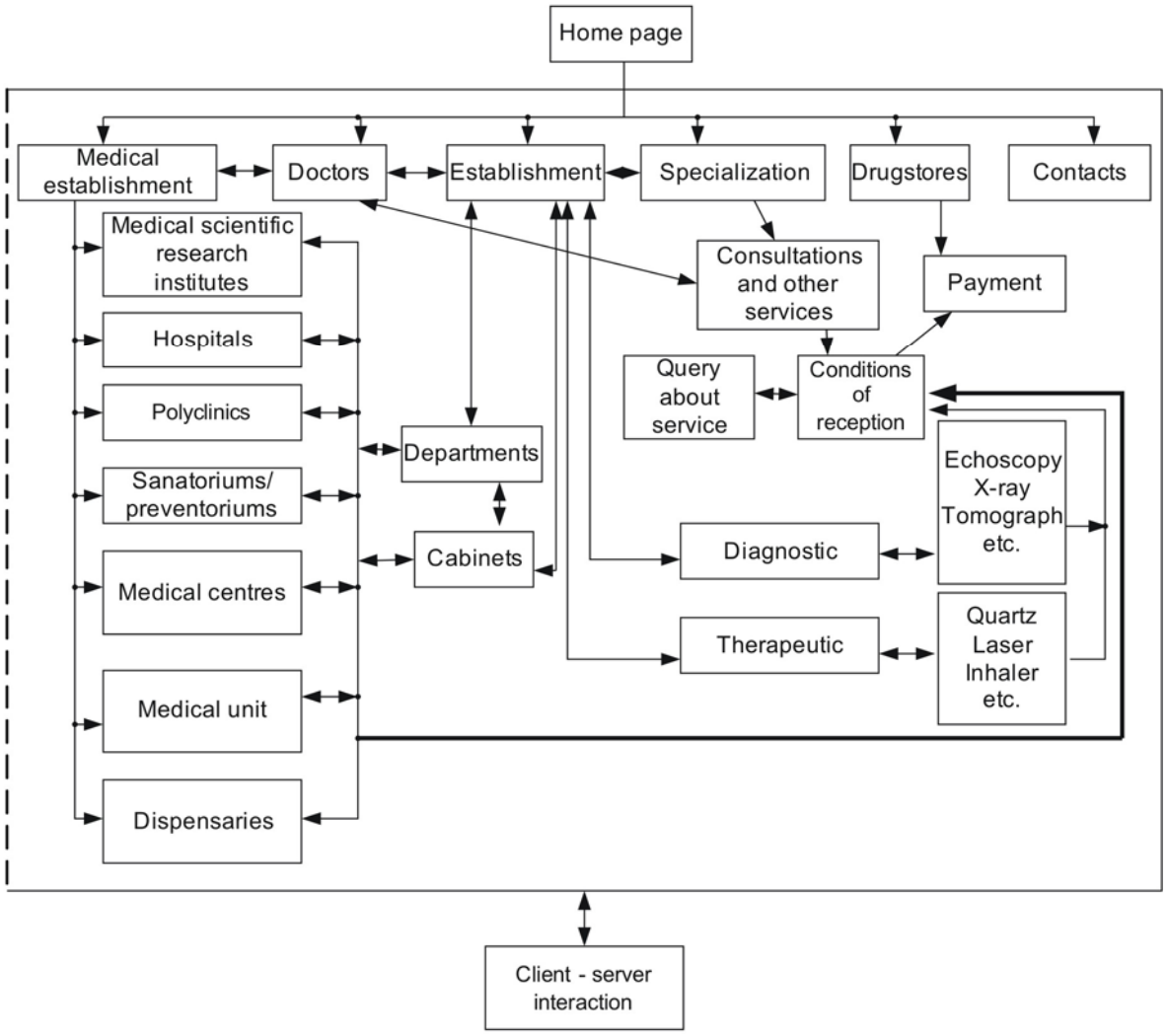


Figure 2 - The block diagram of the medical data-advisory web-resource “”

The section “Medical establishments” comprises the information about all existent medical institutions (medical scientific research institutes; hospitals; polyclinics; sanatoriums/preventoriums; the medical centers; medical units; dispensaries. This section contains the information about a type of the establishment, its personnel; the contact information; the other help information concerning the medical establishment.

The section “Doctors” contains the personal information about doctors, a place of job, their medical specializations, the experience of job, the achievements, used methods of treatment, the contact information. The opportunity to find out about consultations and other kinds of the services offered by the doctor, including services online is given in this section.

The section “Medical specialization” includes the basic directions of diagnostics and the therapies selected in activity of medical establishments and doctors (for example, cardiology, surgery, ophthalmology, etc.).

The section “Medical equipment” contains the information about the equipment contained in concrete medical establishments, with the description of its capabilities and services given with this equipment; the information of firms-manufacturers about the medical equipment produced by them.

The section “Drugstore” includes all assortments of the goods of medical purpose offered by drugstores, as the information about an address of the drugstore. The opportunity of orders of the goods through the Internet with home delivery is given.

The section “Contacts” contains the information for communication with administration of the site at occurrence of questions or offers.

The created Web-resource is designed as for those who cares of the health and interests in novelties in the field of medicine and for experts of a medical structure and other visitors. The user of a web-resource who is the consumer of the services given by medical institutions with the help of the offered web-site can receive if necessary the information about what medical institutions accessible to this user renders service required to him (service of diagnostic, therapeutic or advisory type), what conditions of rendering of this service, what type of the equipment is used there. Due to complete volume of the necessary information, the consumer chooses service optimally suitable for him on quality and availability. Medical institutions receive an opportunity to illuminate all spectrums of the services rendered by these establishments for a wide range of consumers of these services. Besides as on the same site firms-manufacturers of the medical equipment place advertising of their production, medical institutions get access to the information on the newest medical technologies, that allows to improve quality of medical services.

Conclusion

The offered medical information resource provides two-way contact between consumers of medical services and the establishments giving these services, and also manufacturers of the equipment and medical preparations. Complete resource of this kind is interest for various groups of users, first of all for those who wish to choose necessary and optimally accessible medical service among great number of the services given by different medical institutions. Besides this, resource enables medical institutions to give the information about them. Firms-manufacturers of the medical equipment and of the pharmacological preparations also receive an opportunity to acquaint the broad audience of consumers and experts with produced production that enables medical institutions to have the information on the newest medical technologies.

Authors’ Information

A. I. Bykh – Doctor of Physics and Mathematics, professor, Head of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

E. V. Visotska – PhD, associate professor of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

T. V. Zhemchuzhkina – PhD, senior lecturer of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

A. P. Porvan – engineer of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

A. V. Zhuk – student of Biomedical Electronics sub-faculty of Kharkov National University of Radio Electronics

Kharkov National University of Radio Electronics, Ukraine, 61166, Lenin Avenue, 14, Biomedical Electronics sub-faculty, e-mail: diagnost@kture.kharkov.ua

OPEN SOURCE INFORMATION TECHNOLOGIES APPROACH FOR MODDELING OF ANKLE-FOOT ORTHOSIS

Slavyana Milusheva, Stefan Karastanev, Yuli Toshev

Abstract: *Computer modeling is a perspective method for optimal design of prosthesis and orthoses. The study is oriented to develop modular ankle foot orthosis (MAFO) to assist the very frequently observed gait abnormalities relating the human ankle-foot complex using CAD modeling. The main goal is to assist the ankle-foot flexors and extensors during the gait cycle (stance and swing) using torsion spring.*

Utilizing 3D modeling and animating open source software (Blender 3D), it is possible to generate artificially different kind of normal and abnormal gaits and investigate and adjust the assistive modular spring driven ankle foot orthosis.

Keywords: *biomechanics; 3D computer modeling, ankle-foot orthosis*

ACM Classification Keywords*: I.3.7 Three-Dimensional Graphics and Realism, I.6.5 Model Development*

Introduction

Open source software refers to computer software available with its source code and under an open source license. Such a license permits anyone to study, change, and improve the software, and to distribute the unmodified or modified software. It is the most prominent example of open source development. This software gives an outstanding flexibility in terms of extensibility and modularity.

The study is based of 3D modeling technology provided by one of the most advanced open source software – **Blender.** Blender is a free 3D modeler program. It is used for modeling and rendering three-dimensional graphics and animations. Blender is available for several operating systems, including FreeBSD, IRIX, GNU/Linux, Microsoft Windows, Mac OS X, Solaris, SkyOS, and MorphOS. In addition, Blender's recent burst of new features in the last few versions has actually brought it close in feature set comparison to high-end 3D software such as 3D Studio Max and Maya. Among these features and user interface ideas are, for example, complex fluid and cloth effects, a comprehensive and well-thought out hotkey program, which rivals that of most higher end applications, and a wide range of easily accessible and creatable extensions using Python scripting. Regardless of lack of natively implemented CAD functionality there are a lot of possibilities for development of Python based helper scripts for precise engineer modeling.

Ankle-foot orthosis (AFO) is commonly used to help subjects with weakness of ankle dorsiflexor muscles due to peripheral or central nervous system disorders. Both these disorders are due to the weakness of the tibialis anterior muscle which results in lack of dorsiflexion assist moment. The deformity and muscle weakness of one joint in the lower extremity influences the stability of the adjacent joints, thereby requiring compensatory adaptation.

During level plane ambulation the ankle should be close to a neutral position (right angle) each time the foot strikes the floor. Insufficient dorsiflexion may be the result of hyperactive plantarflexors that produce very high plantarflexion moment at the ankle, or weakness of the dorsiflexion muscles. This affects the ability of the ankle to dorsiflex. As result the patient make a forefoot contact instead of normal “heel-strike”. If there is a weak push-off, the stride length reduces, and the gait velocity decreases. Similarly, during the gait swing phase, the ankle is dorsiflexed to allow the foot to clear the ground while the extremity is advanced. Hyperactive or weak dorsiflexors may result in insufficient dorsiflexion, which must be compensated by alterations in the gait patterns so that the toes do not drag. This insufficient dorsiflexion during the gait swing phase is termed as “foot-drop”. In addition to

the toes dragging, the foot may become abnormally supinated, which may result in an ankle sprain or fracture, when the weight is applied to the limb. Foot-drop is commonly observed in subjects after a stroke or personal nerve injury.

There are several possible treatments for foot-drop - medicinal, orthotic, or surgical. It is to note that the most common is the orthotic treatment. Orthotic devices are intended to support the ankle, to correct deformities, and to prevent further occurrences. A key goal of orthotic treatment is to assist the patient achieving normal gait patterns.

Different orthoses are used to enhance the ankle-foot position and mobility. The most common types are hingeless and hinge orthoses. Using springs, the hinge orthoses could assist ankle flexion/extension during gait, i.e. they are pseudo-active orthotic devices. The standard ankle foot orthoses (AFO) is a rigid polypropylene structure which prevents any ankle motion.

Methods

The study is oriented to develop modular ankle foot orthosis (MAFO) with two units (shank brace and foot brace) connected with lateral and medial adjustable hinged joints.

Gait analysis

Gait analysis is useful in objective documentation of walking ability as well as identifying the underlying causes for walking abnormalities in patients with cerebral palsy, stroke, head injury and other neuromuscular problems. The results of gait analysis are useful in determining the best course of treatment in these patients.

Normal gait

The gait cycle begins when one foot contacts the ground and ends when that foot contacts the ground again. Thus, each cycle begins at initial contact with a stance phase and proceeds through a swing phase until the cycle ends with the limb's next initial contact. Stance phase accounts for approximately 60 percent, and swing phase for approximately 40 percent, of a single gait cycle.

Each gait cycle includes two periods when both feet are on the ground. The first period of double limb support begins at initial contact, and lasts for the first 10 to 12 percent of the cycle. The second period of double limb support occurs in the final 10 to 12 percent of stance phase. As the stance limb prepares to leave the ground, the opposite limb contacts the ground and accepts the body's weight. The two periods of double limb support account for 20 to 24 percent of the gait cycle's total duration.

Stance phase of gait is divided into four periods: loading response, midstance, terminal stance, and preswing. Swing phase is divided into three periods: initial swing, midswing, and terminal swing. The beginning and and ending of each period are defined by specific events.

Each subphase is accompanied by a change in position, ground reaction force, and/or internal muscular forces. Gait cycle analysis in this sense is essentially a sagittal plane function.

The ankle is plantarflexed 10 degrees at heelstrike, with further plantorflexion dampened by the ankle dorsiflexors, aiding with shock absorption. At midstance, ground reaction tends to dorsiflex the ankle which is held rigid by the plantarflexors, controlling forward thrust of the tibia. Ground reaction continues to push the ankle toward dorsiflexion in terminal stance, resisted by the plantarflexors. The ankle passively dorsiflexes as it is unloaded in preswing.

Ankle joint motion (sagittal plane):

- between initial heel contact and foot flat ankle undergoes ~3-4° plantar flexion (first 6% of stride);
- after foot flat ankle dorsiflexes until a little beyond 40% of stride (as hip moves over ankle), reaching maximum of 8-10°;

- ankle plantar flexes for remainder of stance phase until after push-off (reaches maximum plantar flexion of 16-19° just after toe-off);
- after push-off ankle rapidly dorsiflexes during early swing for toe clearance;
- ankle dorsiflexion slows or stops during mid-swing but may continue to dorsiflex slightly in late swing until just prior to heel contact when plantarflexion begins.

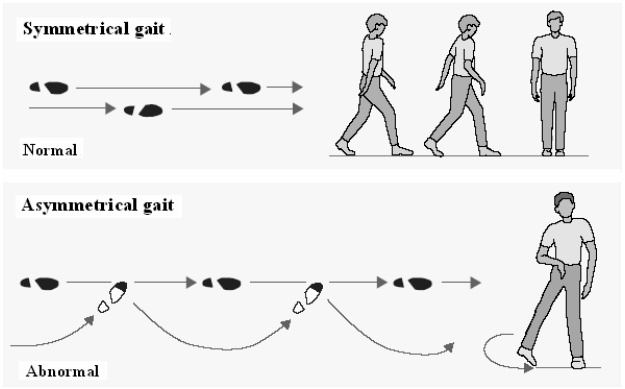


Fig.1.Gait analysis

Abnormal gait

Pathological gait describes altered gait patterns that have been affected by deformity (usually in the form of contractures), muscle weakness, impaired motor control. Any alteration affecting one or more motion or timing pattern can create a pathological gait pattern.

Deviations to normal gait patterns can be observed during both swing and stance phases and requires systematic evaluation for assessment of functional compensations and/or neuromuscular-skeletal factors. Functional compensations are voluntary posturings that attempt to substitute for specific motor weaknesses and joint instabilities. It is important to identify functional compensations from imposed mechanisms for appropriate orthotic design and therapeutic considerations.

Gait analysis can be used to evaluate more objectively the dynamic basis for an observed gait deviation in the patient requiring a lower limb orthosis. It also can be a valuable tool in objectively assessing the impact of different orthotic interventions.

Ankle-foot orthosis

A standard polypropylene AFO is a rigid polypropylene structure which prevents any ankle motion.

Different orthoses are used to enhance the ankle-foot position and mobility. The most common types are hingeless and hinge orthoses. Using springs, the hinge orthoses could assist ankle flexion/extension during gait, i.e. they are pseudo-active orthotic devices.

Which are to assist the quite popular gait abnormalities inherent to a spring-controlled human ankle-foot complex.

Previous studies have shown the DACS (dorsiflexion assist controlled by spring) AFO to have the following desirable characteristics: 1) the magnitude of the dorsiflexion assist moment and the initial ankle angle of the AFO can be changed easily, and 2) no plantarflexion assist moment is generated.

Torsion springs

Torsion springs (Fig.2) can store and release angular energy or statically hold a mechanism in place by deflecting the legs about the body centerline axis. They offer resistance to twist or rotationally applied force.

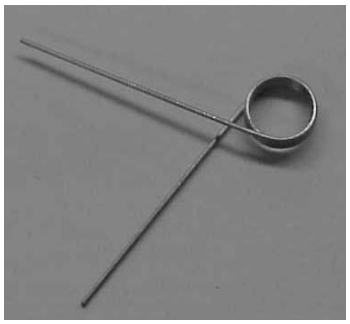


Fig.2. Torsion springs

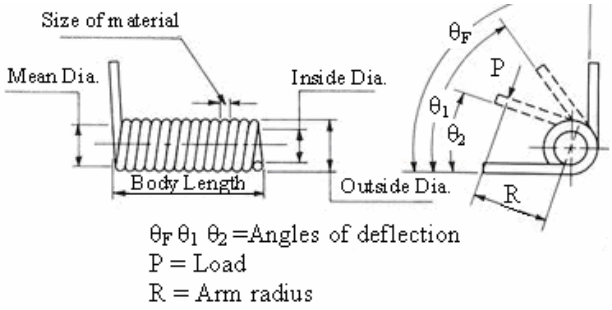


Fig. 3.Linear characteristic of the spring

Figure 3. shows the linear characteristic of the torsion spring

$M_{spring} = -k\theta$ linear spring

where M_{spring} is the spring torque, k is the spring constant and θ is the angular displacement from its rest angle

Dynamic equilibrium at joints

For normal leg, the dynamic equilibrium at each joint can be expressed as:

$M_i = M_g + M_s + M_d + M_a$

where M_i, M_g, M_s, M_d and M_a represent the torque due to moment of inertia of the rotating segment, gravity, joint stiffness, joint viscosity and muscle activation respectively. As there will be no muscle activity during the period considered, M_a becomes zero.

Results

Using the advanced open source 3D modeler (Blender3D) with outstanding user script capabilities (by Python script language extensions), different 3D computer solid models of MAFO were developed (Fig.4, Fig.5, Fig.6 and Fig.7). The main idea was to design two personalized AFO parts - lower (foot) and upper (calf), using 3D human model (artificially generated by specialized Blender3d script).

On the base of the obtained 3D surfaces two personalized AFO parts were designed and different variants of elastic elements (torsion spring) connecting the two parts of the orthosis.

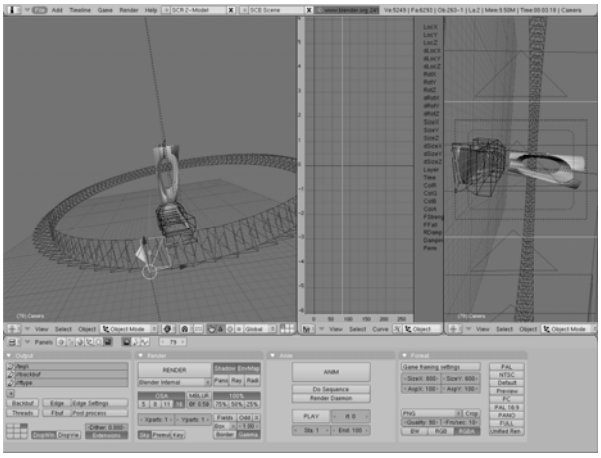


Fig.4. Dump screen of Blender3D during modeling process

A modular ankle-foot orthosis (MAFO) with one degree-of-freedom (dorsiflexion-plantarflexion motion) has been developed. The flexion/extension is controlled by springs. Dorsiflexion correction is achieved via the compression force of springs within the assistive device.

The big advantage of the 3D model is the possibility for further dynamics and kinematics development in field of more precise simulation of the real orthosis behavior. This could be achieved by combination of build-in physics engine of Blender3D, which covers both rigid and soft body simulation and user developed scripts.

Fig.5. 3D computer solid model of MAFO normal state

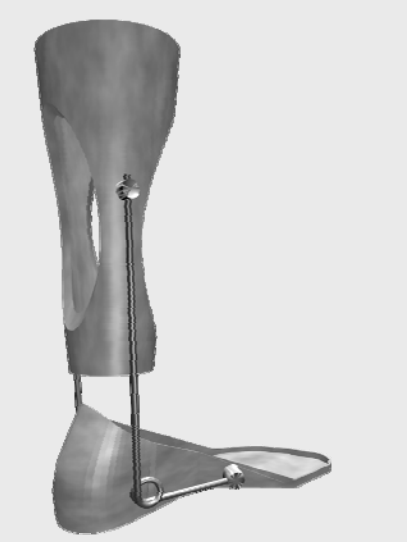


Fig.6. 3D computer solid model of MAFO in state of tension includes two parts: lower part (foot) and upper part (calf) with two torsion spring (lateral and medial).

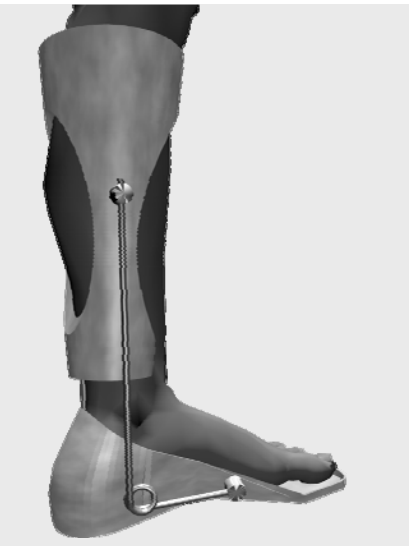


Fig.7. 3D model of the human ankle-foot segment with a model of orthoses.

Discussion

The magnitude of the MAFO dorsiflexion assist moment and the initial ankle angle can be modified by variation of the spring parameters (spring constant, spring rest angle). Regardless the simplicity of MAFO, the results in improvement of plantarflexion during swing phase are similar to the results obtain using commercial AFOs. The proposed modular ankle-foot orthosis is currently under additional mechanical durability tests. Continuous plantarflexion was applied to MAFO to check the durability of each part. At present, more than two million repetitions of plantar flexion have been applied and no serious problems have arisen. The results of the durability test will be use to improve the design of MAFO. Our MAFO can be used by the patients daily, and is also useful for gait training, since various characteristics can be easily modified. Moderately large dorsiflexion assist moment and small dorsiflexion initial ankle angle facilitates the increase of knee extension muscle forces, thus preventing forward foot slap during the initial stance phase.

References

Milusheva S., Tochev D., Karastanev S. (2005) Ankle-foot orthosis with spring elements. Proceedings of the 10th Congress on Theoretical and Applied Mechanics, Varna, pp. 145-149

http://www.fleshandbones.com/readingroom/pdf/1162.pdf

Acknowlegements

The study was supported by grant TN 1407/04 - Bulgarian National Research Fund.

Authors' Information

Slavyana Milusheva – e-mail: slavyana@imbm.bas.bg

Stefan Karastanev – Bulgaria; e-mail: stefan@info.imbm.bas.bg

Yuli Toshev – e-mail: ytoshev@imbm.bas.bg

Institute of Mechanics and Biomechanics, BAS, Acad.G.Bonthev St., bl.4, Sofia-1113, Bulgaria.

APPLICATION OF THE ARTIFICIAL INTELLIGENCE ALGORITHMS FOR SYSTEM ANALYSIS OF MULTI DIMENSION PHYSIOLOGICAL DATA FOR DEVELOPING POLYPARAMETRIC INFORMATION SYSTEM OF PUBLIC HEALTH DIAGNOSTICS

Nina Dmitrieva, Oleg Glazachev

Abstract. *The polyparametric intelligence information system for diagnostics human functional state in medicine and public health is developed. The essence of the system consists in polyparametric describing of human functional state with the unified set of physiological parameters and using the polyparametric cognitive model developed as the tool for a system analysis of multi dimension data and diagnostics of a human functional state. The model is developed on the basis of general principles geometry and symmetry by algorithms of artificial intelligence systems. The architecture of the system is represented. The model allows analyzing traditional signs - absolute values of electrophysiological parameters and new signs generated by the model – relationships of ounces. The classification of physiological multidimensional data is made with a transformer of the model. The results are presented to a physician in a form of visual graph – a pattern individual functional state. This graph allows performing clinical syndrome analysis. A level of human functional state is defined in the case of the developed standard ("ideal") functional state. The complete formalization of results makes it possible to accumulate physiological data and to analyze them by mathematics methods.*

Keywords*: information medical systems, system analysis of multi metric data and electrophysiological processes, diagnostics of adaptation processes, application of artificial intelligence algorithms.*

Introduction

One of problems of the contemporary preventive medicine is the development of an informational system of health diagnostics, which could be able to conduct a system analysis of multi dimension data, while could be comparable with existing clinical functional diagnostics and corresponding to the modern requirements [Hummel et al. 2000]. The experience obtained by us through the use of the visualized patterns and graphic modeling of functional states of an organism under activity of physiological substances [Dmitrieva et al. 1982] created the basis for development of the polyparametric method for evaluation of a human functional state in terms of the pattern recognition theory [Dmitrieva et al. 1989]. Patient data are presented in graphical formats as visual patterns, which permit to interpret these data in clinical-physiological terms. According to the recommendations of the Word Health Organization we have conducted the comparative research of a health state of students by polyparametric and clinical physiological methods [Dmitrieva, Glazachev, 2000]. These results demonstrated advantages and disadvantages of polyparametric method and lead us to development of new model on the basis of an artificial intelligence algorithms to improve one [Pospelov,1992; Zenkin, 1991].

Case-Based Reasoning

The gist of the polyparametric method for diagnostics of a human functional state consists in polygraphic recording (0.5 minutes) and data processing of objective physiological characteristics (electrocardiogram, electromiogram, electrovasogram and others), parameterization of analog signals, polymetric description of a functional state with the unified set of the time - amplitude matrix, using artificial intelligence algorithms and graphical modeling and methods of pattern recognition for an analysis of multi dimension data on line mode. Necessity and sufficiency of the set of parameters for a description of functional state have been grounded earlier. The novelty of the new variant of the polyparametric method contains in the original polyparametric cognitive model presenting the intelligence image system as the tool for the system analysis of multi dimension data.

The Intelligence Image System.

Absolute values of the unified set of physiological parameters mentioned above are represented as vectors in the system of polar coordinate (Fig.1). Each parameter has its own scale determined by modal level (middle circle). The contour limited with external and internal circumferences (maximal and minimal values of the parameters without pathognomonic signs) is the intelligence transformer performing analysis and classification of every parameter and whole shape, for example nosologic diagnosis. The active part of the intelligence transformer provides relationships of parameters as additional new signs generating new knowledge about a subject.

The changes of sympathetic or parasympathetic regulation are reflected in a displacement of a pattern (dotted circumference on fig.1) to the left and to the right correspondingly.

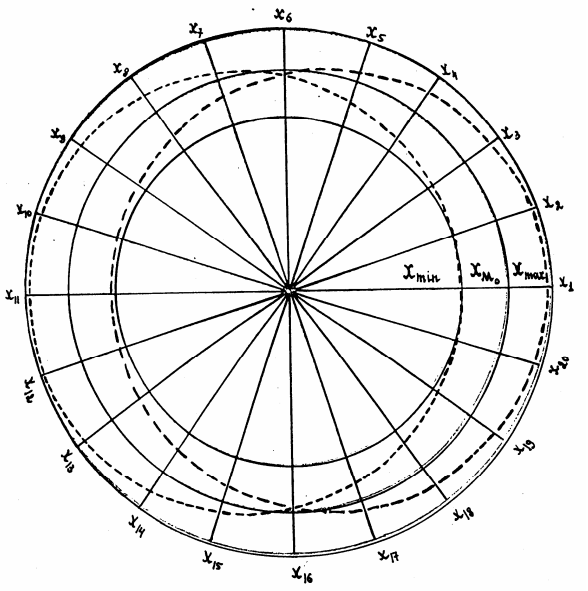


Fig.1. Artificial intelligence image model for analysis of physiological sings.
(Vectors X_1 – X_{20} are physiological parameters; the method of construction is described in the text).

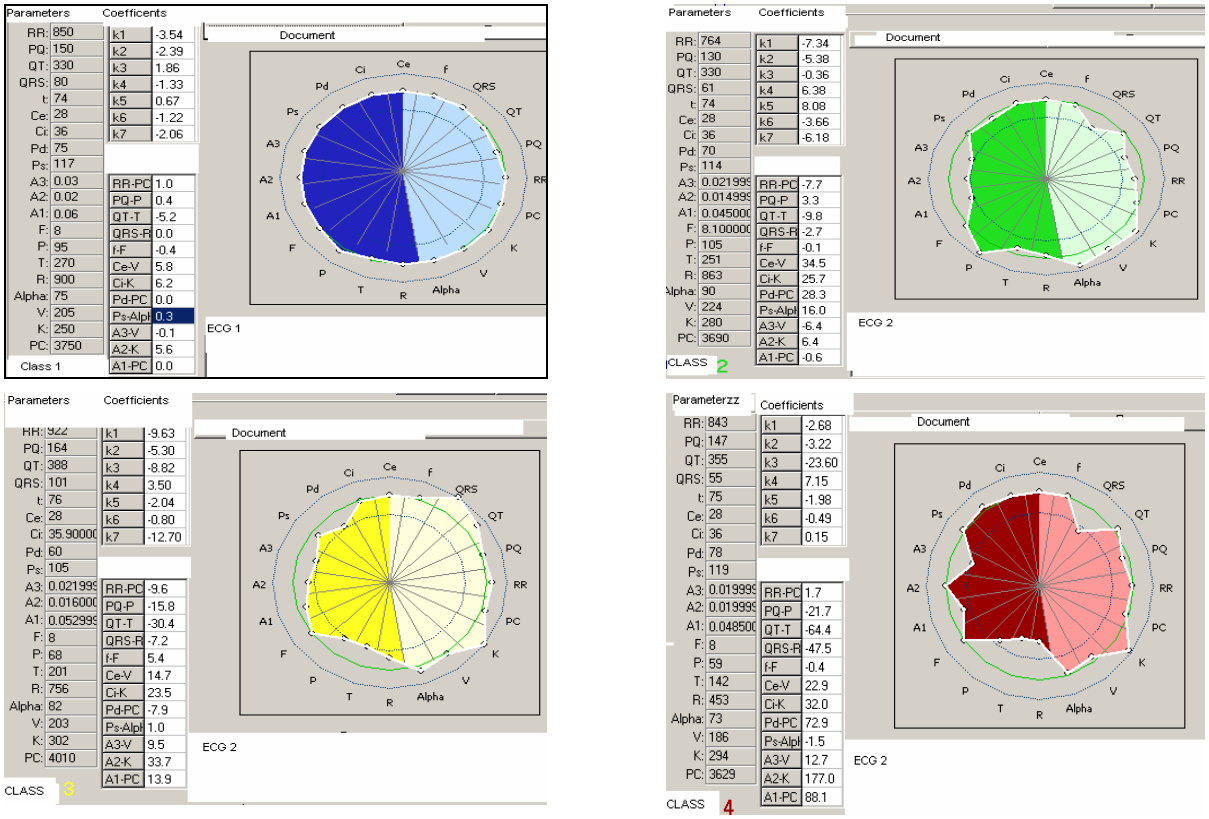


Fig.2. Patterns of different functional states. (1– satisfactory functional state, 2 – strained functional state, 3 – overstrained functional state, 4 – stress. Description of structure is in text)

The model of "ideal functional state" is characterized by the invariant relationships of all parameters (fig.2, left top). This model and patterns of individual functional states are constructed on the basis of the general intelligence model.

The results of the polyparametric examination are presented to physician in tabular form and as a pattern of functional state. On fig.2 there are 4 protocols of polyparametric examination patients with different functional sate: the first column is the list of physiological parameters and theirs values (in physical dimensions), the second column represents the relationships of parameters giving as the values of deviation (percent) from the invariant, visual graph (the pattern) of individual functional state to perform clinical analysis of multiple data in interactive mode.

The polyparametric method allows evaluating level of a functional state in on-line mode during 3-5 minutes.

The main characteristic of satisfactory state class is relationships of parameters closed to invariant: deviation is less than 5-7% whereas the absolute values of parameters can be in wide range between maximal and minimal rings. Thus the pattern of satisfactory functional state has a round form or similar to it. The deeper the adaptation syndrome the greater the misbalance of parameters is become. The pattern state assumes an irregular form of a different kind (Fig.2). It means that parameter relationships of the vital physiological functions are supplementary diagnostics signs of changes of human functional states. This is the new knowledge about information connections of physiological functions.

The state patterns are classified on decision rules in the PC programs and graded into four classes according to the main stages of adaptation processes development The results of the polyparametric examination according to the classes of functional state were controlled by criterion χ^2 (by program "S-Plus 2000 professional") and discriminative analysis (table 1).

Table1. Differences between functional classes by criterion χ^2 (by program "S-Plus 2000 professional")

Classes	4-1	4-2	4-3	3-1	3-2	2-1
Criteria	113.6	82.5	33.95	44.86	26.35	97.2
Mean	0.001	0.001	0.0034	0.001	0.034	0.009

Thus, with χ^2 criterions the main stages of adaptation process have objective differences between classes of functional states.

The discriminative analysis of the polyparametric data has confirmed their subdivision into 4 prescribed classes of functional states with satisfactory differences (under 9%).

As using parameters allow characterizing the autonomic regulation, the patterns of functional states can be considered as syndrome of autonomic status [Veyn, 1998]. For a definition of autonomic regulation every time parameter in the pattern marked with light color and amplitude parameters are darkened. This gives possibility for a physician to definite autonomic status ease and quickly [Dmitrieva, 1999].

Special investigations were found out significant individual differences of the patterns. The results of statistical analysis (mean values, mode, standard deviation and coefficient of variation) of polyparametric data demonstrated highly variable of some parameters for different people. It means the number of combinations of the changed signs can be high. The research of individual variations revealed that they can be satisfactorily systematized into major classes of states in respect of the standard model.

In support of that the polyparametric data were analyzed with the cluster method by the strategy of Word. On the Fig.3 the results of cluster analysis of polyparametric documents (on left) and statistic refined data (on right) are represented.

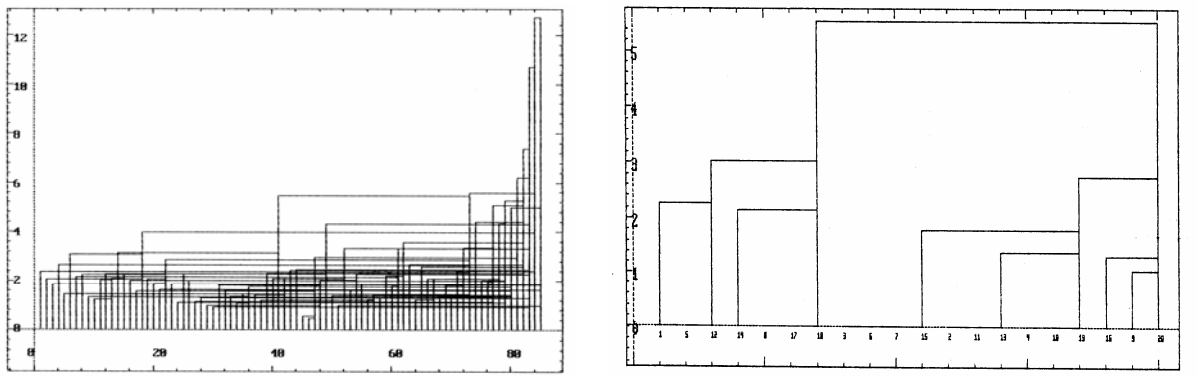


Fig3. Cluster analysis of documents of polyparametric examinations of students (axis X –individual documents, axis Y –cluster ranges).

There were singled out three main and four added clusters. The interpretation of clusters was performed by the visual analysis of individual documents. It was found out that same people examined in different time formed same clusters. Thus there were demonstrated that some syndromes are stable because of their patterns are fluctuated close to errors of a measure of parameters (± 7-8 %). Thus the reproducibility of derived parameters and their relationships has been confirmed. It was shown that a high formalization of the obtained results of the polyparametric examination makes to do systematization and mathematical analysis of multidimensional physiological data.

The patterns of typical adaptation syndromes were selected for using in the information support of physician decisions in diagnostics of a person adaptation syndrome [Seley H. 1976]. These patterns were fixed in phase 2 (Fig.4) to use them as the conceptual models of different adaptation syndromes for comparative analysis of newcomer patterns. The gist of this procedure is “data mining”. The scheme of DATA MINING of polyparametric technology is represented on fig.4.

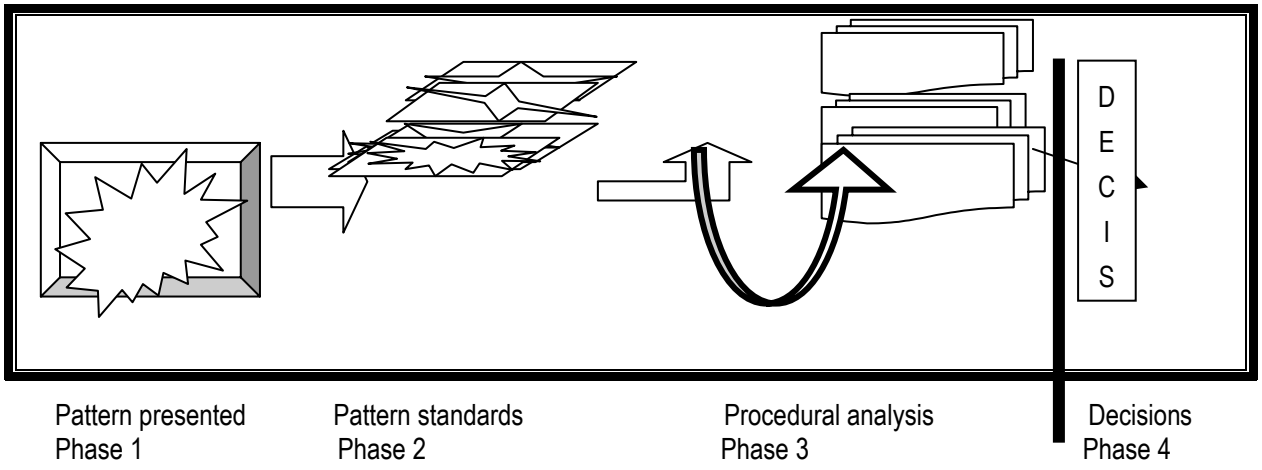


Fig.4. The scheme of polyparametric technology of diagnostics of multitude human functional states.

The phase 3 of the technology is intended for a predictable decision by using the known rules of procedure “if S then A” [Newell, Simon, 1972]. It is easy to see changes of any pattern and to interpret a dynamic of a functional state using these rules working with the model.

But the obtained results do not allow making a definite conclusion about exact role of concrete signs in the forming of pattern.

The polyparametric method and technology are open for further development and improvement based on the bank-accumulated data of polyparametric examinations and evaluations of a functional state under therapeutics and correction mode.

Application of the polyparametric method in condition of the comparative analysis of the results of examination of students with the clinical methods was conducted. The findings demonstrated that 44% of students during their term are in a state of overstrain and 40% in adaptation failure according to the classification of stages of adaptation development processes. Thus the comparative analysis showed the good correlation and correspondence.

Conclusion

The information intelligence image polyparametric system represents the instrument to analyze multidimensional data, performing knowledge engineering and data mining. It allows very quickly define level of a human functional state relatively to the developed standard of functional state. The image of functional state is very easy for interpretation. Complete formalization of the polyparametric results makes it possible to accumulate data and analyze them by math methods. This tool allows determining and evaluating relationships of electrophysiological parameters, which became a new diagnostics signs. The conceptual model of standard ("ideal") functional state was formed. This model allows to circumvent the indefinite notions of "norm" and "ordinary man" and work out the technique for measurement of various deviations from the "ideal functional state" as changes of the functional state.

Today the first experimental model of computer polyparametric system is installed in the Moscow State University, where the health of students has been examined for some years. The polyparametric method is open for further development of the preclinical diagnostics of functional disorders on the "training-with-a teacher" basis.

Bibliography

[Hummel et al.2000] Hummel M, van Rossum W., Verkerke G., Rakhorst G. Assessing medical technologies in development. //Intern. J. of Technology Assessment in Health Care. 2000; 16:4.

[Dmitrieva et al. 1982] Dmitrieva N.V., Nizhnii S.V., Ivanova I.V. Chemical (drug) stress – qualitative assessment of toxic effect of physiologically active substances // Izvestiya AN SSSR. Ser. Biol. – 1982. – N 3. – P. 398

[Dmitrieva et al. 1989] Dmitrieva N.V., Voronov E.B., Yakovlev U.V. at al. The Polyparametric method of evaluation of human functional state with image recognition method. // Phisiologiay cheloveka. 1989. 4. P. 103-112.

[Dmitrieva, Glazachev, 2000] Dmitrieva N. V., Glazachev O.S. Individual Health and Polyparametric Diagnosis of Organism's functional state. M.: Gorizont, 2000. – 214 p.

[Pospelov, 1992] Pospelov D.A. Cognitive graphic – a window in new world // News Art, Intel.. 1992 .3.4.

[Zenkin, 1991] Zenkin A.A. The cognitive computer's graphic. "Nauka", Novosybirsk. 1991. -186p.

[Dmitrieva, 1999] Dmitrieva N.V. Syndrome analysis of polyparametric images of functional states of organisms // News of Artificial Intellect. – 1999. – N 1. – P. 120–129.

[Seley, 1976] H. Seley, Stress in health and disease. Boston-London, 1976, -1256 p.

[Newell, Simon, 1972] Newell A., Simon H. Human problems solving. Englewood Cliffs: Prentice Hall, 1972.

Author's Information

Nina V. Dmitrieva – M.D. Prof. Head of the System Analysis Laboratory Institute of Normal Physiology Russian Academy of Medical Sciences, Mochovay str.11.Moscow, 125009 Russia; e-mail: nvdmitrieva@mtu-net.ru

Oleg S. Glazachev - M.D. Prof. of Moscow Medical Academy, , Mochovay str.11.Moscow, 125009 Russia.

Knowledge Engineering

ON LOGICAL CORRECTION OF NEURAL NETWORK ALGORITHMS FOR PATTERN RECOGNITION

L.A. Aslanyan, L.F. Mingo, J.B. Castellanos, V.V. Ryazanov, F.B. Chelnokov, A.A. Dokukin

Abstract: *The paper is devoted to the description of hybrid pattern recognition method developed by research groups from Russia, Armenia and Spain. The method is based upon algebraic correction over the set of conventional neural networks. Output matrices of neural networks are processed according to the potentiality principle which allows increasing of recognition reliability.*

Keywords: *Pattern recognition, forecasting, neural networks, algebraic correction.*

Mathematical recognition theory has long history and the variety of its reality modeling methods is quite wide. Every research group has its own traditions and usually works in specific area of mathematics. There are two basic approaches which are commonly said to be different. They are functional and algorithmic ones. For example, neural networks approximate output function but their parameters has no appropriate interpretation. Algorithmic models as for example algorithms of estimates calculating provide interpretable parameters though may have high calculation difficulty. Integration of scientific schools and small groups of "particular specialists" in the framework of joint projects provide possibilities for revealing potentials of different methods and their combinations. Developing of one such integrated approach is connected to the execution of series of INTAS projects by research groups from Russia, Spain, Armenia and some other countries.

Algebraic theory of pattern recognition based upon discrete analysis and algebra [1] is the basic approach which has been being used for 35 years in the Computing Centre of RAS under the direction of academician Yu.I. Zhuravlev. Research activities of the Institute for Informatics and Automation Problems of NAS Armenia lie in the same area of discrete recognition models. Their specific is the use of optimization structures of discrete isoperimetric tasks, discrete topology and hierarchical class searching [2]. Neural network models especially ones with polynomial output and linear activation functions [3] are the main area of interest of the Spanish group. In particular, they research temporal signal delays in recognition tasks. Good results have been achieved in forecasting of stock exchange and similar problems.

Some hybrid methods and applications for pattern recognition have been developed by these groups in the framework of INTAS projects 96-952, 00-367, 00-636 and 03-55-1969. One of them is based on assembling of neural networks and logical correction schemes. The main cause of this research was the idea of creating such pattern recognition and forecasting application which requires minimal human intervention or no intervention at all. It should be possible for the operator with no specific knowledge in mathematics to use that software. Such NNLC (Neural Networks with Logical Correction) application has been developed in the framework of INTAS projects 03-56-182 inno and 03-55-1969 YSF. Now we are proud to say that it has justified our expectations in a great extent. The method has shown high and stable results in many practical tasks.

Further we shall describe general training and recognition scheme for the l-classes task. The notation from [1] will be used. Let the training sample be $S_1, S_2, ..., S_m$ and the testing one $S'_1, S'_2, ..., S'_q$:

$$S_{m_{i-1}+1}, S_{m_{i-1}+2}, ..., S_{m_i} \in K_i, i = 1,2,...,l, m_0 = 1, m_l = m,$$

$$S'_{q_{i-1}+1}, S'_{q_{i-1}+2}, ..., S'_{q_i} \in K_i, i = 1,2,...,l, q_0 = 1, q_l = q.$$

For simplicity sake let us also suppose the task is solved without denials.

Finally, let us have N neural networks $A_j(S) = (\alpha_1^j(S), \alpha_2^j(S), ..., \alpha_l^j(S))$ trained for this task. It will give us the following matrix of recognition results:

$$A_j(S'_t) = (\alpha_1^j(S'_t), \alpha_2^j(S'_t), ..., \alpha_l^j(S'_t)),\ \alpha_i^j(S'_t) \in \{0,1\}, i = 1,2,...,l, \cdot j = 1,2,...,N, \cdot t = 1,2,...,q.$$

Algorithm of recognition by the group of neural networks will be designed according to the principle of potential correction [4]. New object will be assigned to the class of maximum estimation which is calculated according to the following formula:

$$\Gamma_i(S) = \frac{1}{q_j - q_{j-1}} \sum_{t=q_{j-1}+1}^{q_j} \Phi_i(S'_t, S), \cdot i = 1,2,...,l.$$

The variable $\Phi_i(S'_t, S)$ is called the potential between S'_t и S and is calculated as follows:

a) $\Phi_i(S'_t, S) = \begin{cases} 1, & \left|\{\alpha_i^j(S) \ge \alpha_i^j(S'_t), j = 1,2,...,N,\}\right| / N \ge \delta, \\ 0, & otherwise. \end{cases}$

b) $\Phi_i(S'_t, S) = \{the\ number\ of\ correct\ inequalities \cdot \alpha_i^j(S) \ge \alpha_i^j(S'_t), \quad j = 1,2,...,N\}.$

A-type potential we will call monotonous, b-type one will be called weekly monotonous with monotony parameter $\delta, \quad 0 < \delta \le 1$.

Thus, training phase consists of training of N neural networks (with no denials) and consequent calculation of binary matrix $\left\|\alpha_i^j(S'_t)\right\|_{l \times N \times q}$. New object S is classified by calculating its binary matrix $\left\|\alpha_i^j(S)\right\|_{l \times N}$ and its estimates for each class according to either a-type or b-type potential. As we have already mentioned software realization of the method has been made by means of NNLC application. By the grant system of INTAS organization the NNLC application has been qualified as innovation software.

Acknowledgements

The authors are glad to acknowledge support of the following organizations for execution of the described research: INTAS (projects 03-56-182 inno, 04-77-7076, 03-55-1969 YSF), RFBR (projects 05-07-90333, 06-01-00492, 05-01-00332). The work has been also supported by the program N 14 of RAS Presidium's.

Bibliography

[1] Zhuravlev Yu.I., On algebraic approach for pattern recognition or classification // Cybernetics problems (Problemy kibernetiki), Nauka, Moscow, 1978, N33, pp. 5-68. (In Russian).

[2] Aslanyan L., Zhuravlev Yu., Logic Separation Principle, Computer Science & Information Technologies Conference // Yerevan, 2001, pp. 151-156.

[3] Luis Mingo, Levon Aslanyan, Juan Castellanos, Miguel Diaz and Vladimir Riazanov // Fourier Neural Networks: An Approach with Sinusoidal Activation Functions. International Journal Information Theories and Applications. Vol. 11. ISSN: 1310-0513. 2004. Pp. 52-53.

[4] Zuev Yu.A., Method for increasing of classification reliability based upon monotony principle in case of multiple classifiers // Journal of Calculating mathematics and Mathematical Physics, 1981, T.21, N1, pp. 157-167.

Authors' Information

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

L.F. Mingo – Dpto. 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

J.B. Castellanos – Dpto. Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid; Boadilla del Monte – 28660 Madrid, Spain; e-mail: jcastellanos@fi.upm.es

V.V. Ryazanov – Department of Mathematical Pattern Recognition and Methods of Combinatorial Analysis, Computing Centre of the Russian Academy of Sciences; 40 Vavilova St., Moscow GSP-1, 119991, Russian Federation; e-mail: rvvccas@mail.ru

F.B. Chelnokov – Department of Mathematical Pattern Recognition and Methods of Combinatorial Analysis, Computing Centre of the Russian Academy of Sciences; 40 Vavilova St., Moscow GSP-1, 119991, Russian Federation; e-mail: fchel@mail.ru

A.A. Dokukin – Department of Mathematical Pattern Recognition and Methods of Combinatorial Analysis, Computing Centre of the Russian Academy of Sciences; 40 Vavilova St., Moscow GSP-1, 119991, Russian Federation; e-mail: dalex@ccas.ru

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

Levon Aslanyan, Juan Castellanos

Abstract*: Pattern recognition (classification) algorithmic models and related structures were considered and discussed since 70s: – one, which is formally related to the similarity treatment and so - to the discrete isoperimetric property, and the second, - logic based and introduced in terms of Reduced Disjunctive Normal Forms of Boolean Functions. A series of properties of structures appearing in Logical Models are listed and interpreted. This brings new knowledge on formalisms and ontology when a logic based hypothesis is the model base for Pattern Recognition (classification).*

1. Introduction

Pattern Recognition is in reasonable formalization (ontology) of informal relations between objects visible/measurable properties and of object classification by an automatic or a learnable procedure. Among the means of formalization (hypotheses) - metric and logic based ones are the content of series of articles started by the current one. The stage of pattern recognition algorithmic design in 70s dealt with algorithmic models – which are huge parametric structures, combined with diverse optimization tools. Algorithmic Models cover and integrate wide groups of existing algorithms, integrating their definitions, and multiplying their resolution power. Well known example of this kind is algorithmic model of estimation of analogies (AEA) given by Yu. I. Zhuravlev [1]. This model is based indirectly on compactness hypothesis, which is theoretically related to the well known discrete

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

isoperimetric problem (3). The optimization problem of isoperimetry is a separate theoretical issue and its pattern recognition implementations are linked alternatively to the general ideas of potential functions [4]. We present the logical separation (LSA) algorithmic model, as it is described below, to be one of the generalizations of algorithmic model of estimation of analogies. For AEA models a number of useful combinatorial formulas (algorithms) to calculate the analogy measure of objects and of objects and classes were proven [2]. These are the basic values for the decision making rules in AEA. In these models large number of parameters appears, being consecutively approximated using the appropriate optimization procedures. For this reason, a special control set besides the learning set is considered having the same formal structure as the learning set. Considering classification correctness conditions for the set of given objects by the decision procedure we get a system of restrictions/inequalities, which may not be consistent. In the simplest case a system of linear inequalities appear and then we receive a problem of approximating the maximal consistent subsystem of this basic requirements system. In terms of Boolean functions this is equivalent to the well known optimization problem of determining of one of the maximal upper zeros of a Monotone Boolean function when it is given by an operator.

LSA is based on implementation of additional logical treatments on learning set elements, and above the AEA specific metric considerations. Some formalization of additional properties on classification in this case is related to the terms of Boolean functions and especially - to the reduced disjunctive normal forms of them. Let us consider a set of logical variables (properties) $x_1, x_2, ..., x_n$ and let we have two types/classes for classification: K_1 and K_2. Let $\beta \in K_1$, and $\gamma \in K_2$, and α is an unknown object in the sense of classification. We say, that γ is separated by the information of β for α if $\beta \oplus \gamma \le \beta \oplus \alpha$, where $\oplus$ is summation by $mod\,2$ operation. After this assumption we get, that the reduced disjunctive normal forms of two complementary partially defined Boolean functions describe the structure of information enlargement of the learning set. This construction is extending the model of estimation of analogies. It was shown that the logical separators divide the object sets into three subsets, where only one of them needs the treatment by AEA. This set is large enough for almost all weakly defined Boolean functions, but for the functions with the property of compactness it is small. Let, for $0 \le k_0 < k_1 \le n$ F_{n,k_0,k_1} is the set of all Boolean functions consisting of pair of k_0 and $n - k_1$ spheres centered at 0 and 1 respectively as the sets of zeros and ones of the function. On the remainder part of vertices of n-cube the assignment/evaluation of the functions are arbitrary. This functions satisfy the compactness assumptions, 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, also, it is enough learning set, consisting of any $n2^{n-\varepsilon(n)\sqrt{n}}$ or more arbitrary points for recovering the full classification by means of logical separators procedure. This is an example of postulations considered. The given one is relating the metric and logic structures and suppositions, although separately oriented postulations are listed as. 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. Logic Based Model

Solving the main problem of pattern recognition or classification assumes that indirect or informal information or data on classification $K_1, K_2, ..., K_l$ is given. Often this information is in form of appropriate conditions in an analogy to the compactness hypothesis, which in the very common shapes assumes, that given a metric in the space of all objects M and that closer values of classification predicate K corresponds to the pairs of "near" objects of M. We assume that objects of M are coded - characterized by the collections of values of n properties $x_1, x_2, ..., x_n$. Then each object is identified with the corresponding point of the n-dimensional characteristic space. So under the compactness of classes we assume the geometrical compactness of sets of points in the characteristic space, which corresponds to the elements of classes $K_1, K_2, ..., K_l$ and the

consecutive adjustments of this property can be given in the following descriptive form: closer neighborhoods of class elements belong to the same class; the distance increase from a class element increases the class change probability; for elements pairs of different classes there exist simple paths in three parts – classes and a limited transition area in the middle.

Above we already considered the general formalization models of hypothesis by metrics and by logic. More formalizations move to more restricted sets of allowable classifications and in this regard it is extremely important to determine the level of formalisms applied. During the practical classification problem arrangements it is to check the satisfaction level of the application problem to the metric and/or logic hypothesis. Resolution is conditioned by the properties of the given learning set $\bigcup_{i=1}^{l} \mathrm{M}_i$. On the other hand there are more different conditions and methods of classification, which are very far in similarity to the model of compactness. These structures require and use other formalisms, providing the solution tools to the wide amounts of practical of pattern recognition problems. Such are the classes of algorithms of estimation of analogies, test's algorithms [2] potential function methods [4], etc.

Note that the arbitrary pattern recognition class problems can be reduced to the others, with the satisfaction of compactness type hypothesis. However this doesn't mean that the compactness hypothesis is universal because the pattern recognition problem's solution for the given space or creation of appropriate transformations to the other problems are the equivalent problems.

Now let us formulate the condition F_0, which will formalize the additional to the compactness hypothesis properties of classes. We'll consider the case of two classes ($l = 2$) intending the formalism simplifications. Particularly, in case of completing of partially defined predicate P, we will base on condition F_0. We'll apply a correspondence of considered object properties and the set of binary variables $x_1, x_2, ..., x_n$, and the same time between the partial predicate P - and it's characteristic function $f(x_1, x_2, ..., x_n)$, and will solve the modeled problem of determining (completing) of the target Boolean function F.

Let $\tilde{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_n) \in \mathrm{M}$.

Consider the determination (completion) of function f in $\tilde{\alpha}$. Take the arbitrary $\tilde{\gamma}$, $\tilde{\gamma} \in \mathrm{M}_1$. If there exists such a point $\tilde{\beta}$, $\tilde{\beta} \in \mathrm{M}_0$, that $\tilde{\gamma} \oplus \tilde{\beta} \le \tilde{\gamma} \oplus \tilde{\alpha}$ (so the $\tilde{\beta}$ is different of $\tilde{\gamma}$ on a subset of the set of properties, where are different $\tilde{\alpha}$ and $\tilde{\gamma}$), then we conclude that $\tilde{\beta}$ logically separates $\tilde{\alpha}$ from $\tilde{\gamma}$, and the information, that $f(\tilde{\gamma}) = 1$ doesn't affect on the determination of the value of the function f in the point $\tilde{\alpha}$ by 1. In the opposite case we'll call $\tilde{\alpha}$ allowable in respect to the point $\tilde{\gamma}$ and to the set M_1 and decide, that information $f(\tilde{\gamma}) = 1$ influence on the determination of $\tilde{\alpha}$ by one, and the real measure of that is given by the value of the object similarity measures.

Consider the following classes of points of the n--dimensional unit cube:

N_0^f -- the set of all $\tilde{\alpha} \in \mathrm{M}$, which are allowable for the set M_0 and not allowable for M_1,

N_1^f -- the set of all $\tilde{\alpha} \in \mathrm{M}$, which are allowable for the set M_1 and not allowable for M_0,

N_2^f -- the set of all $\tilde{\alpha} \in \mathrm{M}$, which are not allowable for the sets M_0 and M_1,

N_3^f -- the set of all $\tilde{\alpha} \in \mathrm{M}$, which are allowable for both the M_0 and M_1.

[3] pointed out the general relation of condition F_0 with the notion of the reduced disjunctive normal form of Boolean functions. To see this relation let us consider the functions f and its negation $\bar{f}$, and let $\Re_f$, $\Re_{\bar{f}}$ correspondingly are the reduced forms for these functions. Denote by:

M_0^f --the collection of all points $\tilde{\alpha}$ for which $(\Re_f)_{\tilde{\alpha}} = 0$, $(\Re_{\bar{f}})_{\tilde{\alpha}} = 1$,

M_1^f --the collection of all points $\tilde{\alpha}$ for which $(\Re_f)_{\tilde{\alpha}} = 1$, $(\Re_{\bar{f}})_{\tilde{\alpha}} = 0$,

M_2^f --the collection of all points $\tilde{\alpha}$ for which $(\Re_f)_{\tilde{\alpha}} = 0$, $(\Re_{\bar{f}})_{\tilde{\alpha}} = 0$,

M_3^f --the collection of all points $\tilde{\alpha}$ for which $(\Re_f)_{\tilde{\alpha}} = 1$, $(\Re_{\bar{f}})_{\tilde{\alpha}} = 1$,

Proposition 1. $\mathrm{N}_i^f \equiv \mathrm{M}_i^f, i = 0.1.2.3.$

Proposition 2. If $\mathrm{M}_0 \cup \mathrm{M}_1 \neq 0$, then M_2^f is empty, in opposite case $\mathrm{M}_2^f \equiv \mathrm{M}$.

It is simply to prove this and some of the consecutive propositions and by this reason we omit the complete proofs and give the main idea of that. So, to prove proposition 2 consider an arbitrary point $\tilde{\alpha} \in \mathrm{M}$. If $\mathrm{M}_0 \cup \mathrm{M}_1 \neq 0$, then let us take the distance of $\tilde{\alpha}$ to the set $\mathrm{M}_0 \cup \mathrm{M}_1$ (which equals the minimal possible distance of $\tilde{\alpha}$ from any of the points of $\mathrm{M}_0 \cup \mathrm{M}_1 \neq 0$), which is in some point $\tilde{\beta} \in \mathrm{M}_0 \cup \mathrm{M}_1$. Suppose, without loss of generality, that $\tilde{\beta} \in \mathrm{M}_0$. Then the interval (binary subcube) $\mathrm{E}(\tilde{\alpha}, \tilde{\beta})$, constructed on base of points $\tilde{\alpha}$ and $\tilde{\beta}$ does not contain points from the set M_1. From here, on base of definition of reduced disjunctive normal form implies, that the point $\tilde{\alpha} \in \mathrm{M}$ is allowable in respect to the set M_0.

Proposition 3. If f_0 is an appropriate completion of function f, constructed on base of condition F_0, then $\forall \tilde{\alpha} \in \mathrm{M}_0^f (f_0(\tilde{\alpha}) = 0)$ and $\forall \tilde{\beta} \in \mathrm{M}_1^f (f_0(\tilde{\beta}) = 1)$.

Proposition 4. $\mathrm{M}_0 \subseteq \mathrm{M}_0^f$ and $\mathrm{M}_1 \subseteq \mathrm{M}_1^f$.

As a consequence from these two propositions we get, that the arbitrary completion of function f, which is made on base of condition F_0, constructs the function, allowable in respect of f. In terms of pattern recognition problems this means that arbitrary methods of recognition, which are based on the condition F_0, couldn't be "false" on the elements of the learning set $\mathrm{M}_0 \cup \mathrm{M}_1$. Write out the minimal completions of the function f, constructed on base of the condition F_0:

$$f_1(x_1, x_2, ..., x_n) = \begin{cases} 0, & if\ (x_1, x_2, ..., x_n) \in \mathrm{M}_0^f \\ 1 & if\ (x_1, x_2, ..., x_n) \in \mathrm{M}_1^f \\ is\ not\ determined & if\ (x_1, x_2, ..., x_n) \in \mathrm{M} \setminus (\mathrm{M}_0^f \cup \mathrm{M}_1^f) = \mathrm{M}_3^f \end{cases}$$

So we get some "enlargement" for the basic function f. There arose a question -- might f_1 be the new starting point (learning set, function) for the completion on base of condition F_0, and how close we can approach by this steps the final goal?

The answer gives the

Proposition 5. If f_{i+1} is completion of partial function f_i, constructed on base of condition F_0, $i=1,2,...$, then $f_1 \equiv f_k, k=1,2,...$.

Let us now analyze the conditions, related to the successful continuation on base of F_0 of a partial Boolean function (that is the case of the solvable problems). Let f -- is a partially defined Boolean function and $\varphi_1,\varphi_2,...,\varphi_\tau$ -- all that functions of class $P_2(n)$ which might appear as a continuation of function f, constructed by the given assumptions. Then we are interested in conditions, when extension f_1 is allowable in respect to each of functions $\varphi_1,\varphi_2,...,\varphi_\tau$.

Consider the function f_0, defined in the following way:

$$f_0(x_1,x_2,...,x_n)=\begin{cases} 0, & if\ \varphi_i(x_1,x_2,...,x_n)=0, i=1,2,...,\tau \\ 1, & if\ \varphi_i(x_1,x_2,...,x_n)=1, i=1,2,...,\tau \\ is\ not\ defined & in\ other\ cases \end{cases}$$

Denote by $M_0(f_0)$ and $M_1(f_0)$ sets of all n--cube vertices, where function f_0 achieves values 0 and 1 respectively. Then our requirement can be formulated as the following: $M_0^f \subseteq M_0(f_0)$ and $M_1^f \subseteq M_1(f_0)$. Here $M_3^f = M \setminus (M_0^f \cup M_1^f)$ and $M_3^f \supseteq M \setminus (M_0(f_0) \cup M_1(f_0))$ so that this partial continuation doesn't violate the continuality of starting function to the each of functions $\varphi_1,\varphi_2,...,\varphi_\tau$. It is to mention that the conditions $M_0^f \subseteq M_0(f_0)$ and $M_1^f \subseteq M_1(f_0)$ are not convenient, which is related to the applied information on the final goal (the functions $\varphi_1,\varphi_2,...,\varphi_\tau$). Supposing the case of continuation for needs of pattern recognition problems let us show that practically useful conditions of the given type might be formulated.

Consider the structural behavior, when $n \to \infty$ and suppose a parameter $\theta \to 0$ given as. Suppose $f_0 \in P_2(n)$ (note, that the results below are true in more general cases and in more general forms). Here are some preliminary results.

1. Consider the concept $H_k^-(f_0)$ introduced by Glagolev [7]. $H_k^-(f_0)$ equals the number of vertices $\tilde{\alpha} \in E^n$, where $f_0(\tilde{\alpha})=1$, and which are covered by (involved in) maximal intervals of function f_0 of sizes not exceeding k. It was proven [7] that for almost all functions $f_0 \in P_2(n)$ $H_k^-(f_0) = o(2^n)$ when $n \to \infty$ and $k \le k_1 = log\,log\,n - 1$.
2. We'll say, [5] that the cube vertices prick the intervals including these vertexes. The set A of vertices of n-dimensional unit cube is a pricking set for the set B_k -all of the k -size intervals, if each k -subcube is pricked at least by one of the vertices of A. Denote by $K(n,k)$ the minimal number of vertices, forming a pricking set for k -subcubes. By [5] $2^{n-k} \le K(n,k) \le (n+1)2^{n-k}$. We will use the upper bound by this formulae but in our case $k \le k_1 = log\,log\,n - 1$ and a better estimation is possible as follows [4] (an extended survey on pricking is included in [6]). Let us denote by $A_i(\tilde{\alpha})$ the set of all of n-cube vertices, which lay in respect to the given vertex $\tilde{\alpha}$ on layers with numbers $\tau, \tau \equiv i(mod\,k+1)$, $i=0,1,...,k, k \le n$. Let E^k -is an arbitrary k -subcube of an n-cube. Points of subcube E^k are placed exactly in the $k+1$ consecutive layers of E^n in respect to it's arbitrary vertex $\tilde{\alpha}$. It is correct to post the

Proposition 6. Each of the sets $A_i(\tilde{\alpha}), \tilde{\alpha} \in E^n, i = 0,1,...,k$ are pricking for the set B_k -all of the k -subcubes of n-cube, and $2^{n-k} \le K(n,k) \le 2^n / k + 1$.

Proposition 7. F_0 implemented in continuation of almost all functions $f_0 \in P_2(n)$ yields the accuracy, tending to 1 as $n \to \infty$, if for the initial function f holds the condition $\mathrm{M}_0 \cup \mathrm{M}_1 \supseteq A_i(\tilde{\alpha})$ at least for any $i, i = 1,2,...$ and vertices $\tilde{\alpha} \in E^n$, where $A_i(\tilde{\alpha})$ is constructed for a $k \le [\log\log n] - 1$.

Note, that the proposiition 7 postulation is constructive, and it implies to the "sufficient" learning set, consisted no more that from $2^n / \log\log n$ points (which is $o(2^n)$) as $n \to \infty$. However, basically, in the pattern recognition problems it is impossible to obtain the learning set arbitrarily. Often it is formed as a random collection of any fixed size of the main collection of the studying objects.

Conclusion

Logic Separation is an element of pattern recognition hypotheses and formalisms. Structures appear in this relation based and introduced in terms of Reduced Disjunctive Normal Forms of Boolean Functions. An initial set of properties of these structures were introduced in propositions 1-7.

Bibliography

1. Zhuravlev Yu. I. On an algorithmic approach to the problems of recognition and classification. Problemi Kibernetiki, 33, (1978) 5--68.
2. Zhuravlev Yu. I. Selected Scientific Works, Publishing House Magister, Moscow, (1998) 417p.
3. Aslanyan L. H. On a pattern recognition method based on the separation by the disjunctive normal forms. Kibernetika, 5, (1975), 103--110.
4. Vapnik V. and Chervonenkis A. Theory of Pattern Recognition. “Nauka”, 1974.
5. Aslanyan L. H. The Discrete Isoperimetric Problem and Related Extremal Problems for Discrete Spaces. Problemy Kibernetiki, 36, (1979), 85--128.
6. Nechiporuk E. I. On topological principles of self-correcting. Problemy Kibernetiki, 21, (1969), 5--102.
7. Graham N., Harary F., Livingston M. and Stout Q. Subcube Fault-Tolerance in Hypercubes. Information and Computation 102 (1993), pp. 280{314.
8. 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

Juan Castellanos – Dpto. Intelligencia Artificial – Facultad de Informatica – Universidad Politecnica de Madrid 28660 Boadilla del Monte – Madrid, Spain; e-mail: J.Castellanos@fi.upm.es

DNA SIMULATION OF GENETIC ALGORITHMS: FITNETSS COMPUTATION[1]

Maria Calvino, Nuria Gomez, Luis F. Mingo

***Abstract**: In this paper a computational mode is presented base on DNA molecules. This model incorporates the theoretical simulation of the principal operations in genetic algorithms. It defines the way of coding of individuals, crossing and the introduction of the individuals so created into the population. It resolves satisfactorily the problems of fitness coding. It shows also the model projection for the resolution of TSP. This is the basic step that will allow the resolution of larger examples of search problems beyond the scope of exact exponentially sized DNA algorithms like the proposed by Adleman [Adleman, 1994] and Lipton [Lipton, 1995].*

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

Introduction

Deoxyribonucleic acid (DNA) is the chemical inside the nucleus of all cells that carries the genetic instructions for making living organisms. A DNA molecule consists of two strands that wrap around each other to resemble a twisted ladder. The sides are made of sugar and phosphate molecules. The "rungs" are made of nitrogen-containing chemicals called bases. Each strand is composed of one sugar molecule, one phosphate molecule, and a base. Four different bases are present in DNA - adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases arranged along the sugar - phosphate backbone is called the DNA sequence; the sequence specifies the exact genetic instructions required to create a particular organism with its own unique traits. Each strand of the DNA molecule is held together at its base by a weak bond. The four bases pair in a set manner: Adenine (A) pairs with thymine (T), while cytosine (C) pairs with guanine (G). These pairs of bases are known as Base Pairs (bp).

DNA and RNA computing is a new computational paradigm that harnesses biological molecules to solve computational problems. Research in this area began with an experiment by Leonard Adleman [Adleman, 1994] in 1994 using the tools of molecular biology to solve a hard computational problem. Adleman's experiment solved a simple instance of the Travelling Salesman Problem (TSP) by manipulating DNA. This marked the first solution of a mathematical problem with the tools of biology.

Computing with DNA generated a tremendous amount of excitement by offering a brand new paradigm for performing and viewing computations. The main idea is the encoding of data in DNA strands and the use of tools from molecular biology to execute computational operations. Besides the novelty of this approach, molecular computing has the potential to outperform electronic computers. For example, DNA computers may use a billion times less energy than electronic computers, while storing data in a trillion times less space. Moreover, computing with DNA is highly parallel: in principle there could be billions upon trillions of DNA or RNA molecules undergoing chemical reactions, that is, performing computations, simultaneously. Some advantages of DNA are that it is both self-complementary, allowing single strands to seek and find their own opposite sequences, and can easily be copied. Also, molecular biologists have already established a toolbox of DNA manipulations, including restriction enzyme cutting, ligation, sequencing, amplification, and fluorescent labelling, giving DNA a head start in the arena of non-silicon computing.

[1] This work has been partially supported by Spanish Project TIC2003-9319-c03-03 "Neural Networks and Networks of Evolutionary Processors".

Despite the complexity of this technology, the idea behind DNA computing springs from a simple analogy between the following two processes, one biological and one mathematical: the complex structure of a living organism ultimately derives from applying sets of simple instructed operations (such as copying, marking, joining, inserting, deleting, etc.) to information in a DNA sequence, and computation is the result of combining very simple basic arithmetic and logical operations.

The fact that the definition of gene implies the concept of a unit of minimum relative information as far as a functional unit and that it corresponds to the structural unit of basic molecular DNA and by association can be considered as the basic unit of mutation and of heredity, has taken it directly to trying to simulate genetic algorithms using DNA.

Overview of Genetic Algorithms

Genetic Algorithms are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. The basic concept of GA is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve a problem.

First pioneered by John Holland in the 60s [Holland, 1975], Genetic Algorithms has been widely studied, experimented and applied in many fields in engineering worlds. Not only does GAs provide an alternative method to solving problem, it consistently outperforms other traditional methods in most of the problems link. Many of the real world problems involved finding optimal parameters, which might prove difficult for traditional methods but ideal for GAs.

GAs are based on an analogy with the genetic structure and behaviour of chromosomes within a population of individuals using the following foundations:

- Individuals in a population compete for resources and mates.
- Those individuals most successful in each 'competition' will produce more offspring than those individuals that perform poorly.
- Genes from ‘good' individuals propagate throughout the population so that two good parents will sometimes produce offspring that are better than either parent.

Thus each successive generation will become more suited to their environment. After an initial population is randomly generated, the algorithm evolves the through three operators: selection which equates to survival of the fittest; crossover which represents mating between individuals; mutation which introduces random modifications.

Selection Operator: gives preference to better individuals, allowing them to pass on their genes to the next generation. The goodness of each individual depends on its fitness. Fitness may be determined by an objective function or by a subjective judgement.

Crossover Operator: prime distinguished factor of GA from other optimization techniques. Two individuals are chosen from the population using the selection operator. A crossover site along the bit strings is randomly chosen. The values of the two strings are exchanged up to this point. If S1=000000 and s2=111111 and the crossover point is 2 then S1'=110000 and s2'=001111. The two new offspring created from this mating are put into the next generation of the population. By recombining portions of good individuals, this process is likely to create even better individuals.

Mutation Operator: with some low probability, a portion of the new individuals will have some of their bits flipped. Its purpose is to maintain diversity within the population and inhibit premature convergence. Mutation alone induces a random walk through the search space. Mutation and selection (without crossover) create parallel, noise-tolerant, hill-climbing algorithms.

One of the basic arguments of the theory of evolution is that individuals that show similarities are related. Based on this principle, Holland observed that certain groups of individuals with particular similarities in some positions in their chains had good common properties whilst others were worse. Abstracting this idea Holland defines the concept of scheme (H) in one binarian coding with chains of length ℓ, thus;

$$H = h_{\ell-1} \dots h_0 \in \{0, 1, *\}^{\ell} \leftrightarrow H = \{ s_{\ell-1} \dots s_0 / h_j \neq * \rightarrow s_j = h_j \}$$

In other words, one scheme represents a certain subgroup of the population in which the individuals differentiate themselves at most in the position of the asterisks.

For example, the scheme H = 10 $*$ 01$*$ correspond with the chain group

{100010, 100011, 101010, 101011} At the same time any group of chains defines a scheme, suffise to consider the J-th Coordinate.

$$\begin{array}{l} \pi_j : \{ 0, 1 \}^{\ell} \rightarrow \{ 0, 1 \} \\ s_{\ell-1} \dots s_0 \quad |\rightarrow s_j \end{array}$$

And to define the scheme H thus;

$$h_j = \begin{cases} 0 \text{ if } \pi_j (H) = \{ 0 \} \\ 1 \text{ if } \pi_j (H) = \{ 1 \} \\ * \text{ if } \pi_j (H) = \{ 0, 1 \} \end{cases}$$

In fact, the group of chains that can be generated by crossing the elements of the group C is exactly H. For example if C = {001011, 011111} then each chain of H = 0 * 1 * 11 can be generated by crossing the elements of C, even 011011 y 001111, that were not initially in C.

Obviously in some schemes their elements show more likeness between themselves than in others. To quantify this idea there are two concepts; The order of a scheme which is the number of fixed alleles in the scheme and the length of definition which is the distance between the first and the last of the fixed alleles. For example if H = 00 * *1*, then o(H) = 3 y δ(H) = 4.

In essence, Holland's scheme theorem affirms that the algorithm drives the search for the optimum through certain subgroups of the population. In other words, explores the space of search through those areas that on average are more adequate.

Given that during the reproduction a chromosome is selected with a probability proportional to fitness

$$\frac{f (s)}{\sum_{s \in P(t)} f (s),}$$

where P(t) denotes the population t-th. Then if we start from a population of N elements the spected number of representatives of H in the following iteration is

$$E (n (H, t + 1)) = n (H, t). \ N. \ \frac{f (H, t)}{\sum_{s \in P(t)} f (s),}$$

where **E** denotes the operator hope, n(H, t) is the number of chains in the scheme

In the generation t-th and

$$f(H,t) = \frac{\sum_{s \in H} f(s)}{n(H,t)}$$

is the average value of the scheme in that generation. If we now consider the action of the operators of crossing and mutation the previous equation is transformed in:

$$E(n(H,t+1)) = n(H,t).\ \frac{f(H,t)}{\overline{f}}\ [1-\alpha(H)]$$

where:

$$\overline{f} = \frac{\sum_{s \in H} f(s)}{N}$$

Represents the average fitness of the population and **α(H)** depends of the structure of H and the probabilities of crossing and mutation, **pc** y **pm,** but not of **t**, Because:

$$\alpha(H) = p_c.\ \frac{\delta(H)}{\ell - 1}.\ (1-p_m)^{o(H)}$$

This expression, ignoring the terms of grade ≥ 2 in Newton's binomial, and because pm _ 0.01, turns into the final formula of the fundamental Theorem of genetic algorithms (or Theorem of schemes):

$$E(n(H,t+1)) = n(H,t).\ \frac{f(H,t)}{\overline{f}}\ [1-p_c.\ \frac{\delta(H)}{\ell - 1} - o(H).\ p_m]$$

Thus, we have that if H is a scheme with a fitness level greater tan the average of the population it is hoped to increase the number of chains with the structure of H in the following generation as long as a(alpha) is small. In other words, the principle of this theorem's can be interpreted saying that "short schemes of lower order with greater fitness than the average increase the number of representatives in the successive generations" This type of schemes that seem to play an important role in the way that GAs act, are known as building blocks.

It seems then tha by juxtaposing solid blocks of small size, increasingly better individuals could be built. This leads us to think that functions that can be defined utilising short schemes of lower order and high suitabilty would be easy to optimise for the Gas [Mitchell, 1994]

This affirmation, known as the hypothesis of the building blocks [Holland, 2000] seems very reasonable. In fact, GAs have been designed for various applications are empirical evidence that for different types of problems such hypothesis is correct.

DNA Simulation of Genetic Algorithms

The construction of a genetic algorithm for the resolution of an optimisation problem requires the definition of the genetic architecture. In this sense the election the manner of coding of the individuals represents an important point to obtain the correct solution.

The said coding must be done in a way that each chain allows the storage of information corresponding to an individual of a genetic algorithm. [González, 2002] Later on the bits will be represented independently of the position that they are in. [Lipton, 1995]

Given that the genetic composition of an individual constitutes a multyfactorial variable, the definition of the individual within a population will be dependent of the problem to be dealt with, that is, an individual has

a genome that must comply with a number of prerequisites subordinate to the problem, to be considered suitable within that population.

The process of representation of the genes as a minimum unit of information requires the analysing of the problem in question with the purpose of stipulating the number of bits assigned to the mentioned gene.

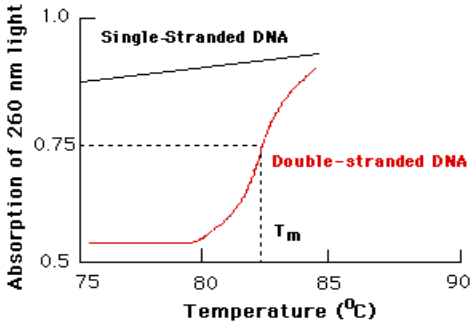
A gene will be represented as a whole of three fields; the percentage of Timine of the gene will be proportional to the fitness that represents the central field:

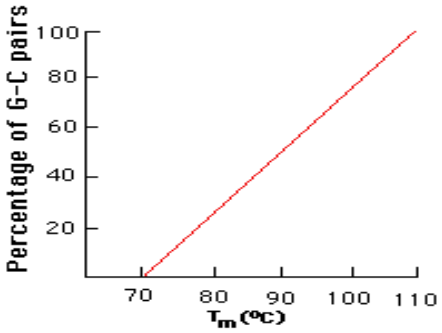
ENC(X) FITNESS xy ENC(Y)

With the purpose of mapping such fields for later processing the recombining DNA will bring a linker between fields in a way that can be recognised by restriction endonucleases. [Bonen, 1996]

Given that PCR will be used for the amplification of the DNA, each individual will carry at the same time, at both ends of the chain a sequence that will hybridise with a specific primer in the annealing phase (hybridization).

The fitness must be embedded in the coding of the individuals and given its definition will be determined by the content in G+C which implies that the fitness of an individual will be directly related with the fusion temperature and hence would be identifiable by spectophotometry (A_{260}) and separable by electrophoresis techniques (DGDE) [Macek 1997].





It is possible then to detect the perfect candidate by means of DGDE as it would be the one the possible candidates to present the greater number of GC pairs and therefore has the greater electrophotometric mobility.

The identification of the individuals of the population requires the tailing of the recombinant with a specific field that identifies the individual in a unique manner, the coding of this field will be done by means of the following function:

$$CODE: N \rightarrow \{G, C\}^*$$

Where N is the totallity of the natural numbers. It will receive a number corresponding to an individual within a population and returns a sequence of nucleotides.

The identification of the individuals in the mating zone requires again the inclusion of a field (Nm) which will be determined again by the function CODE.

The generation of the initial population has as an objective obtaining an aleatory population with a number of individulas equal to the size of the population. The complexity of the sinthesis of the sequence will be directly related to the number of genes used for the representation of the individuals within the genetic algorithm.

Basically it consists of a recombinant by means of a union of compatible fragments digested with restriction endonucleases.

The final format would look as follows:

PCR-primer Np REp XY RE0 XY RE1 …….. REn-1 XY REp Np -1 PCR-primer

The selection of individuals will be done by means of specific probes of the problem in question and the isolation of the individuals will be achieved by means of electrophoretic techniques (DGDE).

After selection, the individual will be introduced in the mating zone. For this he must be modified adding a specific field of such zone. In the event that a crossing of individuals is required, it is done in a temporal test tube containing the pair of individuals.

The mutation will be induced on each of the individual's results of the crossings operation in the genes in which the mutation frequency surpasses others obtained at ramdom and consists in the substitution of a gene by its complement in bases.

Later ill adapted individuals from the mating zone will be eliminated and will be substituted by the created recombinants. To determine the finalisation of the algorithm in each iteration the average of population adaptation is calculated. Once the convergence of the population is reached the best individual will be analysed by means of radioactive marking (o etching) techniques.

Fitness Computation on TSP Problem

The TSP is interesting not only from a theoretical point of view, many practical applications can be modelled as a travelling salesman problem or as variants of it, for example, pen movement of a plotter, drilling of printed circuit boards (PCB), real-world routing of school buses, airlines, delivery trucks and postal carriers. Researchers have tracked TSPs to study biomolecular pathways, to route a computer networks' parallel processing, to advance cryptography, to determine the order of thousands of exposures needed in X-ray crystallography and to determine routes searching for forest fires (which is a multiple-salesman problem partitioned into single TSPs). Therefore, there is a tremendous need for algorithms.

In the last two decades an enormous progress has been made with respect to solving travelling salesman problems to optimality which, of course, is the ultimate goal of every researcher. One of landmarks in the search for optimal solutions is a 3038-city problem. This progress is only party due to the increasing hardware power of computers. Above all, it was made possible by the development of mathematical theory and of efficient algorithms.

There are strong relations between the constraints of the problem, the representation adopted and the genetic operators that can be used with it. The goal of travelling Salesman Problem is to devise a travel plan (a tour) which minimises the total distance travelled. TSP is NP-hard (NP stands for non-deterministic polynomial time) - it is generally believed cannot be solved (exactly) in time polynomial.

TSP Solution using a DNA Genetic Algorithms Simulation. Fitness Computation.

Applying the previous protocol to the TSP of four cities in which the total size of the population is 256 (N) and the number of arches 6 (M) the individuals will be coded with an amount T inversily proportional to the length of the arches. The resulting sequence is shown in fig 1:

arch	Distance	Number of nucleotides	Fitness	Nucleotides of G
AB	1	40	1	40
BC	1	40	1	40
CD	4	40	4	10
BD	3	40	3	13
AD	2	40	2	20
AC	6	40	6	6

Figure 1.- Results of TSP simulation

The final format would look like this:

PCR-primer Np REp XY RE0 XY RE1 …….. REn-1 XY REp Np -1 PCR-primer

where XY ϵ { AB, BC, CD, AC, BD }

The selection criteria of the individuals in this case involves the encoding for an existing way to do this the strands of selection join two vertex of the graph neither initial nor final.

To be discarded are those individuals that do not start in the original city, those that do not finish in the final city as well as those that are found in repeated cities.

Selected are those structures which in the stretch O to N are covered by the strands of selection.

The format for the introduction of individuals in the mating zone is the following:

PCR-primer Np Nm REp XY RE0 XY RE1 …….. REn-1 XY REp Np -1 PCR-primer

where XY ϵ { AB, BC, CD, AC, BD }

We proceed then to the crossing mutation and evaluation of the degree of adaptation of the individuals prior to their introduction to the population and the process is repeated until the convergence of the population obtaining in each iteration the degree of adaptation

Conclusion

The generation of this work has produced a new approach to the simulation of genetic algorithms with DNA. The problem of fitness evaluation in a parallel form has been resolved satisfactorally. This does not imply that the definition would be independent of the problem at hand even though there are rules that facilitate such definitions and that can be solved by means of genetic algorithms.

In a GA simulated with DNA the concept fitness field disappears.

The coding of the individuals is closely related with the characteristics of electrophoretic migration.

The fitness of the individual is embedded in his coding and any attempt to add such field represents a grave error. The addition of such field would mean a personalisation of the individual thus preventing a massive and anonymous parallelism.

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.

[Amos, 2005] Martyn Amos.Theoretical and Experimental DNA Computation, Springer. ISBN 3-540-65773-8. 2005.

[Baum, 1996] Eric B. Baum, Dan Bohec. Running dynamic programming algorithms on a DNA computer. Proceedings of the second Annual Meeting on DNA Based Computers. 1996.

[Bonen, 1996] Dan Boneh, Christopher Dunworth, Richard J. Lipton, and Jiri Sgall. On the Computational Power of DNA. DAMATH: Discrete Applied Mathematics and Combinatorial Operations Research and Computer Science 71. 1996.

[Darwin C] Charles Darwin. On the Origin of the Species by the Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. Murray, London, 1859.

[González, 2002] Eduardo González Jiménez, Valery I. Poltev. La simulación computacional de procesos genéticos a nivel molecular. Ciencia y Cultura. Elementos, 479(47): September- November. 2002.

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

[Holland, 2000] Holland, J. H. (2000). Building blocks, cohort genetic algorithms, and hyperplane-defined functions. Evolutionary Computation, 8:4, 373-391.

[Kari, 2000] Lila Kari, Greg Gloor, Sheng Yu. Using DNA to solve the Bounded Post Correspondence Problem. Theoretical Computer Science 231 (2): 192–203. 2000.

[Lipton, 1995] Richard J.Lipton. Using DNA to solve NP-Complete Problems. Science, 268:542-545. April 1995. [Paun, 1998] Gheorge Paun, Grzegorz Rozenberg, Arto Salomaa. DNA Computing - New Computing Paradigms, Springer-Verlag. ISBN 3540641963. 1998.

[Mitchell, 1994] Mitchell, M., Holland, J. H., & Forrest, S. (1994). When will a genetic algorithm outperform hill climbing? IAdvances in Neural Information Processing Systems 6, 51-58, MIT Press.

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

Authors' Information

Maria Calvino – Dep. Inteligencia Artificial. Facultad de Informatica, Universidad Politecnica de Madrid, Boadilla del Monte, 28660 Madrid, Spain; e-mail: maria.calvino@medicimage.com

Nuria Gomez – Dep. Organizacion y Estructura de la Informacion, Escuela Universitaria de Informatica, Universidad Politecnica de Madrid, 28031 Madrid, Spain; e-mail: ngomez@dalum.eui.upm.es

Luis Fernando de Mingo Lopez – Dep. Organizacion y Estructura de la Informacion, Escuela Universitaria de Informatica, Universidad Politecnica de Madrid, 28031 Madrid, Spain; e-mail: lfmingo@eui.upm.es

ESTIMATING THE VOLUME FOR AREA FOREST INVENTORY WITH GROWING RADIAL BASIS NEURAL NETWORKS

Angel Castellanos, Ana Martinez Blanco, Valentin Palencia

Abstract*: This paper proposes a new method in order to compute clusters and centers for classification and approximation of wood volume using a set of data that can be easily obtained such as: height, radius, surface, surfaces used in cubical proofs and in the forest inventory built process. The proposed model, using radial basis function Neural Networks, achieves a rapid convergence dealing these tasks. This method is compared to other methods getting better results concerning the fewer training patterns; it also classifies the trees under a valid cluster set. Some figures are shown in order to better explain the learning procedure and results of this clustering process*

Keywords*: Neural Networks, Radial Basis Functions, Clustering, Forest Inventory.*

Introduction

Generally, to estimate wood volumes some standard formulas such as Huber's and others have been used. Because of simplicity and practicality the Huber's formula is frequently used to volume estimation. A new approach was developed to predict volume when there are a few data and when different species of trees are combined and it is necessary to obtain the volume of wood, using radial basis function Neural Networks.

The research community has developed several different neural network models, such as backpropagation, radial basis function, growing cell structures [Fritzke 1994] and self-organizing feature maps [Kohonen 1989]. A common characteristic of the aforementioned models is that they distinguish between learning and a performance phase. Neural networks with radial basis functions have proven to be an excellent tool in approximation with few patterns. Most relevant research in theory, design and applications of radial basis function neural networks is due

to Moody and Darken [Moody and Darken, 1989], Renals [Renals, 1989] and to Poggio and Girossi [Poggio and Girosi, 1990].

Radial basis function (RBF) neural networks provide a powerful alternative to multilayer perceptron (MLP) neural networks to approximate or to classify a pattern set. RBFs differ from MLPs in that the overall input-output map is constructed from local contributions of Gaussian axons, require fewer training samples and train faster than MLP. The most widely used method to estimate centers and widths consist on using an unsupervised technique called the k-nearest neighbour rule (see figure 1). The centers of the clusters give the centers of the RBFs and the distance between the clusters provides the width of the Gaussians. Computation of the centers, used in the kernels function of the RBF neural network, is being the main focus to study in order to achieve more efficient algorithms in the learning process of the pattern set. The choice of adequate centers implies a high performance, concerning the learning times, convergence and generalization.

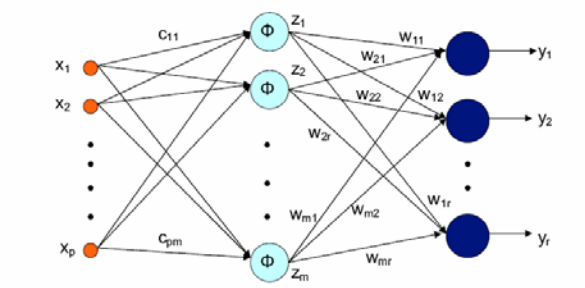


Figure 1.- Radial Basis Function Neural Network.

Problem Description

Volume parameter is one of the most important parameters in forest research when dealing with some forest inventories [Schreuder, H.T., Gregoire, T.G. and Word, G.B. 1993]. Usually, some trees are periodically cut in order to obtain such parameters using cubical proofs for each tree and for a given environment. This way, a repository is constructed to be able to compute the volume of wood for a given area and for a given tree specie in different environments. Stem volume is function of a tree's height, basal area, shape, etc. It is one of the most difficult parameters to measure, because an error in the measure or assumptions for any one of the above factors will propagate to the volume estimate. Volume is often measured for specific purposes, and interpretation of the volume estimate will depend on the units of measurement standards of use, and others specifications. Calculations of merchantable volume may also be based on true cubic volume. Direct and indirect methods for estimating volume are available [Hamilton, F. and Brack, C.L. 1999].

The method to estimate volumes used in forest are the tree volume tables or tree volume equations. Huber's volume equations are a very common equation used to estimating volume:

$$V = h\pi\left(\frac{d}{2}\right)^2$$ *V* denotes volume, *h* denotes length, *d* denotes diameter.

Another form of previous equation is:

$$V = \eta h\pi\left(\frac{d}{2}\right)^2$$ η = factor for the merchantable volume

Here, we proposed a study of the potential wood forest amount, that is, the maximum amount of wood that can be obtained. All data are taken from an inventory of the M-1019 area at "Elenco" in Madrid (Spain), at "Atazar" village. Most of the trees belongs to the Pinus Pinaster family and a small amount to the Pinus Sylvestris family. All this area is focused on the wood production. The area is divided into two different sub areas with a surface of 55.6 Ha and 46.7 Ha respectively. The main aim is to be able to forecast the wood volume and detect relationships between all the variables that are in our study. Variables taken into account are: normalized

diameter, total height, surface thickness, and radial growth in the last ten years. Normalized diameter has been measured in the whole feet of the two sub areas that made up the samples, provided they are larger than 12.5 cm till the last cluster of 60 cm. A parabolic regression analysis has been performed in order to obtain the cubical proofs to be compared to obtained results using neural networks.

Radial Basis Function Networks as Classifiers for Prediction in the Forest Products

A radial basis function neural network has been implemented with four input neurons: diameter, thickness surface, diameter and height in order to estimate the volume of wood that can be used. The net uses a competitive rule with full conscience in the hidden layer and one output layer with the Tanh function, all the learning process has been performed with the momentum algorithm. Unsupervised learning stage is based on 100 epochs and the supervised learning control uses as maximum epoch 1000, threshold 0.01. We have performed an initial study using 260 patterns in training set; after a 90 patterns in training set and finally with only 50 patterns in training set, and the error MSE, are similar in three cases. Problem under study is prediction of volume of wood, and it is compared to other methods such as the Huber's formula and the statistical regression analysis in order to estimate the amount of wood using typical tree variables: diameter, thickness and diameter growth. Neural networks had approximated in a good manner tested examples, getting a small mean squared error, see table below. Radial basis function neural network learns with only a few patterns, that is the way results using only 50 patterns are really excellent. For each of the tree species tested, the RBF gives less MSE estimated than the standard formulas Huber and Multivariate Analysis Regression.

	Error-Huber	**Error-RBF**	**error-regression multivariate**
MSE	0.05	0.007	0.01

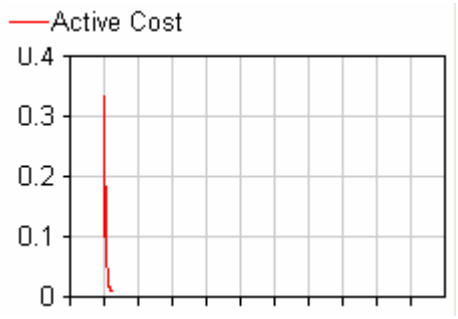
Next step consists on forecasting the input variable importance (sensitive analysis) in the learning process. Neural networks is a mapping $f(x_1, x_2, x_3, x_4) : \Re^4 \to \Re$ where $x_1 = diameter(cm)$, $x_2 = thickness(cm)$, $x_3 = growth\ of\ diameter(cm)$, $x_4 = height(cm)$ in order to forecast variable $x_5 = volume(dm^3)$. All center are stable in two points, that are those who signal the two main clusters, and that the net has been able to detect the two tree species.

Several matrixes have been computed; where columns are input variables to forecast and rows are hidden neurons. These matrixes show the center values. Variable $X_3 = diameter growth$ takes the same value in both centers what it means that the study can be done without such variable obtaining similar values of MSE. Main centers of RBF approximate real clusters in the two forest areas, following table shows the real clustering.

Zone - species	$\overline{d}_R$ (cm)	esp (cm)	$growing - d_R$ (cm)	**High** $\overline{a}_R$ (cm)
1	**19,49**	**5,28**	**3,19**	**6,45**
2	**33,71**	**7,38**	**3,91**	**10,66**

Previous table shows the matrix where the columns represent the input variable and the rows represent the hidden neurons. The hyperspace is divided into different regions or clusters starting from 16. Later, the number of clusters has been decreased till the minimum number of possible clusters is reached in order to solve the problem minimizing the mean squared error. Two main centers are found in the hyperspace, see following figures.

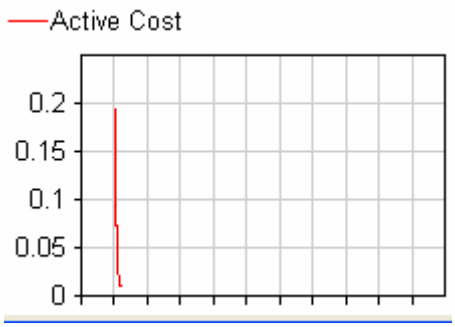
Four input variables and 16 clusters MSE=0.0079



Weights of unsupervisedSynapse

	0	1	2	3
0	0.036942655721	-0.238822595904	0.072954496902	0.093981749931
1	0.432889797662	-0.102832117679	-0.499664296396	-0.114246040223
2	0.095477156096	0.016403698043	0.159566026795	-0.156392101810
3	19.225327376276	5.288927609702	3.919955962094	6.339057415563
4	-0.168355357524	0.201620532853	-0.307824945830	0.172383800775
5	0.115833002716	-0.085833307901	-0.209097567675	0.372219000824
6	33.012244023029	7.304901050401	3.203351306184	10.479091408200
7	-0.233115634632	-0.324335459456	0.319177831355	0.477904599139
8	-0.361873226112	0.036851100192	0.336542863247	0.047074800867
9	0.295220001416	0.206065240573	-0.015004004102	-0.251121555223
10	-0.233970152898	-0.384365367595	0.357264931181	-0.219901120029
11	-0.488586077456	-0.010910367138	0.026932584613	-0.456450086978
12	0.475059059004	-0.305219805861	-0.256553049910	0.161805433515
13	-0.329340495010	-0.179799790570	-0.061021759697	-0.405140472549
14	-0.052446058535	0.357661671804	-0.064714499344	-0.358485671560
15	0.117053743095	-0.327082125309	0.262565996277	-0.391903439436

Four input variables and 12 clusters MSE=0.0078



Weights of unsupervisedSynapse

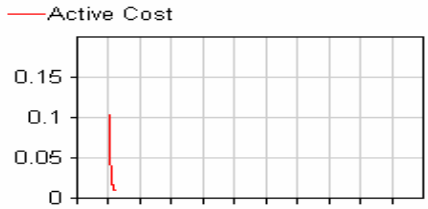
	0	1	2	3
0	0.085497604297	-0.495361186560	-0.428678243355	-0.275597399823
1	-0.271294289987	0.256492812891	-0.242240668966	0.380428479873
2	19.225327368018	5.288927687110	3.919955962077	6.339057412042
3	0.485595263527	0.020432142094	-0.446348460341	0.155293435469
4	0.402279732658	0.038407544176	-0.099566637165	-0.202536088137
5	-0.226248970000	-0.007187108982	-0.057847834712	-0.376430555132
6	0.443571275979	-0.466948454237	-0.310022278512	0.221640675069
7	0.282921842097	-0.360255745109	-0.080614642781	-0.386562700278
8	-0.318018127995	0.275719473861	-0.157948545793	0.363612781152
9	-0.152119510483	-0.382992034669	0.037705618458	-0.090838343455
10	33.012243953694	7.384981854488	3.203351386453	10.479891396617
11	-0.422513504440	0.383114108707	-0.434110538041	-0.229270302438

Four input variables and 8 clusters MSE=0.0075

Weights of unsupervisedSynapse

	0	1	2	3
0	0.139393292032	-0.039048432875	-0.491027558214	0.009964293344
1	0.022415845210	0.086809900204	0.144367809076	0.051072725608
2	0.098223822748	0.138782921842	-0.480559709464	0.225455488754
3	-0.001388592181	0.200430310984	-0.053514206366	0.407345805231
4	19.635513153942	5.354130101370	3.961877017807	6.443362612384
5	0.243858149968	-0.416776024659	-0.162892544328	0.434659871212
6	-0.461210974456	0.113635670034	0.494354075747	0.235343485824
7	33.531236499098	7.460325997065	3.093478779314	10.660911450110

Four input variables and 5 clusters MSE= 0.0073



Weights of unsupervisedSynapse

	0	1	2	3
0	0.225424970244	0.038163396100	-0.008224738304	-0.242484817042
1	-0.326044495987	-0.496215704825	0.115955076754	-0.043595690786
2	-0.041459395123	0.219168675802	-0.375911740471	0.385494552446
3	19.635514146112	5.354130753430	3.961878389731	6.443361133609
4	33.531214807424	7.460317864873	3.093467055318	10.660923756069

Four input variables and 4 clusters MSE=0.1

Weights of unsupervisedSynapse

	0	1	2	3
0	25.641225758815	6.267486735499	3.586643466294	8.269150221939
1	-0.469908749657	0.400753807184	-0.449949644459	0.216422009949
2	0.180684835353	-0.020706808679	0.264641254921	-0.400875881222
3	-0.177663502915	-0.422421948912	0.222281563768	-0.498290963469

Three input variables and 4 clusters MSE=0.0078

Weights of unsupervisedSynapse

	0	1	2
0	-0.497070223090	-0.439146092105	0.293725394452
1	33.371678352723	7.454034626552	10.546546589181
2	19.492469146051	5.318156886818	6.451743076843
3	-0.314081240272	0.053331095309	0.129078035829

Three input variables and 3 clusters MSE=0.104

Weights of unsupervisedSynapse

	0	1	2
0	-0.451506088443	0.463896603290	0.436979277932
1	-0.466643269143	0.177205725272	0.006576738792
2	25.641225758815	6.267486735499	8.269150221939

Two input variables and 4 MSE=0.0079

Weights of unsupervisedSynapse

	0	1
0	-0.219809564501	0.099536118656
1	19.351824656801	6.425514018775
2	33.205695811530	10.478871907419
3	-0.105517746513	-0.362300485244

Two input variables and 3 MSE=0.008

Weights of unsupervisedSynapse

	0	1
0	19.351820360090	6.425511846266
1	33.205662556926	10.478868449637
2	-0.358394116031	0.100421155431

All these tasks performed by RBF neural networks permits to classify the input patterns in the two main clusters belonging to the two tree species. Also, the variable representing the diameter growth is the less important one. If a neural network without such variable is implemented and with only 50 input patterns then results concerning the mean squared error are similar. When decreasing the number of input variables then the mean squared error increases, but forecasting results are still good if the importance of input variables is considered.

Results and Conclusions

A radial basis function neural network has been trained with a few patterns in order to forecast the volume of wood in a given forest area. The network performs a clustering process of the trees using different input variables. A sensitive analysis can be computed observing the weight of unsupervised synapse. To achieve a valid forecasting stage a previous classification must be performed. Let x_{ij} data corresponding to the matrix in the training process, where $i =$ *number of input variables* and $j =$ *number of hidden neurons,* then the following constraint must be verified $j \geq i+1$: it is needed that the number of hidden neurons must be greater than the number of input variables to perform a correct learning. A previous clustering process of input data permits a better forecasting process in the output variable, in our case the amount of volume of wood in a forest area. These results improve commercial and classical forecasting methods in forest inventories, and proposed method can be applied to any tree specie or forest area without taking into account environment variables that appears in classical mathematical equations. As the number of classes that needs to be discriminated decreases, classifier accuracy increases; until obtain the real number of classes. Once the correct number of classes has been obtained using the RBF and with a supervised learning the volume of wood for a forest inventory can be estimated.

Bibliography

[Fritzke 1994] Fritzke B. Supervised learning wiht growing cell structures. Advances in neural information processing system 6, Pp. 25. Morgan Kaufmann Publishers 1994.

[Hamilton, F. and Brack, C.L. 1999] Hamilton, F. and Brack, C.L. 1999. Stand volume estimates for modelling inventory data. Australian Forestry 62(4): 360 - 367

[Kphonen 1989] Kohonen T. Self organization and associative memory. Springer-Verlag 1989

[Moody and Darken, 1989] Moody J., Darken C. Fast learning in networks of locally-tunned processing units. Neural computation, vil1, pp 281-294. 1989.

[Poggio and Girosi, 1990] Poggio T. and Girossi F. Networks for approximation and learning. IN procceding of the IEEE, volume 78, pages 1481-1497.

[Schreuder, H.T., Gregoire, T.G. and Word, G.B. 1993] Schreuder, H.T., Gregoire, T.G. and Wood, G.B. 1993. Sampling Methods for Multiresource Forest Inventory. John Wiley and Sons, Inc. New York.

Authors' Information

Angel Castellanos – Dpto. de Ciencias Basicas aplicadas, a la Ingeniería Forestal. Escuela Univ. de Ingeniería Técnica Forestal. Universidad Politécnica de Madrid, Avda. de Ramiro de Maeztu s/n Madrid 28040, Spain; e-mail: acaste@forestales.upm.es

Ana Martinez Blanco – Dpto. de Ciencias Basicas aplicadas, a la Ingeniería Forestal. Escuela Univ. de Ingeniería Técnica Forestal. Universidad Politécnica de Madrid, Avda. de Ramiro de Maeztu s/n Madrid 28040, Spain; e-mail: amartinez@forestales.upm.es

Valentin Palencia – Dpto. de Arquitectura y Tecnología de Sistemas Informáticos. Facultad de Informática. Universidad Politécnica de Madrid, Campus de Montegancedo. Madrid 28660, Spain; e-mail: vpalencia@fi.upm.es

DECISION TREES FOR APPLICABILITY OF EVOLUTION RULES IN TRANSITION P SYSTEMS

Luis Fernandez, Fernando Arroyo, Ivan Garcia, 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 active evolution rules subset inside each membrane of the P system. But, to establish the active evolution rules subset, it is required the previous calculation of useful and applicable rules. Hence, computation of applicable evolution rules subset is critical for the whole evolution process efficiency, because it is performed in parallel inside each membrane in every evolution step. The work presented here shows advantages of incorporating decision trees in the evolution rules applicability algorithm. In order to it, necessary formalizations will be presented to consider this as a classification problem, the method to obtain the necessary decision tree automatically generated and the new algorithm for applicability based on it.*

***Keywords**: Decision Tree, ID3, Evolution Rules, Applicability, Transition P System.*

***ACM Classification Keywords**: I.2.6 Learning – Decision Tree; D.1.m Miscellaneous – Natural Computing*

Introduction

Membrane computing is a new computational model based on the membrane structure of living cells [Păun, 1998]. This model has become, during last years, a powerful framework for developing new ideas in theoretical computation. Main idea was settled in the base of connecting the Biology with Computer Science in order to develop new computational paradigms.

An overview of membrane computing software can be found in literature, or tentative for hardware implementations [Fernández, 2005], or even in local networks is enough “to understand how difficult is to implement membrane systems on digital devices” [Păun, 2005].

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 an evolution rules subset inside each membrane of the P system. Evolution rules subset we are studding here will be composed by applicable rules. Moreover, It exist algorithms of application for evolution rules [Fernández, 2006] that, recurrently to its end, need the computation of applicable evolution rules subset. Hence, computing applicable evolution rules is critical for the whole evolution process efficiency, because it is performed in parallel inside each membrane in each one of the evolution steps.

At the present time, computation of applicable evolution rules subset falls on redundancies in a directly or indirectly way. Incorporating decision trees in this computation avoids these redundancies and improves global efficiency of P system evolution.

This work is structured as follows: firstly, evolution rules applicability over a multiset of objects problem is formalized together with its corresponding traditional algorithm. Following section, briefly describes essential elements of decision trees. Afterwards, they are presented new formalizations that permit considering applicability problem as a classification problem solvable through decision trees. In next section, it is presented the algorithm based on decision trees. Finally, efficiency between both algorithms is compared and we expose our conclusions.

Applicability of Evolution Rules

This section defines concepts about multisets, evolution rules and applicability which are needed to follow the developed work presented here. Moreover, it is presented the traditional algorithm, without decision trees, for applicability evolution rules on multisets and its complexity.

From now on, let U be a finite and not empty set of symbols with $|U| = m$.

Let ω be a multiset over U, where ω is a mapping from U to N. Hence, $\omega(u) = p / \forall u \in U \; \exists! \, p \in N$. Let us present the set of all multisets as $\mathcal{M}(U) = \{\omega / \omega \; is \; a \; multiset\}$.

Weight of a symbol $u \in U$ is defined over a multiset $\omega \in \mathcal{M}(U)$ as $\omega(u)$ and it is represented by $|\omega|_u$.

Inclusion of multiset is a binary relation defined as $\omega_1 \subset \omega_2 \Leftrightarrow |\omega_1|_u \leq |\omega_2|_u$, $\forall u \in U \;\; \forall \omega_1, \omega_2 \in \mathcal{M}(U)$.

Any $\omega \in \mathcal{M}(U)$ can be represented as the m-tuple of natural number by the Parikh vector associated to the multiset w with respect to U. The problem is that the Parikh vector representation depends on the order of the elements of U. To avoid this problem, an order over the set U is defined as an ordered succession of symbols through a one to one mapping Φ from $\{1..m\}$ to U that is:

1. $\forall i \in \{1,...,m\} \; \exists u \in U / \Phi(i) = u$
2. $\forall u \in U \; \exists i \in \{1,...,m\} / \Phi(i) = u$
3. $\forall i, j \in \{1,...,m\} / \Phi(i) = \Phi(j) \Rightarrow i = j$

This fact permits us to represent every $\omega \in \mathcal{M}(U)$ as an element of N^m in a congruent manner. Hence,

$\omega = (p_1,..., p_m) \in N^m \; / \; |\omega|_u = p_{\Phi(u)} \;\; \forall u \in U$.

On the other hand, let T be a finite and non empty set of targets.

Evolution rule with symbols *in* U *and targets in* T *is defined by* $r = (a, c, \delta)$ where $a \in \mathcal{M}(U)$, $c \in \mathcal{M}(U \times T)$ and $\delta \in$ {dissolve, not dissolve}. The set of evolution rules is defined as $\mathcal{R}(U,T) = \{r / r \; is \; a \; evolution \; rule\}$.

Antecedent of $r = (a, c, \delta) \in \mathcal{R}(U,T)$ is defined as $input(r) = a$.

Finally, it is said that $r \in \mathcal{R}(U,T)$ is applicable over $\omega \in \mathcal{M}(U)$ if and only if $input(r) \subset \omega$.

Applicability Algorithm. On the one hand, a set of useful evolution rules R and a multiset of objects ω, will be the input to the process. On the other hand, output of process will be R_A, the evolution rules subset of R that are applicable over the multiset. Traditional algorithm [Fernández, 2005] checks weights of each evolution rules antecedent symbol with the corresponding from multiset of objects.

(1) $R_A \leftarrow \emptyset$
(2) **FOR-EACH** r_i **IN** R **DO BEGIN**
(3) $\quad j \leftarrow 1$
(4) $\quad$ **WHILE** $j \leq |\omega| - 1$ **AND** $|input(r_i)|_j \leq |\omega|_j$ **DO**
(5) $\quad\quad j \leftarrow j + 1$
(6) $\quad$ **IF** $|input(r_i)|_j \leq |\omega|_j$ **THEN**
(7) $\quad\quad R_A \leftarrow R_A \cup \{r_i\}$
(8) **END**

Algorithm 1. Evolution rules applicability (without decision trees).

Complexity of algorithm 1 consider, in the worst case, situation in which every evolution rule are applicable over the multiset of objects: loop in (4) will reach as many iterations as symbols exists in U on each iteration of loop (2) to each evolution rule present in R. In the worst case, complexity order will be $O(n)$ being $n = | R | \cdot | U |$

Analysis of previous algorithm will reveal possible redundancies in checks: in a direct and indirect way. So,

- A redundant check in a direct way will occur when weight of a same symbol is equal in more than one evolution rule antecedent, executing several times the same comparison (for example, let be $input(r_1) = (3, 1, 4, 1)$, $input(r_2) = (3, 2, 4, 4)$, and $\omega = (7, 3, 5, 4)$ where comparisons for the fist and third symbol of $input(r_2)$ are redundant in a direct way with its respective symbols in $input(r_1)$).
- A redundant check in an indirect way will occur when, after result of a checking which is false, it will be performed checks between greater weights of that symbol in others evolution rules antecedent (for instance, let it be $input(r_1) = (3, 1, 3, 1)$, $input(r_2) = (5, 2, 5, 1)$, and $\omega = (1, 3, 5, 4)$ where comparison for first symbol of $input(r_2)$ is redundant in an indirect way with its respective symbol in $input(r_1)$).

Furthermore, any checking of the weight of a symbol from an evolution rule antecedent with 0 will be unnecessary because $0 \leq n \ \ \forall n \in N$.

Decision Trees

A decision tree is a tree that permits us to determine the class which one element belongs to, depending on the values that take some attributes of it. Every internal node represents one attribute and edges are possible values of that attribute. Every leaf node in the tree represents one class. So, one unknown element can be classified processing the tree: every internal node studies the value of one attribute for the element and takes the appropriate edge, depending on its value; it continues until a leaf node is reached and, therefore, to the element classification.

There are a lot of algorithms to generate decision trees [Rasoul, 1991]. In particular, ID3 algorithm is based on entropy information and it generates an optimum decision tree from a non incremental set of instances and without repetitions.

ID3 algorithm requires as input (Fig. 1): let E be a finite set of instances $\{e_1,...,e_p\}$; let A be a finite set of attributes $\{a_1,...,a_q\}$;let V_j be a finite set of values $\{v_{1j},...,v_{rj}\}$ for each attribute a_j, (where a_j attribute value for instance e_i fulfils $v_{ij} \in V_j$); and finally, let C be a finite set of classes for the classification $\{c_1,...,c_s\}$. On the other hand, ID3 algorithm outputs the optimum decision tree for any element classification (Fig. 2).

E	a_1	...	a_j	...	a_q	C
e_1	v_{11}	...	v_{1j}	...	v_{1q}	c_1
...	...	...	...	...	...	...
e_i	v_{i1}	...	v_{ij}	...	v_{iq}	c_j
...	...	...	...	...	...	...
e_p	v_{p1}	...	v_{pj}	...	v_{pq}	c_s

Figure 1. Example of values table for ID3 algorithm input

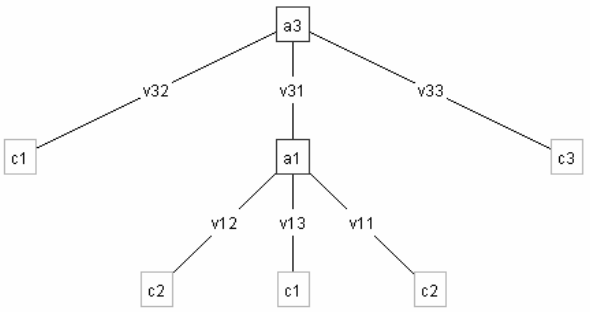


Figure 2. Decision tree generated by ID3 for values table from fig 1.

Decision Trees for Applicability of Evolution Rules

This section presents evolution rules applicability as a classification problem. This way, it will be posible to design a new algorithm that, being based on a decision tree, avoid direct and indirect redundancies of algorithm 1 presented above.

In order to it, we invert evolution rules applicability problem terms: for a given multiset, we compute the applicable evolution rules subset. Hence, we consider:

- Multisets of objects will be the elements to be classified: $\omega = (p_1,..., p_m) \in \mathcal{M}(U)$;
- The set of attributes will be a settled as a set of checks between the objects weights from the multiset and the same object from the evolution rules antecedents having a non null weight. Hence, the finite set of attributes will be: $A = \{a \equiv \mid \omega \mid_u \geq k \ / \mid input(r) \mid_u = k \wedge k \neq 0 \ \ \exists r \in R \ \forall u \in U \}$;
- Consequently, the finite set of values for every attribute will be true or false, result from comparison relationship between weights.
- Finally, classes to consider will be the different applicable evolution rules subsets. Therefore, the finite set of classes will be: $C = \{c \equiv R_A / \exists R_A \subset R \}$.

To obtain automatic generation of decision tree from ID3 algorithm, it will be necessary a non incremental and without repetitions battery of finite instances. In order to it, domain is defined as a set of multisets having the same values for all of their attributes. Consequently, each domain is characterized because every multiset responds to the same applicable evolution rules subset, that is, to the same class. Finally, examples battery will be formed by a representative from each domain.

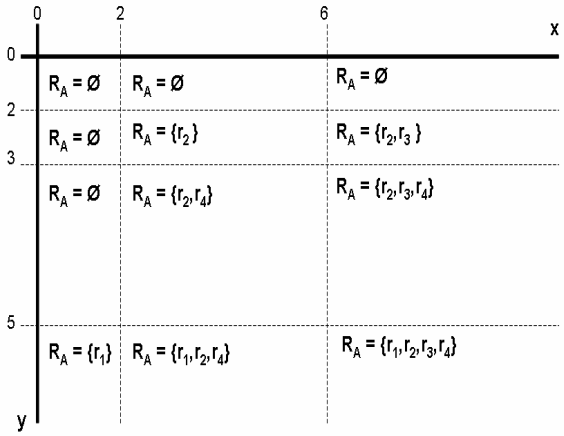


Fig. 3. Disjoint domains of multisets of objects for rules set and its corresponding applicable evolution rules.

Fig. 3 shows an example with disjoint domains of multisets of symbols for $U = \{x, y\}$ and rules set: $R = \{r_1, r_2, r_3, r_4\}$ where *their antecedents are:*

$r_1 = (y^5)$, $r_2 = (x^2, y^2)$, $r_3 = (x^6, y^2)$, $r_4 = (x^2, y^3)$.

Next, they are presented necessary definitions for formalizing the finite set of representative domains that are needed for the generation of decision trees.

It is defined projection of $u \in U$ over $R \subset \mathcal{P}(\mathcal{K}(U,T))$ as:

$$P_u(R) = \{n \in N / \exists r \in R \wedge \mid input\ r \mid_u = n\} \subset \mathcal{P}(N)$$

Hyperplane of $d \in \{1,...,m\}$ in $k \in N$ over N^m is defined as:

$$H_d^k(N^m) \rightarrow \{(x_1,...,x_d,...,x_m) / x_d = k\} \subset \mathcal{P}(N^m)$$

Thus, it is considered the grid over $R \subset \mathcal{P}(\mathcal{K}(U,T))$ as:

$$\mathcal{H}(R) = \{h / h = H^k_{\Phi(u)} \ \forall u \in U \wedge \forall k \in P_u(R)\}$$

Moreover, $\mathcal{D}(R)$ is defined as the partition N^m in disjoint subsets formed from every hyperplane of the grid $\mathcal{H}(R)$. It is named domain D to each one of the elements from partition $\mathcal{D}(R)$. Where it is fulfilled:

$$d = |\mathcal{D}(R)| = \prod_{\forall u \in U} |P_u(R)| < \infty$$

$$N^m = \bigcup_{k=1}^{d} D_k \ \wedge \ D_i \cap D_j \neq \emptyset, \forall i, j \in \{1,...,d\} \wedge i \neq j$$

Finally, it is defined representative of $D \in \mathcal{D}(R)$ as

$$\Lambda(D) = \min(\{dist(m,(0,...,0)) \ \forall m \in D\})$$

Fig. 4 shows an example obtained from values of table for the evolution rules set of figure 3. This table includes a row by each representative of the domains, its values for checking relations and applicable evolution rules subset that corresponds to each domain. Fig. 5 shows the classification tree generated by ID3 algorithm for the corresponding figure 4 values table.

Incorporation of decision trees avoids unnecessary null weights comparisons from algorithm 1 because they are not incorporated as in starting instances. Same, direct way redundancies are avoided, the weight of a symbol is compared with the same value just once. Finally, indirect way redundancies are also avoided due to the optimum decision tree ensured by ID3 algorithm, avoiding relations of transitive comparisons.

E	x≥6	x≥2	y≥5	y≥3	y≥2	C
$x^0 y^0$	no	no	no	no	no	Ø
$x^0 y^2$	no	no	no	no	yes	Ø
$x^0 y^3$	no	no	no	yes	yes	Ø
$x^0 y^5$	no	no	yes	yes	yes	$\{r_1\}$
$x^2 y^0$	no	yes	no	no	no	Ø
$x^2 y^2$	no	yes	no	no	yes	$\{r_2\}$
$x^2 y^3$	no	yes	no	yes	yes	$\{r_2,r_4\}$
$x^2 y^5$	no	yes	yes	yes	yes	$\{r_1,r_2,r_4\}$
$x^6 y^0$	yes	yes	no	no	no	Ø
$x^6 y^2$	yes	yes	no	no	yes	$\{r_2, r_3\}$
$x^6 y^3$	yes	yes	no	yes	yes	$\{r_2,r_3,r_4\}$
$x^6 y^5$	yes	yes	yes	yes	yes	$\{r_1,r_2,r_3,r_4\}$

Fig. 4. Examples battery for the evolution rules set of figure 3: in each row there is representative of each domain, values it takes for comparison relations and corresponding applicable evolution rules subset.

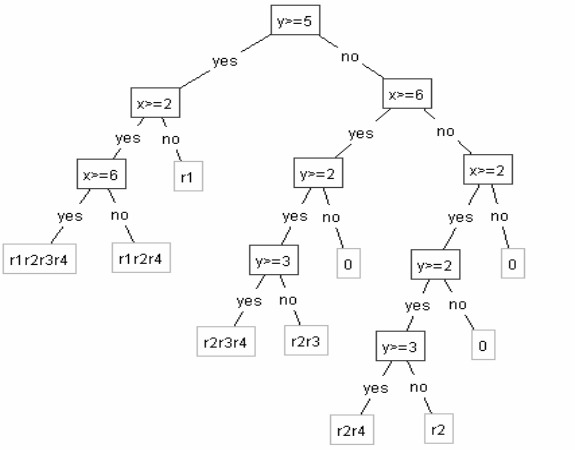


Fig. 5. Example of decision tree generated by ID3 algorithm for the examples battery from figure 4

Applicablity Algorithm based on Decision Trees

Previously to the algorithm presentation, we will expose the appropriate data structure for supporting the decision tree.

- On the one hand, they are disposed four correlative tables *left*, *symbol*, *value* and *right* for attribute nodes, with one cell in each table by each attribute node; root node is located in position 0 cells;
- On the other, It is disposed a table *classes* for classification nodes, with one cell for each classification node;
- Correlative cells of tables *symbol* and *value* determine which object weight from the multiset of objects has to be compared with which weight. Cells of tables *left* y *right* indicate, whether or not it is respectively accomplished previous relation comparison, which cell is the following attribute node in, whether index is positive; otherwise, indicate which cell of classification nodes table is the solution in.

Figure 6 shows an example of data structures of corresponding generated decision tree from figure 3.

Then, the input to applicability algorithm is ω , multisets of objects, and the supporting decision tree data structure: *left*, *symbol*, *value*, *right* and *classes*. On the other hand, output is A, an evolution rules subset of R that is applicable over that multiset. Following code processes rows of the indexes of branches *left* or *right*, depending on the comparison of symbol indicated by *symbol* weight with established value on *value* until it is reached a classification leaf, indicated by a negative value.

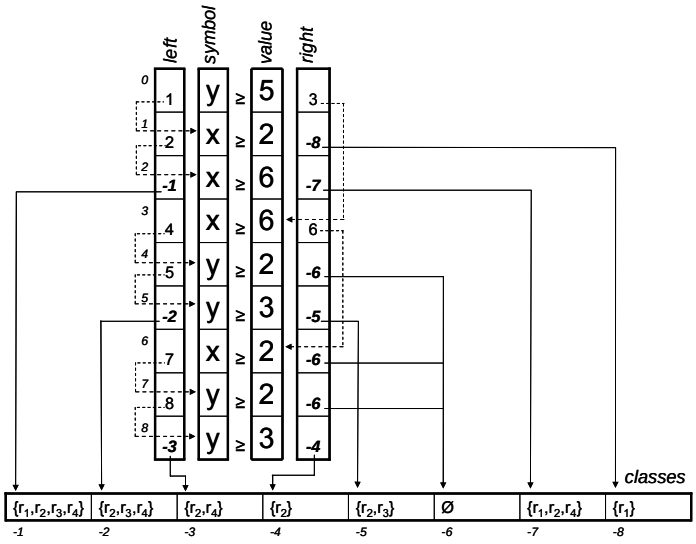


Fig. 6. Data structure for decision tree of figure 5.

(1) $f \leftarrow 0$
(2) **WHILE** $f \geq 0$ **DO**
(3) **IF** $|\omega|_{\Phi^{-1}(symbol[f])} \geq value[f]$ **THEN**
(4) $f \leftarrow left[f]$
(5) **ELSE**
(6) $f \leftarrow right[f]$
(7) $A \leftarrow applicable[f]$

Algorithm 2. Evolution rules applicability based on decision trees.

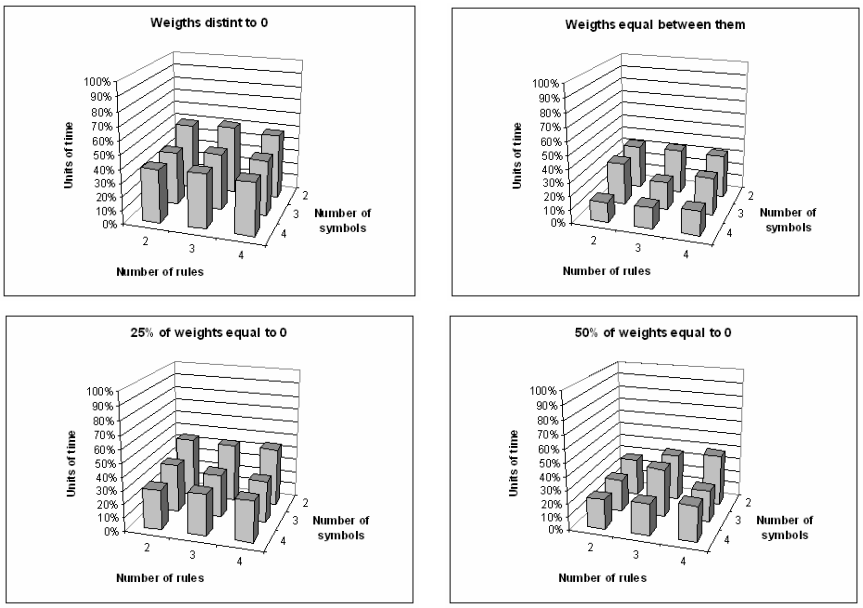


Fig. 7. Execution time reduction carried out by applicability algorithm based on decision tree respect traditional algorithm.

In the worst case, complexity of algorithm 2 considers to process the longest branch of the decision tree which length will always be lesser than $n = | R | \cdot | U |$. We will reach this conclusion by reduction to the absurd: in order to the longest branch of the decision tree requires n attribute nodes, It must be carried out the following a) every weight of each symbol in the antecedent of evolution rules is not null and different between them and, b) the d existing domain leads to applicable evolution rules different subsets. That is impossible because, in such circumstances, always exist more than one domain that would lead to the empty set. Specifically, a number of domains equal to $\left(\sum_{i=1}^{|U|} | P_i(R) | \right) - | U | + 1$.

Comparative

This section presents the experimental results obtained from evolution rules applicability using the two algorithms presented here. The test set has been randomly generated and it is composed by 48 different evolution rules sets (composed between 2 and 4 evolution rules composed between 2 and 4 symbols per antecedent), over these tests, it has been calculated the applicability of more than a million of randomly generated symbols multisets.

A first global analysis presents a reduction of execution time of this new algorithm based on decision trees respect to traditional algorithm in an average of 33%, with a variance of 7%.

Particularly, they has been made tests directed to four different situations to analyse the behaviour of new algorithm in extreme cases: with every different weight in antecedents of evolution rules, with every weight of same value, and with presence of 25% and 50% null weights.

In the worst case, with every weight being different between them, it has been reached at least 50% of execution time reduction. With all weights with same value, execution time is reduced to a 15% (for 4 evolution rules with 4 symbols per antecedent). In presence of 25% and 50% of null weights in antecedents of evolution rules, time is reduced to a 35% and a 29%, respectively, always in favor of the new algorithm with decision trees.

Conclusions

This work presents a new approach to the calculus of evolution rules applicability over a symbols multiset. This approach is based on decision trees generated from the set of evolution rules of a membrane. This way, they are avoided unnecessary and redundant checking in a direct or indirect way. Consequently, It is always obtained a lesser complexity than the corresponding traditional algorithm. So, execution time is optimized in the calculation of evolution rules applicability over a symbols multiset. All of this has repercussions in global efficiency of the P System evolution, because applicability calculation is carried out in parallel in each membrane in each evolution step.

Bibliography

[Fernández, 2005] L.Fernández, V.J.Martínez, F.Arroyo, L.F.Mingo, *A Hardware Circuit for Selecting Active Rules in Transition P Systems*, Workshop on Theory and Applications of P Systems. Timisoara (Rumanía), september, 2005.

[Fernández, 2006] L.Fernández, F.Arroyo, J.Castellanos, J.A.Tejedor, I.García, *New Algorithms for Application of Evolution Rules based on Applicability Benchmarks*, BIOCOMP06 International Conference on Bioinformatics and Computational Biology, Las Vegas (USA), july, 2006 (accepted).

[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.

[Păun, 2005] Gh.Păun, *Membrane computing. Basic ideas, results, applications*, Pre-Proceedings of First International Workshop on Theory and Application of P Systems, Timisoara, Romania, September 26-27, 2005, 1-8

[Rasoul, 1991] S.R.Safavian, D.Landgrebe, *A Survey of Decision Tree Classifier Methodology*, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 21, No. 3, pp 660-674, May 1991.

Authors' Information

Luis Fernandez – e-mail: setillo@eui.upm.es

Fernando Arroyo – e-mail: farroyo@eui.upm.es

Ivan Garcia – e-mail: igarcia@eui.upm.es

Gines Bravo – e-mail: gines@eui.upm.es

Natural Computing Group of Universidad Politécnica de Madrid (UPM); Ctra. Valencia, km. 7, 28031 Madrid, Spain.

ADVERGAMES: OVERVIEW

Eugenio Santos, Rafael Gonzalo, Francisco Gisbert

Abstract*: Advergame is a new marketing concept that has appeared due to the fact that young people are always connected to the Internet, are using mobile services such as SMS and MMS, or are chatting with instant messenger services and they spend too much time just playing in a stand alone way or in a network game. A new revolutionary service is the advergame one; that is a game with advertisment capabilities. Any company can develop an advergame that is, a game with some kind of advertising process of this company. This paper introduces some idea and concepts when developping an advergame..*

Keywords*: Advergames, Mobile Computing, Games Development.*

ACM Classification Keywords*: D.1.m Miscellaneous.*

Introduction

Online games are the future of the interactive entertainment industry, seeing the convergence between the traditional media, and entertainment industry, and the gaming industry in an effort to develop new and sustainable business models and revenue streams in an increasingly online world. They move the gaming industry into a more functionally rich online environment from which the majority of the revenue stream will come -- an e-business environment. But moving to this new model presents a number of challenges to the games developers, the players, and the service providers who ultimately will need to support this new environment.

However, it also presents a number of exciting opportunities for new business models, new markets, and new growth. The main problem faced is a solution integration issue. The player wants to pay for online content with their existing channels, but they also want security and privacy. The developers need cross-platform integration and support for multiple services, channels, and providers. The service providers need to build reusable business function that is robust, efficient, and generic -- it should work for all business models, not just the gaming industries.

Online games come in many forms. Perhaps the most recognized are the highly visual, action-oriented pop-ups familiar to NYTimes.com users. They're primarily used by advertisers for branding purposes and are generally delivered via pop-ups and in various other ad formats on third-party sites. The objective is to attract traffic and acquire new customers.

Instant-win promotions and contests requiring some level of consumer participation are increasingly popular. Again, their purpose goes beyond branding into acquisition and building databases of customers and prospects.

These games can take many forms, from a roulette-style wheel spun by the user to determine whether she's got a winning game card to an online drag race in which consumers challenge an automated car for a chance to win a related prize.

One marketer specializing in such online instant promotions had noteworthy results with incentive-based online games and contests.

It's hard to say, but most marketers' resistance to the game industry probably has to do with the pimply, geeky gamer stereotype. If we're not in an industry targeted to the teen market, we probably don't believe that games can have any impact on our marketing efforts. That's wrong, and more companies than ever out there are trying to change our collective marketing perceptions -- and convince us to start using games as advertising.

The concept's called advergaming, and regardless of its clunky moniker, it's a concept that has worked well for brands such as Nike (with its Nike Shox email game campaign), Ford (which used a racing game to promote its new Escape), and Pepsi. Companies such as YaYa, WildTangent, and The Groove Alliance have used stunning 3-D technology to create games that rival many commercial desktop and console games. Other companies, such as XI Interactive have brought together some of the finest minds in the gaming industry to create killer sports games. They combine single-player fun with innovative viral techniques that get consumers engaged with brands.

Even now, most of these games can be played over dial-up connections with middle-of-the-road computers. But with higher-speed computers and broadband connections becoming commonplace, these games are destined to become killer marketing vehicles for the future.

Simply put, games engage users for long periods of time, immersing them in an environment where they can develop an affinity for the brand. Rather than merely watching the action (as they do when viewing a sponsored sporting event on TV), advergame consumers actually become part of the action. Also, since the experience can be closely scripted in a near TV-like manner, the action can be interrupted to show TV-like commercials, or the views can be scripted to ensure advertiser messages are seen. It's a great combination of interactivity (for the user) and control (for the advertiser).

Some companies (such as Life Savers) create destination sites (check out Candystand) that host heavily brand-identified games. Others (such as Nike) have created effective viral campaigns in which users can play each other via email, inviting friends to beat their scores. Games have also been incorporated into banners and other rich media ad vehicles.

The market (and audience) for games is huge now and is going to continue to grow in the future as today's kids become tomorrow's sophisticated consumers. It's time to consider games as a viable marketing vehicle.

Advergames

The strict definition of an advergame is a Web-based computer game that incorporates advertising messages and images. However, we like to think of it as a tool that adds stickiness to your site as well as a little fun. Advergames allow you to market your product or brand subtly. Benefits of an Advergame are:

- Brand image reinforcement.
- Databases created from the advergame can be used for demographics research.
- Targeted markets can be reached by your advertising (when the game link is emailed).
- Visitors may spend more time on your site.
- Increased traffic due to viral marketing.

An advergame is not just for kids anymore - many surfers play advergames. These surfers include but are not limited to:

- 59\% of the boys ages 13 to 17 who go online.
- 62\% of the men ages 18 to 24 who go online.
- The largest group of women game players are between the ages of 45 and 54.

Like other advertising promotions not all advergames are equal. Here are some advices:

- Add a sweepstakes element to your game to provide an incentive for return visits or multiple plays.
- Include a viral component to encourage individuals to e-mail the game or its URL to friends to maximize word-of-mouth marketing.
- Consider a partnership with a company that can market your game to your desired audience. Simply building a great game and placing it on your website won't maximize impact.
- Make the game part of a larger media buy, such as by placing it within an interstitial and using buttons or banners on other sites to raise awareness and drive traffic.
- Take measures to reach your target audience during their leisure time. It's more likely they will play your game, and studies show people are more receptive to marketing messages when they are having fun

An advergame can help generate leads, build long-term brand awareness, and increase site stickiness as well as repeat web site visitors. It is a cost effective tactic that any sized e-business should at least check into for possible inclusion in marketing efforts.

Games in general are undeniably popular. Some of the more successful interactive games used for advertising purposes are parodies of the tried-and-true brand name games such as Jeopardy, Wheel Of Fortune, Who Wants To Be A Millionaire? and Roulette. In the end, what makes for a good interactive game for the player is a compelling, immersive experience.

What makes for a good advergame from the perspective of the advertiser footing the bill is the incorporation of a seamless data capture device. That data is then leveraged to build relationships. Even better are games in which in addition to data being captured, the player is learning about a product or service, its features and benefits, while playing the game.

An advergame can be used as part of an email campaign to a qualified, rented 3rd party opt-in list. The idea is to tease the recipient enough in the email to cause click-through to a special landing page. It is there at the landing page that the forging of an emotional bond with the prospect starts, and that the game is played.

In order to play, or to get to the exciting or more challenging levels of play, information is required of the prospect. At the very least, this includes name and email address. Some interactive games require additional data for each game level. The greater the prospect's desire to continue to play, the greater the likelihood the prospect will willingly provide additional data. It is critical that the prospect not be made to feel he has been brought to the game for the sole purpose of providing personal information. The prospect must feel he wants to be there of his own free will playing that game, or there will be no data at all forthcoming.

Another manner in which leads are generated is through a tell-a-friend viral device. Encouraging such forwarding can result in a viral factor, if you will, of 25\%. This means the amount of leads a game ultimately produces is that much greater than it would have been without the forwarding, or word-of-mouth.

It is the repetitive nature of games that drives up awareness of the sponsor. The same player, the frequency of the brand and/or the brand’s message increases, plays each time the game. The greater the frequency, the greater the chances of being remembered.

Research on the specifics of advergaming's ability to build brand awareness does not yet exist. Common sense will tell you that brand awareness does, indeed, increase, but to what degree isn't the point. Awareness should not be sole nor primary objective of an advergaming campaign; it is more of an added benefit to the advertiser. Advergaming's strength is in immediate, and for however long of a shelf life the game may have, lead generation.

Each time the game is played by the same player is also a repeat visit to the site. This is where web site navigation, page design, merchandising techniques, etc. come in to play. A sharp marketer will employ every technique possible to move the player to and through product when the player has completed the gaming session.

The agency with whom an advertiser works to create, implement and manage an advergaming effort will have a mechanism set up to track users (players), qualified leads, viral rate, even viral posting of the game on other venues, if any. The agency will also report on the advertiser's cost per thousand, cost per lead, cost per minute engaged, average engagement time, click-through rate, and conversion rate.

An even more cost effective means of employing advergaming as a marketing tactic is the use of an off-the-shelf game. For this type of game, the agency is re-purposing a game that already exists especially for an advertiser. It is amortizing its fixed production costs over several clients, allowing it to charge a greatly reduced fee for its use. This scenario allows for the simple changing of skins to meet the needs of a particular advertiser, and perhaps a few other adjustments along the lines of incorporating a logo into the game.

The major advantage to a custom interactive game is in the opportunity to tie in product or service features and benefits - even news - to the game play. The obvious disadvantages are extended production time versus re-purposing, and additional cost.

Advergame Architecture

There is a broad range of personal computing devices in the market. Personal computers are the most extended ones, but nowadays there are turning out smaller devices such as handhelds and mobiles phones. These three devices (PCs, handheld and mobile phones) are in a convergence process since they all have a real operating system with graphical capabilities and can be connected to Internet. All of them have an {\it Internet} browser (HTML, WAP and XHTML) and also a Java virtual machine, even SMS services. This is the main reason advergames should be developed using the Java technology or a markup language in order to obtain a portable solution that can be run in any platform (Windows, Linux or Symbian). Maybe simple advergames can make use of WML technology or SMS connecting to a server. Advergames must be implemented in an efficient way mainly due to the limited resources in mobile phones. Connection capabilities of all devices include HTTP and sockets, so client-server and P2P solutions can be used.

A number of different technologies, listed below, are used for games on mobile devices. Some games are programmed to run natively on a phone's chipset, installed on the phone at the factory, and shipped with it. New embedded games cannot be installed by the consumer, and they are becoming less prevalent. Short Message Service is used to deliver short text messages from one phone to another. Users typically pay about 10 cents per message. SMS games are played by sending a message to a phone number that corresponds to the game provider's server, which receives the messages, performs some processing, and returns a message to the player with the results.

Just about every phone shipped since 1999 includes a Wireless Application Protocol (WAP) browser. WAP is, in essence, a static browsing medium, much like a vastly simplified form of the Web, optimized for the small form factors and low bandwidth of mobile phones. WAP games are played by going to the game provider's URL (usually through a link on the carrier's portal), downloading and viewing one or more pages, making a menu selection or entering text, submitting that data to the server and then viewing more pages. Phones will continue to contain WAP browsers, and developers may find WAP useful to deliver more detailed help or rules to players than can be contained in a game application, since most games are still subject to strict memory limits.

Java 2 Micro Edition (J2ME) is a form of the Java language that is optimized for small devices such as mobile phones and PDAs. Nokia (and most other phone manufacturers) have made a strong commitment to Java phone deployment. Tens of millions of Java-enabled phones are already in consumers' hands. J2ME is limited by comparison to desktop Java, but it vastly improves the ability of mobile phones to support games. It allows far

better control over interface than either SMS or WAP, allows sprite animation, and can connect over the air network to a remote server. Because of its capabilities and the widespread and growing deployment of Java-enabled phones, it is a natural for mobile game development today.

It is important to take into account the programming language when developing an advergame. This software must be installed in many different devices with different operating systems and of course it must not be recoded everytime a new device appears in the market. This is the main reason to choose JAVA as the programming language since most of the today devices have a Java Virtual Machine:

- Mobile Phones based on Symbian Operating System have an embedded virtual machine and
- those non based on Symbian also have one, both of them using the J2ME architecture.
- HandHeld devices can install a software tool such as Jeode from Insignia or J9 VM from IBM.
- Personal computers usually have a java runtime-environment that can be obtain from java.sun.com for the Microsoft Windows or Linux platforms.

A developped advergame with the J2ME with the MIDP profile technology can be run in a mobile phone, in a handheld device and in a personal computer. But it the advergame is developped with Standard Java maybe only just a few mobile phones will be able to run it, and of course handhelds devices and personal computers will be able to run it.

The MIDP APIs are logically composed of high-level and low-level APIs. The APIs are designed for applications or services where the handset functions as the client device. The user gains access to applications and services that run on the handset through a network service provider. The high-level APIs are designed for applications where software portability across a range of handsets is desired. This is important if you are writing an application or service that a network service provider plans to deploy to a selected set of handsets. To achieve this portability, the APIs use a high level of abstraction. The trade-off is that the high-level APIs limit the amount of control the developer has over the human interface look and feel. The underlying implementation of the user interface APIs, which is accomplished by the handset manufacturer, is responsable for adapting the human interface to the device hardware and native user interface style.

Among the most interesting capabilities of MIDP are the following ones:

- Persistent storage in the device.
- Graphical libraries.
- Connection via sockets or http.
- Portability amond different devices.

J2ME is not the only interpreted language deployed on phones, but it is an industry standard backed by many manufacturers and therefore offers a large and growing installed base. Some proprietary interpreted languages have significant regional presence, including Qualcomm's Binary Runtime Environment for Wireless in North American and a standard called GVM supported by some Korean carriers. Games initially developed for the large J2ME installed base can be recoded in these proprietary languages if a sound business case presents itself.

A simple advergame can be also developped with the WML language. Only a WAP browser (Opera, EzWap, WinWap, Personal Internet Explorer, etc.) will be needed in onder to play it. The problem is that a connection to a server must be stablished and it is impossible to play off-line. Also WML has neither device storage capabilities nor graphics libraries in order to obtain good results.

MIDP 2.0 is backwards compatible with MIDP 1.0, hence it provides all functionality defined in the MIDP 1.0 specification. In addition it provides Over-The-Air (OTA) provisioning. This feature was left to OEMs to provide in the MIDP 1.0 specification.

An enhanced user interface has been defined, making applications more interactive and easier to use. Media support has been added through the Audio Building Block, giving developers the ability to add tones, tone

sequences and WAV files even if the Mobile Media API optional package is not available. Game developers now have access to a Game API providing a standard foundation for building games. This API takes advantage of native device graphic capabilities. MIDP 2.0 adds support for HTTPS, datagram, sockets, server sockets and serial port communication. Push architecture is introduced in MIDP 2.0. This makes it possible to activate a MIDlet when the device receives information from a server. Hence, developers may develop event driven applications utilizing carrier networks. End-to-end security is provided through the HTTPS standard. The ability to set up secure connections is a leap forward for MIDP programming. A wide range of application models require encryption of data and may now utilize the security model of MIDP 2.0 based on open standards.

There are many connection methods when accessing a server or client depending on the device or the programming language when running an advergame. It is clear that some information needs to be sent among players or among clients-servers in order to accomplish the advergame, keep information tracking of the users, and the flow advertisment control.

Short messages or multimedia ones are a useful tool when sending information, only telephony services are neccesary. Mobile telephones, PDA's with expansion cards or personal computers can be used to send messages. One of the main disanvantages is that a non real-time game can be developed, but it is a good method to keep track information about the users. When a message reaches the game server some infomation (telephone number, provider, user status, etc.) can be updated in a database in order to send push messages. Typical advergames that use SMS technology are:

- Test questions, when there are enough messages in the server a prize can be randomly delivered to any mobile phone number.
- Last minute offers, when there are some tickets (cinema, football, theater, etc.) that are not sold, they can be delivered to mobile phones with a previous registration or asking the server with a SMS.

Only a SMS server and a agreement with an SMS provider is needed to implement this service. There are many companies that are actually offering such service. Even a Java implementation could be coded but it is needed a special agreement with the SMS provider in order to deliver all the messages to a given IP. SMS is not a particularly good technology for games, because it is dependent on text entry by the user, and thus is, in essence, a command-line environment. It is also expensive for a game of any depth, since a mere 10 exchanges with the server will cost a user 1 dollar or more. Although the deployment of Multimedia Message Service (MMS) technology makes message-based games more appealing, this is still not a great gameplay environment.

A more complex service than SMS can be implemented if devices support HTTP connections. All new devices are capable of such services. This is an approach similar to WWW services in personal computers. A server is necessary in order to control all the information since the IP address is usually dinamic. Java Servlets technology is an usefull tool to implement such services.

This service can use WML and WMLScript to connect to a server or use J2ME to stablish HTTP connections. In any case schema is similar to normal web pages. A real-time advergame can be achieved with this method since it does not depend on the message delivering just as in the previous section. The information is sent in real-time. This is a flexible method since only a browser is needed (or a JVM). The main advantage is the graphical user interface. But no P2P communication can be carried out. Either version of WAP offers a friendlier interface than SMS, and is generally less expensive for consumers who pay for airtime only, rather than by the message. But it is a static browsing medium; little or no processing can be done on the phone itself, and all gameplay must be over the network, with all processing performed by a remote server.

Socket use gives J2ME developers the flexibility to develop all kinds of network applications for wireless devices. However, not every wireless manufacturer supports socket communication in MIDP devices, which means that wireless applications developed using sockets could be limited to certain wireless devices and are less likely to

be portable across different types of wireless networks. To use a socket, the sender and receiver that are communicating must first establish a connection between their sockets. One will be listening for a request for a connection, and the other will be asking for a connection. Once two sockets have been connected, they may be used for transmitting data in either direction. All today devices are using an embedded JVM, tipically supporting J2ME version 2.0 (with sockets). Main advantages of this implementation are:

- Sockets management.
- 2D and 3D graphical {\it APIs}.
- Persinstent storage on the client.

The use of sockets is usefull when dealing P2P services. The only problem is that the IP address is dinamically assigned to the client so a server is needed. This P2P service needs to send some information to an advergame server to keep track information of the players. This solution is the best one since with sockets all previous schemas can be implemented.

Conclusions

Advergame is a new marketing concept that brings users a way to interact with others and also to participate in quizs. User information can be update in a database in order to send push messages or does mailing while the client in playing some game. Some advertisements can appear in the game or even play with advertisements. The user can win prizes to keep his attention.

This paper has presented some tecnologies that can be used to develop an advergame. Java services are the best solution since it is a portable solution and all today devices have an embedded virtual machine.

Bibliography

[1] Blockdot (2001). Advergaming 101. Available online: http://www.blockdot.com/advergaming/stats.cfm.

[2] Chen, J., Ringel, M. (2001). Can advergaming be the future of interactive advertising? Fast Forward. Available online: http://www.kpe.com.

[3] March, T. (2001, Spring). How to bag the elusive human attention span. Digitrends. Available online:

http://www.digitrends.net/marketing/13639_16525.html.

[4] Pintak, L. (2001, May 23). It's not only a game: Advergaming set to become a billion dollar industry. Available online: http://www.turboads.com/richmedia news/2001rmn/rmn20010523.shtml.

[5] Rodgers, A. L. (2002, January). Game theory. Available online: http://www.fastcompany.com/build/build feature/yaya.html.

[6] YaYa (2002a). Why games? Available online: http://www.yaya.com/why/index why.html.

[7] YaYa (2002b). YaYa creates viral Internet games that build brands and drive revenue. YaYa online press kit. Available online: http://reports.yaya.com/presskit.pdf.

Authors' Information

Eugenio Santos Menendez. Dpto. Organización y Estructura de la Información. Escuela Universitaria de Informática de la Universidad Politécnica de Madrid, Ctra. Valencia, km. 7, 28031 Madrid (Spain); e-mail: esantos@eui.upm.es

Rafael Gonzalo Molina – Dpto. Inteligencia Artificial. Facultad de Informática de la Universidad Politécnica de Madrid; Boadilla del Monte, Madrid (Spain); e-mail: rgonzalo@fi.upm.es

Francisco Gisbert - Dpto. Lenguajes, Sistemas e Ingeniería del Software. Facultad de Informática de la Universidad Politécnica de Madrid; Boadilla del Monte, Madrid (Spain); e-mail: fgisbert@fi.upm.es

MODELING AND ANNOTATING THE EXPRESSIVE SEMANTICS OF DANCE VIDEOS

Balakrishnan Ramadoss, Kannan Rajkumar

Abstract: *Dance videos are interesting and semantics-intensive. At the same time, they are the complex type of videos compared to all other types such as sports, news and movie videos. In fact, dance video is the one which is less explored by the researchers across the globe. Dance videos exhibit rich semantics such as macro features and micro features and can be classified into several types. Hence, the conceptual modeling of the expressive semantics of the dance videos is very crucial and complex. This paper presents a generic Dance Video Semantics Model (DVSM) in order to represent the semantics of the dance videos at different granularity levels, identified by the components of the accompanying song. This model incorporates both syntactic and semantic features of the videos and introduces a new entity type called, Agent, to specify the micro features of the dance videos. The instantiations of the model are expressed as graphs. The model is implemented as a tool using J2SE and JMF to annotate the macro and micro features of the dance videos. Finally examples and evaluation results are provided to depict the effectiveness of the proposed dance video model.*

Keywords: *Agents, Dance videos, Macro features, Micro features, Video annotation, Video semantics*.

1. Introduction

Dance data is essentially multimedia by nature consisting of visual, audio and textual materials. Dance video modeling and mining depends significantly on our ability to recognize the relevant information in each of these data streams. One of the most challenging problems here is the modeling of the dance video semantics such that the relevant semantics are consistent with the perception of the real world.

The classical and folk dances are the real cultural wealth of a nation. In India, the most important classical dances are Bharathanatyam, Kadak, Kadakali, Kuchipudi and Manipuri (Saraswathi, 1994). Traditionally, dance learners perform dance steps by observing the natural language verbal descriptions and by emulating the steps of the choreographers. Therefore, the properly annotated dance videos will help the present and future generations to learn dance themselves and minimize the physical presence of the choreographers.

Notations are used everywhere and are most important for the dancers to communicate the ideas to the learners. They use graphical symbols such as vertical lines, horizontal lines, dots, triangle, rectangle etc, to denote body parts' actions on paper. Labanotation (Hutchinson, 1954) and Banesh (Ann, 1984) have been the frontier notational systems to record the dance movements or dance steps. Many western dances are using Labanotation to describe dance steps. However, many choreographers still follow the traditional way of training their students using natural language descriptions, because of the very few recording experts and inherent complexity of reading and understanding the symbols. Moreover, all Indian dances have unique structure and no common notational structure exists, apart from wire-frame stick diagram representing a dance step. Due to lack of notations, it is evident that the complexity of modeling the dance video semantics is relatively high.

Since the dance steps were archived in paper form and many classical dances lack notations, this kind of archival of dance becomes impossible even today. With the advances in digital technologies (Dorai, 2002) nowadays, magnetic tapes and disks record dance presentations efficiently. But, searching a dance sequence from these collections is not efficient, because of the huge volume of video data. The solution is to build a dance video information system so as to preserve and query the different dance semantics like, dance steps, beyond the spatio-temporal characteristics of the dancers and their dance steps.

The dance video database system requires an efficient video data model to abstract the semantics of the dance videos. To be more precise, the dance video data model should:

- abstract the different dance video semantics such as dancers, dance steps, agents (i.e., body parts of the dancers), posture, speed of dance steps, mood, music, beat, instrument used, background sceneries and the costume. More importantly, the spatio-temporal characteristics of the dancers must be incorporated in the model;
- capture the structure of the dance videos such as shot, scene and compound scene abstracting the different components of the accompanying song.

This paper addresses two related issues: modeling the semantics of the dance videos and annotating the dance steps from the real dance videos. The dance video semantics model represents the different types of dance semantics in a simple, efficient and flexible way. The annotation tool manually annotates the semantics (as macro and micro features) for further query processing and video mining. The main contributions of this paper are as follows:

- We propose a generic video data model to describe the dance steps as video events;
- We introduce the Actor entity in order to store the event specific roles of a video object. That is, an actor entity describes the context dependent role of the video object;
- We introduce the Agent entity to describe the context dependent action that is associated with the actor entity;
- We develop a tool that implements the dance video model in order to annotate the different dance semantics.

The rest of the paper is organized as follows: Section 2 presents some related works on video data models. Section 3 describes the different semantics of the dance videos. The DVSM for the dance video is introduced in Section 4. Section 5 illustrates the implementation of the DVSM using Java technologies. The proposed video model is evaluated against a set of conceptual and semantic quality factors in Section 6. Finally, Section 7 concludes the paper.

2. Related Work

Video data modeling is an important component of the dance video database system, as it abstracts the underlying semantics of the dance. This section briefly reviews some of the existing video modeling proposals and discusses the applicability to dance videos.

Colombo(1999) classifies the content-based search as semantic level search (e.g. objects, events and relationships) and low level search (e.g. color, texture and motion). They call the corresponding systems as first and second generation visual information systems. Several key word based techniques are applied to semantic search models, such as OVID (Oomoto, 1993), AVIS (Adali, 1996), Layered model (Koh, 1999) and Schema less semantic model (Al Safadi, 2000). Second generation systems provide automatic tools to extract low level features and subsequently semantic search is performed. Some of these systems include, but not limited to QBIC, Virage, VisualSEEK, VideoQ, VIOLONE, MARS, PhotoBook, ViBE, and PictHunter (Smeulders, 2000; Antani, 2002). However, these systems are either based on textual annotations or purely low level features, but not incorporating the other one.

In (Shu, 2000), Augmented Transition Network based semantic data model is proposed. The ATN models the video based on scenes, shots and key frames using strings as a sequence of characters. The string representation is used to model the spatial and temporal relationships of each object (moving and static) in a shot of the traffic video. Since the semantic features of dance videos are complex, the entire scene or shot cannot be abstracted in a single string.

Translucent markers, reflector costumes, special sensors and specialized cameras are used to capture and track human body parts' movements in some applications such as aerobics, traffic surveillance, sign language, news and sports videos (Vendrig, 2002). In order to record and analyze the dance steps of a dancer, based on this technique requires a special translucent markers or reflector costumes for the dancers. However, dancers do not

prefer to use these costumes as these costumes hide the dancer's make-ups and costumes. Moreover, these markers and reflectors prohibit the realism, affect dancer's comfort as well as reduce the focus or concentration of the dancers. Hence, automatic analysis of dance steps to extract the semantics of the dance steps is very complex.

Recently, the extended DISIMA(Lei Chen, 2003) model expresses events and concepts based on spatio-temporal relationships among salient objects. However, the required dance video database model has to consider not only salient objects, but all objects such as instruments, costumes, background and so on.

Event based syntactic-semantic video model (we call it as, ESSVM) (Ahmet, 2004) proposes Actor entity to specify the context dependent role of a player in soccer sports. This model represents the events such as free kick, goal, penalty etc, in which player assumes different roles such as scorer, assist-maker etc. But in dance videos, the contextual information of the dance events has to be described at multiple levels like actor and agent, rather than at a single granularity of actor entity.

COSMOS7 (Athanasios, 2005) models objects along with a set of events in which they participate, events along with a set of objects and temporal relationships between the objects. This model does not model the temporal relationships between events and the contextual roles. It models the events at a higher level only like speak, play, listen etc, whereas dance video model needs more detailed level of event representation such as agents, their action, speed of action, associated song and so on.

3. The Semantics of Dance Videos

Generally, dance information is dominated by visual content such as steps, posture and costume and the accompanying audio such as song and music. Hence, dance videos are rich in semantics and provide ample scope for the efficient semantic retrieval and dance video mining. This section illustrates the song that accompanies the dance performance, the different dance video types and the features of the dance videos in detail.

3.1. Song Granularity

Dance video contains several dance steps representing each song. In the case of classical dance, it is simply a collection of songs choreographed on the stage or theatre with a single start-stop (Cheng, 2003) camera operation. On the other hand, in a movie dance, a movie contains several songs and for each song dance steps are choreographed by the dancers. A song in a movie may be recorded with multiple start-stop camera operations. For instance, an Indian movie will normally contain about five to six songs. Here, each dance step may represent a step from any of the Indian dances or a new step innovated by the choreographer. Further, it includes the presentation aesthetics such as mood, feelings, emotion and so on.

Song is composed of four parts: Introduction, Additional Introduction, Chores and Stanzas (Web of Indian Classical Dances, 2003). Depending on the type of a song, Additional Introduction and Chores may be optional. Each part has few lines of lyrics for which dance steps are choreographed. In the dance video hierarchy, a shot represents a dance step, a scene represents dance steps of any of the song parts which are recorded in the same location and a video clip represents dance steps of a song. Our DVSM will represent the semantics of one dance step as a dance event. Dance step is the unit of analysis in this paper.

3.2. Features of Dance Videos

There are two types of dance video features- macro features and micro features and are annotated by the human annotators at macro and micro levels (Forouszan, 2004) accordingly. Macro features are general properties of the dance that are event independent and micro features are the properties of the dance step. That is, micro features are spatio-temporal characteristics of the dancers while rendering the dance steps. Micro features can also be called as event dependent features.

Macro features(or Bibliographic features): Date of recording, time of recording, geographic origin of the dance, geographic origin of the dancers, sex, age, number of dancers in a dance, type of performance venue (such as theatre, open-air, etc), type of the accompanying song, type of accompaniment, type of musical instrument used and types of dance videos. The different dance videos are movie dance video, theatre dance video, folk dance video, classical dance video, street dance video and festival dance video. These macro features are independent of the dance steps and are common to all dances.

Micro features (Dance step specific features): Spatio-temporal features classify dance movement behavior which include: movement of one dancer in relation to another dancer, movement of a specific body part (such as eye, leg etc. Refer Appendix-A for a complete list) of a dancer in relation to another part of the body, movement path of the dance (such as circular, linear, serpentine and zigzag), distance between body parts of a dancer while performing a dance step and distance between dancers.

Hence, the proposed video model has to characterize a set of macro features and micro features that exist in the dance videos.

4. The Dance Video Semantics Model

Conceptual model abstracts the dance video data into a structure for later querying its contents and mining some interesting patterns. For efficient conceptual modeling, one should know how choreographers demonstrate a dance to the learners. They are the experts in describing the rhythmic steps to the audience. This section presents a generic dance video model that efficiently describes the dance steps. Every dance step is called as an event and the model represents dance events by a set of micro features. The model is generic in the sense that it is applicable to any type of dance videos. DVSM is an extension of ER (Chen, 1976) with object oriented features. The goal of the model is to describe a dance step as an event.

The main entities of the model are events, objects that participate in these events, actor entities that describe contextual roles of objects in the events, agent entities that represent the action of the actor and concept entities that model the cognitive and affective features of the dancers.

For example, consider a dancer object with name, age, address and all other event independent attributes. The same dancer assumes different roles throughout the dance video. That is, he becomes hero in one dance step, lover in another dance step and so on. Roles are defined as attributes of Actors. Some other examples of actors are heroine, leader, follower, group dancer, friend etc. These context specific object roles form separate actor entities, which all refer to the same dancer object. Although one would say that actor performs the action in an event, finer granularity is necessary as far as dance videos are concerned. Therefore, contextual data of the dancers have to be described in two levels. A particular dance step is characterized by the actions of the agents who belong to the actors. Spatio-temporal characteristics are part of the actors as well as agents. Hence, they are described as attributes of actors and agents.

Apart from the dancer object, DVSM may also represent the ordinary objects with a standard UML class diagram. Some of them are: speed of the action of an agent, instrument used and the posture of the actor. The graph meta-schema of the DVSM is depicted in Figure1.

The graphical notations used in DVSM are described as follows. A rectangle node refers to an entity or an object. A round rectangle node refers to a concept. A dotted rectangle node denotes an actor entity. A thick rectangle node shows an agent object. Event entity is modeled with a trapezoid. Attributes of entities and relationships are represented with oval nodes. Relationships are denoted with directed lines on which the name of the relationship is denoted. Relationships without their names represent the containment type.

The model is instantiated as a directed acyclic graph. The reason for choosing graphs is that it elegantly models repetition of dance steps and has matured as a graph database. If a dance step repeats after some time, it just requires another edge to point to the same node. A graph is formally defined as follows: Let G = (V, E) be a

directed acyclic graph, where V denotes set of vertices and E denotes set of directed edges. The different entity classes, events, actors, agents, concepts, and other basic classes become vertices of the graph. Similarly, the set of interaction relationships will be denoted as directed edges of the graph.

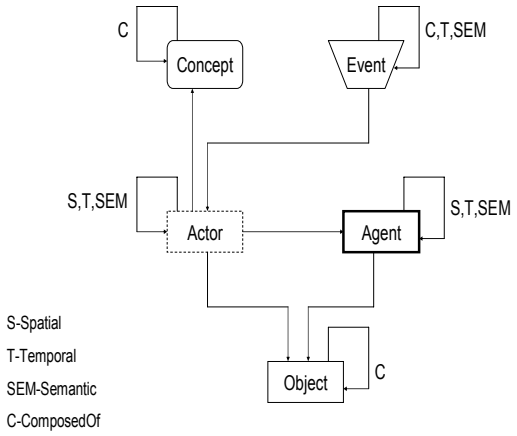


Figure1: Graphical representation of *DVSM*

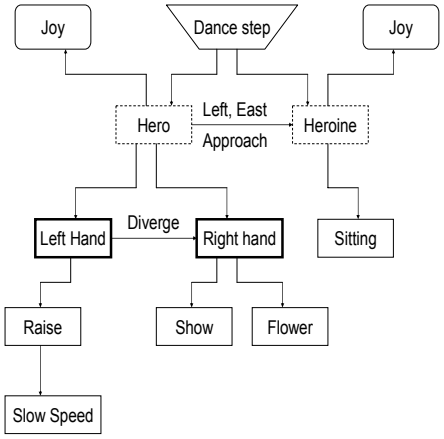


Figure2: Graph of dance step containing actors and agents.

The conceptual representation of an event highlighting a dance step as an instance of the graph, is depicted in Figure2. In this figure, the dance event consists of two actors whose roles are hero and heroine. Hero is standing left to the heroine initially and facing east. Heroine is sitting and facing west. Now hero approaches the heroine. These spatio-temporal semantics are stored as relations. Event independent characteristics of the actor are stored as video objects separately (not shown in the figure). The actors express joy and it is initialized as emotion. Hero raises his left hand to chest level with medium speed and displays a flower to the heroine with his right hand. Heroine remains idle without performing any action. This dance step is choreographed as part of one line of lyrics of a song. Due to overflowing of nodes, attribute nodes are not shown in this figure. The entity classes and relationships of the model are formally defined as shown below:

4.1. Event Entity Class

Dance step of a song is known as a dance event. For instance, consider a Bharathanatyam step Samathristy (Saraswathi, 1994). It is performed with the eyes by keeping them static without blinking. This step represents a thought, firmness, surprise or an image of an angel. Also, dance events can be combined to form a composite dance event. As an illustration, consider a dance step, Chandran. This step represents a moon and is a combination of two other steps: Pathagam and Lola pathmam and should be rendered concurrently. Pathagam is performed by keeping the thumb closed and the other four fingers straight and denotes clouds, air, sword and blessing. Similarly, Lola pathmam is performed by keeping all the fingers open and stretched and represents a sun. Composite dance event many represent events which are rendered concurrently or sequentially by a dancer.

Dance events are composed of actors, posture of the actors, cognitive state of the actors and the interactions in space and time between agents and actors. Formally, a dance event is described as a tuple,

$$\text{Event} = \{ \text{EID}, \text{D}, \text{AL}, \text{ND}, \text{ML} \}$$

where, EID is a unique identifier of the dance step, D is the description of the dance step, AL denotes the list of actors, ND is the number of dancers who are performing steps in the event and ML is the media locator of the video clip. Here, this Event tuple corresponds to Event object of Figure1.

4.2. Basic Entity Class

A dance video object refers to a meaningful semantic entity of a dance video database. It can be described using attributes which can represent macro and micro features. Formally, it is defined as shown below:

$$\text{Object} = \{ \text{OID}, V, \text{TY} \}$$

where OID is a unique object id, V = {a1:v1, ..., an:vn } is n event independent or dependent attribute value pairs and TY = { AID, AGID } denotes the dependency of the object, either actor or agent (to be defined later).

For example, hero is showing a flower in his right hand. Here, ShowFlower is the object in which V denotes attributes action and instrument with values show and flower respectively. TY holds ID of the agent, Right Hand which belongs to the actor Hero. In this case, V represents event dependent (i.e., dance step dependent) values. Similarly, object may also represent any of the macro features.

4.3. Actor Entity Class

Actor is a spatiotemporal entity in dance videos. So the existence time (Vijay, 2004) can be associated with the entity and it represents the life span of it. Actor is also a spatial entity. Therefore actor's displacement in space is modeled using Trajectory Points as in MPEG-7 (Martinez, 2003). Hence, actors are spatio-temporal entities playing context dependent roles in the events. Actors can have spatial, temporal and event specific semantic attributes describing their roles. The roles can be linguistic roles (Martinez, 2003) as in MPEG-7 or any semantic roles, such as loves.

The existence time predicate ΦACTOR, which is associated with the actor entity class, defines life span of the actor in terms of the existence time granularity (e.g. min and sec). ΦACTOR: S(ACTOR) × Z ⟶ B. This predicate takes a particular actor entity and a particular granule (denoted by an integer; say sec) and evaluates to a Boolean. If it is true, then that actor exists in the modeled reality at that granule (sec).

Constraint.1: Life span of an actor can exist only within the defined lifespan of the event to which it belong.

Formally, an actor entity can be described as follows:

$$\text{Actor} = \{ \text{AID}, \text{EID}, \text{DID}, R, L, T, P \}$$

where EID is the event id, DID is the corresponding dancer id, R denotes either semantic or linguistic roles of an actor, L is the existence time or lifespan, T represents the trajectory points(Point Set) as in Mpeg7 and P is the posture of the actor, which is a basic entity.

4.4. Agent Entity Class

Agent entity class represents the finer spatio-temporal semantics of the actions. The agent entity is the one which is most important in dance videos. The essence of a dance step is the actions done by the actors and it is the agent that performs the action. This is an exclusive feature of the dance videos. All other video types possess just one or two agents, which are fixed and do not play any significant role at all. For example, legs are agents in soccer sport videos, bat and ball are agents in cricket sport videos. Agent entity elegantly models the action of the agent which belongs to an actor. For instance, left eye and right eye of a heroine are agents. Formally, it is defined as:

$$\text{Agent} = \{ \text{AGID}, \text{AID}, \text{EID}, L, T, X, S, I \}$$

where AID and EID denote the actor id and event id respectively, X is the action agent performs, S denotes speed of X and I is the instrument held by the agent. Also, L and T depict the lifespan and spatial trajectory, similar to actor objects. Here, X, S and I are all basic entity types as defined earlier.

4.5. Concept Entity Class

The cognitive and affective content of an actor is modeled as a concept object. The concept is modeled as a separate entity type because of its ontological nature, thereby improving the semantic search. Formally, a concept entity can be defined as

Concept = { CID, AID, EID, T, D }

where T = { Emotion, Feeling, Mood } and D denote type of the concept and description as a string using natural language respectively.

4.6. Interaction Relationships

An interaction relationship relates members of an entity set to those of one or more entity sets. The DVSM employs the following set of relationships between the different entity sets. They are Composition (C), Spatial(S): which are topological (Egenhofer, 1994) and directional (Li, 1996), Temporal(T): Allen's interval algebra (Allan, 1983), Spatio-temporal, Motion(M): such as approach, diverge, stationary which are defined over the basic temporal relations (Athanasios, 2005), Semantic(SE) and Ontological(O).

The following section describes the set of relationships that occur between the various dance video entity sets.

4.6.1. Event Relationship

The relations between events are composition and temporal. Intuitively, a dance step of an actor may be followed by another actor immediately. Similarly, a dance step of an actor may be repeated by another dancer some time later. These follows and repeats relations are cues for later retrieval and mining operations. For example the query, find the set of dance steps done by a dancer, that is repeated by another dancer, can be processed by checking the life spans of the corresponding events.

Suppose E1, E2, ..., En are dance events participating in a temporal relationship. Let a1 and a2 be the actors, x1 and x2 be the actions of agents present in E1 and E2 respectively. Then, the predicate ΦREPEATS:S(X)×S(A)→E can take an actor and action and can return a set of events in which the action is performed. There is a constraint on the REPEATS predicate.

Constraint.2. Let LS1 and LS2 be the lifespan of E1 and E2 respectively. Then

$$(x1 = x2) \vee (LS1 < LS1) \Longrightarrow (E1 = E2)$$

Similarly, the other predicates such as performSameStep, performDifferentStep, and observe can be formulated, apart from follows and repeats predicates. Event relationships are formally defined as follows:

EE = { SRC, TAR, LST }

where SRC and TAR denote the source and target event ids and LST is the set of composition and temporal relationships which hold between source and target events.

4.6.2. Object Relationship

Objects can be composed of other objects. For example, consider Figure2 where hero holds a flower in his right hand. Here, flower is an example of an object. Formally, the relationship between objects can be represented similar to event relationships with a restriction that the SRC and TAR can be basic entities and LST will contain only composition relations.

4.6.3. Actor Relationship

Actor relationship represents the relationship between the roles of the objects, such as relation between hero and heroine who are dancer objects. Spatial, temporal and semantic relationships exist between the actors in a particular dance event. For instance, hero standing left to the heroine initially, may approach the heroine. This dance semantic contains spatial and motion relationships left and approach respectively. The actor relationship is formally defined as shown below:

AA = { AID1, AID2, O1, O2, LST }

where AID1 and AID2 are roles of the dancers O1 and O2 respectively and LST is now the set containing spatial, temporal and semantic relationships. Note that O1 and O2 are basic entity types.

4.6.4. Agent Relationship

Agent relationship is a second level semantic relation that describes the spatial and temporal characteristics of the agents. That is, agent relationship represents the finer semantics between the body parts of an actor. For instance, heroine is touching her left cheek with the index finger of her right hand. So, left cheek and right hand are the agents and finger can be the instrument used in the semantic relationship touch. Agent relationship is formally defined as,

$$AGAG = \{ AGID1, AGID2, AID, LST \}$$

where AGID1 and AGID2 are agentIDs of an actor AID and LST is similar to actor relationships.

4.6.5. Concept Relationship

Concept relationship is an ontological relationship (O) between concept entities. Typical ontological relationships (Guiness, 2004) are subClassOf, cardinality, intersection and union. This relationship is similar to event relationship with a modification that the source and target ids represent concepts and LST holds only ontological relations.

All other types of relationships between the different dance video entities are either semantic relationships or composition relationships such as partOf, composedOf, memberOf and so on. Table 1 summarizes the semantics of the kinds of relationships that exist between the dance video entities.

Table 1. Semantics of Relationships.

	Event	Object	Actor	Agent	Concept
Event	C,T,SE		C		C
Object		C	C	C	
Actor	C	C	S,T,SE	C	C
Agent		C	C	S,T,SE	
Concept	C		C		O

5. Implementation of DVSM

We have implemented the model in order to annotate the macro and micro features that are associated with the dance video. The tool has been developed using J2SE1.5 and JMF2.1.1 under Dell workstation. The tool is interactive as it minimizes the hard coding.

The dance video can be annotated by looking at the video clips that is running. Macro features can be annotated initially. The details of the dancers, musician, music, song, background, tempo, dance origin, context (whether live, rehearsal, professional play, competition etc), date and time of recording, type of performance venue and type of dance video are annotated. The screen shot depicting the rendering of the dance and interactive annotation of macro features is shown in Figures 3.

Then, micro features of every dance step of a song have to be annotated. The screen shot depicted in Figure 4 represents events, actors, agents and concepts. The annotator, by looking at the video, annotates the different information pertaining to these entity types in the order: event, actors of this event, agents of the actors, concepts revealed by the actors. But, one can swap the annotation of agents and concepts depending on his interest. The user interface has been carefully designed such a way that it minimizes the hard coding, as many of the graphical components will be populated automatically.

The second part of the micro features annotation involves the description of the various relationships between the entity types. For instance, event relationships, actor relationships, agent relationships and concept relationships describe the spatial, temporal, motion and semantic relations that exist between the entity types. The annotated data are stored in a backend database.

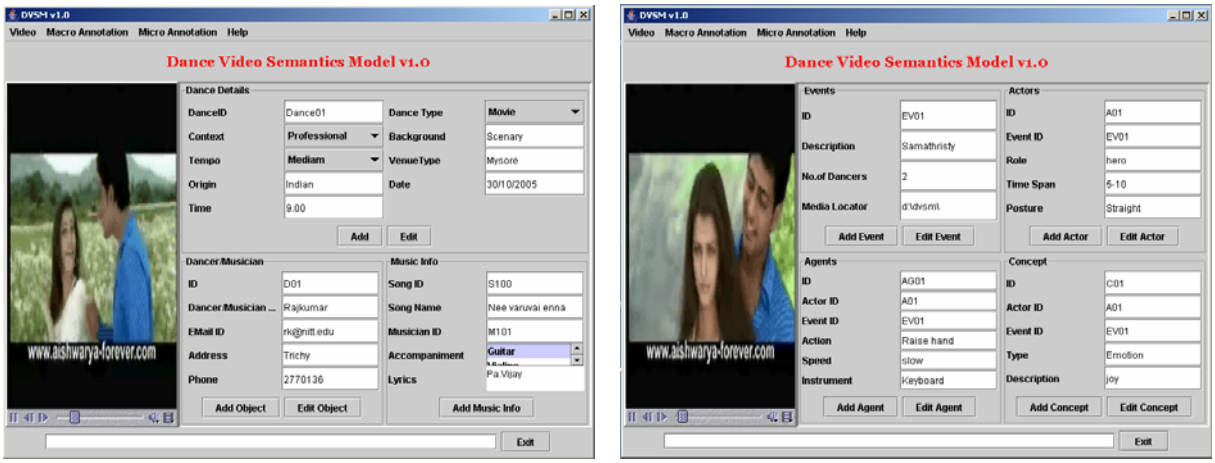


Fig.3 and 4. Screen shot of macro and micro annotation of objects.

6. Evaluation

Batini(1996) posits that conceptual model should possess the basic qualities: expressiveness, simplicity, minimality and formality. Additionally, Harry(2001) outlines other semantic qualities for video types. They are: explicit media structure(M), ability to describe objects (O), events (E), spatial relationships(S), temporal relationships(T) and integration of syntactic and semantic information (I). This paper introduces another factor, contextual description (Actor(A) and Agent(G)) to evaluate the proposed model.

The DVSM satisfies all the semantic quality requirements. Moreover, DVSM is unique in modeling the finer spatio-temporal contextual semantics of the events at finer granularity, with the help of agent entity type. Table 2 contrasts the existing semantic content based models against the semantic quality factors. The table illustrates that some models lack semantic features, some lack syntactic features and only few models integrate both syntactic and semantic relationships. Some applications, like soccer sports video, require the model to represent the contextual features of objects (called actors). The ESSVM proposed by Ahmet et al, describes contextual information at actor level. However, dance videos require contextual description at multiple granularities (called agent), beyond the actor level. Our proposed semantic model possesses both contextual abstractions-actor and agent, apart from the other semantic qualities. Hence, with the agent based approach, the paper claims to have achieved conceptual, semantic and contextual qualities in dance video data modeling.

Table 2. Comparison of semantic video models.

Legend: M-Media structure, O-Object, E-Event, S-Spatial, T-Temporal, I-Syntactic, semantic information, A-Actor, G-Agent.

Model	Semantic Qualities							
	M	O	E	S	T	I	A	G
AVIS	×	×		×	×	×		
OVID	×	×		×	×	×		
QBIC	×	×		×				
DISIMA	×	×	×	×	×	×		
COSMOS7		×	×	×	×	×		
ATN	×	×		×	×	×		
ESSVM	×	×	×	×	×	×	×	
DVSM	×	×	×	×	×	×	×	×

7. Conclusion

Data semantics provides a connection from a database to the real world outside the database and the conceptual model provides a mechanism to capture the data semantics (Vijay, 2004). The task of conceptual modeling is crucial and important, because of the vast amount of semantics that exist in multimedia applications. In particular, dance videos possess several interesting semantics for modeling and mining. This paper described as agent based approach for elicitation of the semantics such as macro and micro features of the dance videos. An interactive annotation tool has been developed based on the DVSM for annotating the dance video semantics at syntactic, semantic and contextual levels. Since dance steps are annotated manually, it is somewhat tedious to annotate dances by the dance expert.

Further work would be useful in many areas. It would be interesting to explore how DVSM can be used as a video model for exact and approximate query processing. As MPEG-7 is used to document the video semantics recently, it is valuable to employ MPEG-7 for representing dance semantics to enable better interoperability. Finally, it will be useful to explore how video mining techniques can be applied to dance videos.

Acknowledgements

This work is supported fully by the University Grants Commission (UGC), Government of India grant XTFTNBD065. The authors thank the editors and the anonymous reviewers for their insightful comments.

Bibliography

Adali, S, Candan, K.S, Chen, S.S, Erol,K & Subramanian, V.S.(1996). The Advanced Video Information System: Database structures and query processing. Multimedia Systems. 4, 172-186.

Ahmet Ekin, Murat Tekalp, A & Rajiv Mahrotra.(2004). Integrated semantic-syntactic video event modeling for search and browsing. IEEE Transaction on Multimedia, 6(6), 839-851.

Allen, J.F.(1983). Maintaining knowledge about temporal intervals. Communication of ACM, 26(11), 832-843.

Al Safadi, L.A.E & Getta, J.R.(2000). Semantic modeling for video content based retrieval systems. 23rd Austral Asian Computer Science conference, 2-9.

Ann Hutchinson, G.(1984). Dance Notation: Process of recording movement. London: Dance Books.

Antani, S, Kasturi, R & Jain, R.(2002). A survey on the use of Pattern recognition methods for abstraction, indexing and retrieval of images and video. Pattern Recognition, 35(4), 945-965.

Athanasios.C, Gkoritsas, & Marios.C.Angelides.(2005). COSMOS-7: A video content modeling framework for MPEG-7. MMM05, 123-130.

Batini, C, Ceri, S & Navathe, S.B.(1992). Conceptual database design: An entity relationship approach, Benjamin/Cummings publishing company.

Chen, P.P.(1976). The Entity-Relationship model-Towards a unified view of data. ACM Transaction on Database Systems, 1(1), 9-36.

Cheng Youg & XU.De.(2003). Hierarchical semantic associative video model. IEEE conference NNSP'03, 1217-1220.

Colombo, C, Del,A, Bimbo & Pala, P.(1999). Semantics in visual information retrieval. IEEE Multimedia, 6(3),.38-53.

Dorai,C, Manthe,A, Nack,F, Rutledge,L, Sikora,T & Zettl,H.(2002). Media Semantics: Who needs it and why?. ACM conf. Multimedia, 580-583.

Egenhofer, M & Franzosa, R.(1994). Point-set topological spatial relations, Conference of GIS, 5(2),161-174.

Forouzan Golshani, Pegge Vissicaro & Yound Choon Park.(2004). A multimedia information repository for cross cultural dance studies. Multimedia Tools and Applications, 24, 89-103.

Guiness, D.M, Van, F & Harmelen(2004). OWL: Web Ontology Language Overview. W3C Recommentation, http:// www.w3.org/tr/owl-features/.

Harry.W.Agius & Marios.C.Angelides(2001). Modeling content for sematic level querying of multimedia. Multimedia Tools and Applications, 15, 5-37.

Hutchinson, A.(1954).Labanotation: A system for recording movement. London: T Art Books.

Koh, J.L, Lee,C.S &.Chen, A.L.P.(1999). Semantic video model for content based retrieval. IEEE Conference MMCS'99, 2, 472-478.

Lei Chen & Tamer Ozsu, M.(2003). Modeling video data for content based queries: Extending DISIMA image data model. MMM'03, 169-189.

Li, J.Z, Ozsu, M.T & Szafron, D.(1996). Modeling of video spatial relationships in a object oriented database management systems. IEEE Workshop on MMDMS, 124.

Martinez, J.M.(2003). MPEG-7 Overview, ISO/IEC JTC1/SC29/WG11-N4980.

Oomoto,E,& Tanaka, K.(1993). OVID: Design and implementation of a video object database system. IEEE Transaction on. Knowledge and Data Engineering, 5(4), 629-643.

Saraswathi.(1994). Bharatha Nattiya Kalai.Chennai: Thirumagal Nilayam.

Shu Ching Chen, Mei Ling Shyu, R.L.Kashyap.(2000). ATN as a semantic model for video data. Journal of Network Information Management, 3(1), 9-25.

Smeulders, A.W.M, Worring, M, Santini, S, Gupta, A &.Jain, R.(2000). Content based image retrieval at the end of early days. IEEE Transaction on. Pattern Analysis and Machine Intelligence. 22(12),.1349-1380.

Vendrig, J.W.(2002). Interactive adaptive movie annotation, Multimedia and Expo conference, 1, 93-96.

Vijay Khatri, Sudha Ram & Richard.T.Snodgrass.(2004). Augmenting a conceptual model with geospatiotemporal annotations. IEEE Transaction on Knowledge and Data Engineering, 16(11), 1324-1338.

Web of Indian Classical Dances, http:// www.narthaki.com/index.htm.

Appendix-A: List of Agents

Head	Hand	Knee	Leg	Foot	Arm
Finger	Ankle	Elbow	Heel	Lower Leg	Wrist
Toe	Hip	Shoulder	Waist	Back	Torso
Forearm	Palm	Pelvis	Thigh	Ball of Foot	Chest

Authors' Information

Balakrishnan Ramadoss – Department of Computer Applications, National Institute of Technology, Trichy 620015, India; e-mail: brama@nitt.edu.

Kannan Rajkumar – Department of Computer Applications, National Institute of Technology, Trichy 620015, India; e-mail: cak0303@nitt.edu.

AUTOMATED PROBLEM DOMAIN COGNITION PROCESS IN INFORMATION SYSTEMS DESIGN

Maxim Loginov, Alexander Mikov

Abstract*: An automated cognitive approach for the design of Information Systems is presented. It is supposed to be used at the very beginning of the design process, between the stages of requirements determination and analysis, including the stage of analysis. In the context of the approach used either UML or ERD notations may be used for model representation. The approach provides the opportunity of using natural language text documents as a source of knowledge for automated problem domain model generation. It also simplifies the process of modelling by assisting the human user during the whole period of working upon the model (using UML or ERD notations).*

Keywords*: intellectual modeling technologies, information system development, structural analysis of loosely structured natural language documents.*

ACM Classification Keywords*: I.2 Artificial Intelligence: I.2.7 Natural Language Processing – Text analysis*

Introduction

The term “Problem domain” is usually used when the problem of Information Systems (IS) design is discussed. This term represents the aggregation of knowledge about objects and active subjects, tied together with specific relations and pertaining to some common tasks.

Usually the scope of the problem domain is not described strictly and different problem domains intersect. Let us take two problem domains for example: a school education service and a public health service.

An information system designed for automating reporting at schools and another one designed for decision-making for health authorities of a city council can not be completely independent. There are medical consulting rooms at schools and the rate of sickness certainly depends on the school working conditions and so on. After all, both information systems share some personality information: many people are citizens and students at the same time.

Nevertheless, a description (a model) of the problem domain is a very important part of an information system project. But, anyway, if this model is not comprehensive then it is incomplete.

Documents and experts usually play a part of the knowledge sources circumscribing the problem domain. There are several types of documents that may be used: legal documents, ones that describe business processes, databases of employees and customers, etc. Human experts may provide information on informal rules, conventions, relative importance of concepts, etc, in the given problem domain. Documents of listed types denote objects and formalize some relations in the problem domain concerned. To a first approximation they may be considered as local models of these relations.

The difficulty is that most local models are built using different approaches, because there is no unified approach that may be applied to a problem domain (excepting some narrow-ranged technical domains, where local models can be combined together into a global model using some strict mathematical rules; information systems built upon such problem domains are called “systems of automatic control”).

We are concerned here about information systems of a different kind – systems where the human element is of primary importance. Investigation into such kinds of problem domains is a type of empirical research, related to the “sciences of the artificial”.

Nowadays most CASE tools (Computer-Aided Software Engineering tools) can automatically build source program code for a projected information system, using some initial formal model of the problem domain (usually the model is represented as a framework, or graph). The urgent problem is to automate the process of building the formal model, e. g. to automate the process of cognition in the given problem domain.

Goals of the Research

The main purpose of this research is development of the special cognitive approach, referred to a class of Intellectual Modelling Technologies (IMT). This approach is designed for automating the process of information system development. Attention is focused on the very early stage of project development, the stage of analysis.

The problem domain of the class of IS under consideration includes a very large amount of legal documents (articles, assizes, bans, etc.), which regulate the status of objects, the behaviour of subjects related to an institution, etc. It also includes a settled system of document circulation. All this information, as a rule, is poorly structured. So, the development of a conceptual model of the problem domain (by means of UML language, for example) using knowledge from documents of these types, is a very difficult task and usually is done manually.

The suggested cognitive approach is aimed toward the problem of automating the conceptual-level model development by using loosely structured natural language documents.

Since the problem under consideration refers to a class of logical lexical systematization problems (as an example from the adjacent area of study we may take translation of natural language text into the language of

predicate logic), it has no solution using only a computation system. That is why the suggested approach is developed to work in conjunction with the human user. Human interference is needed during the automated analysis of problem domain described in source documents. Nevertheless, some self-learning capabilities in the context of approach allow us to depend on the self-development of the analyzer during persistent dialogue with the human user, so that subsequently it could be able to solve similar tasks without direct human assistance.

The suggested approach is oriented to be utilized at the earliest stage of the information system project development process. As indicated on fig. 1, the suggested approach is supposed to be used at the very beginning of the spiral loop, at the boundary between the stages of requirements determination and analysis, also including the stage of analysis.

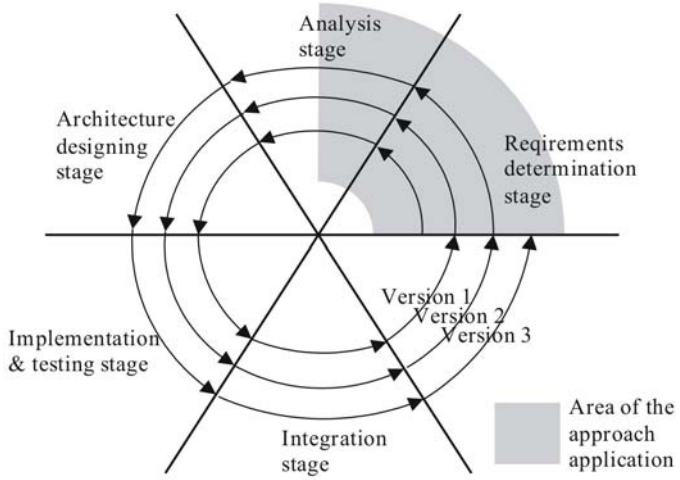


Figure 1. The Spiral Model of Software Life Cycle

It is important to mention that most existing IMT methods, used in CASE systems, automate, in general, stages of projecting, implementing, testing and integrating, but never touch the stages of requirements determination and analysis. Transition from the stage of requirements determination to the early stage of analysis is usually done by hand. Then the user develops a conceptual level model of the problem domain.

The model is developed usually using some special diagram language (UML language, for example). Conceptual level models describe a part of the real world for which the information system is being developed, so conceptual level class diagrams are right for describing the set of terms of the problem domain vocabulary.

When developing a model, the analyst usually processes by hand a large amount of documents referred to the problem domain in order to pick out key terms, properties, functions and relations between them. The proposed approach enables automation of this process. The intellectual cognitive analyzer being implemented according to the described approach should act as a user's assistant. It will do the most routine part of the work in the early stage of analysis.

The suggested approach also includes some other capabilities that let the computer become quite a good assistant for the human user not just at the beginning of analysis, but along its whole length. One of those capabilities, in particular, is the automatic problem domain thesaurus building during interaction with the user. And it is possible to use preinstalled thesauruses too, different ones for each problem domain, describing their specific components, features, etc.

Conceptual-level Modeling

As was said earlier, the purpose of the approach is automated construction of conceptual-level model diagrams of the problem domain. The UML language (static class section), was chosen as a model representing language, because it is the most popular standard for CASE tools nowadays.

UML static class diagrams define object classes and different kinds of static relations between them in the system. Also such things as class attributes, operations and relation limitations are usually shown on class diagrams.

There are three different points of view on how to interpret the sense of class diagrams:

- Conceptual-level point of view. If we take a class diagram from this point of view, then it reflects a set of terms and relations (called vocabulary) of the examined problem domain. Conceptual-level model considered independent from any software programming language.
- Specification-level point of view. In contrast to the above, this affects the software development range, but focuses attention over interfaces, not implementation. Looking at the class diagram from this point of view, designers have to do rather with types, not classes.
- Implementation-level point of view. In this case we really deal with graphical representation of the class structure of software. So the designer goes down to the level of implementation.

Understanding which point of view should be used and when, is extremely important either for developing or for reading class diagrams. Unfortunately, distinctions between them are not understood clearly, so the majority of developers often mix some different points of view when developing a diagram model.

The idea of the point of view on diagrams is not actually a formal part of UML language, but it is extremely important. UML constructions can be used with any of the three points of view in mind.

As has already been said, the suggested approach is going to be used for the automation of the process of conceptual-level problem domain model development. First of all, it is because of the fact that the approach should work at the most initial stage of IS development process. Apart from that, the nature of the documentation used in the problem domain of the considered range of IS (sphere of education) means that the description of objects and their mutual relations is of a sufficiently high level. This fact automatically determines the point of view on a problem domain as conceptual.

However, such a strict binding model to a conceptual level is not obligatory. In some cases the model can get an interpretation from some other point of view. This mainly depends on the nature of the source documents.

Conceptual-level diagrams describe the problem domain vocabulary. Of course, it is doubtful that diagrams developed using the suggested approach could be immediately used for generation of skeleton program code, but it can be used for subject domain database logic structure generation.

IES Architecture

Fig. 2 shows the diagram reflecting the principle according to which the projected system is organized.

Let us consider in more detail the principles assumed for the basis of the suggested approach.

Natural language expresses relations between items in a problem domain in the form of statements. For example, the statement “children study at schools” binds together the concept “school” belonging to the class “educational institutions” and the concept “children” belonging to the class “person”. Any statement can be either correct, or wrong, when established during correlation with reality. So, statements singled out from source documents should be compared to the problem domain thesaurus which reflects the current actuality. In the case of detection of a discrepancy of the obtained propositions to ones from the problem domain thesaurus, the latter should be brought into accord with reality, or the source proposition should be corrected in an appropriate way. When the system cannot make such a decision independently, it can apply for the human user's assistance.

The proposition (statement) is an expression that claims or disclaims the existence of an item, the relation between an item and its attribute, or the relation between two items. A sentence is the language form of the proposition. Propositions in natural language texts are expressed by narrative sentences, for example:

"institutions of primary vocational training may be state, municipal and private". The proposition of connexion of an item and its attribute consists of propositional subject, and a predicate reflecting an attribute of an item. Except for subject and predicate, the proposition includes a copula which can be put into words (for example, "not is", "is", etc.).

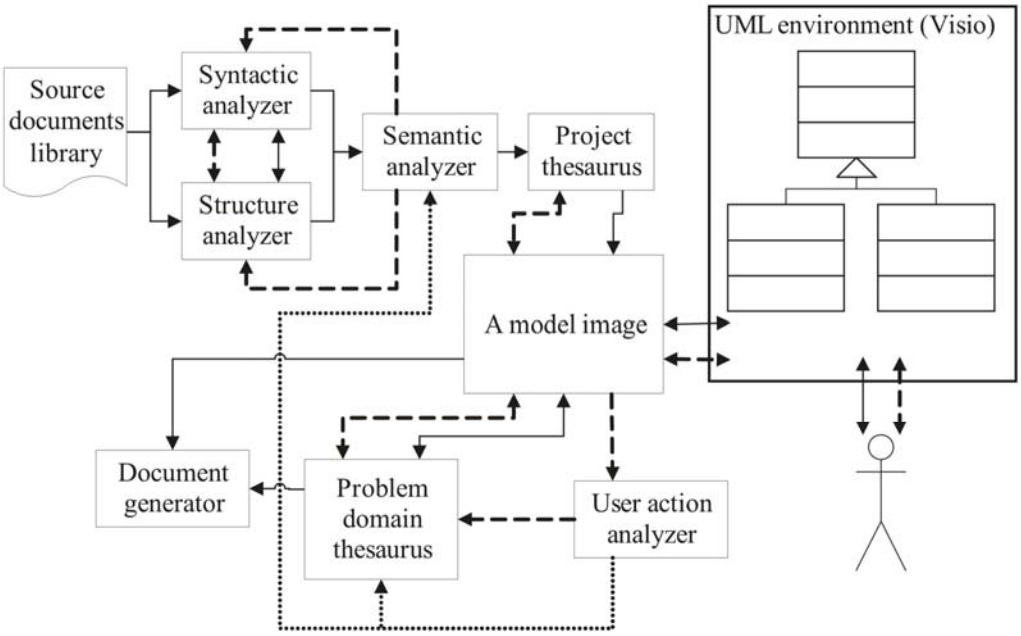


Figure 2. IES Architecture Framework

Depending on what is claimed or disclaimed in the proposition – either the existence of an item, or the relation between an item and its attribute, or the relation between two items – it may be classified as attributive proposition, proposition with relation and existential proposition. A proposition is called compound if it consists of several simple propositions, combined together in an expression.

A conceptual-level model is usually developed using source natural language description. Sentences in this description are propositions of listed types. Some of them concern certain objects; others are general as they concern some class of objects in the problem domain. Source documents consist of compound sentences that describe objects and relations between them in the problem domain.

In the course of linguistic analysis using knowledge of language structure, initial compound sentences are split into simple propositions of three listed types, and the type of each proposition can be determined during the process of decomposition.

The whole totality of concepts and relations between them, expressed by means of natural language, forms a system thesaurus. Thus, we can say it schematically represents the matter of the source documents text. The idea of a thesaurus is frequently applied to problems of semantic search in documents. Within the suggested approach, another variant of its application is offered.

Concepts are extracted from source documents during the linguistic analysis process. One of the basic relation types that make a thesaurus hierarchical, is the relation type named "is a". It realizes the idea of generalization, allowing reference of a narrower concept to a wider one.

Another relation type named "referred to" designates a certain reference between terms. It can be marked by a verb (predicate) extracted from the source sentence (this verb should describe the nature of the relation with certainty). This mark can also be a word-combination, consisting of a verb and a noun: "is operation", "is an attribute".

To avoid the possibility of the appearance of a vicious circle of interdependence of terms in a thesaurus, there are some rules to be obeyed:

- no term can be wider than itself, either directly, or indirectly (this limits the usage of the “is a”);
- no term can be “referred to” a term which is wider or narrower than itself, either directly, or indirectly.

The structure of the thesaurus can be represented by a graph (semantic net), its nodes correspond to terms, and arches are relations between terms. One set of arches forms a directed acyclic graph for the relation of generalization (“is a”). Another set forming the directed graph, represents the relation of referred terms (“referred to”). Relation types “is a” and “referred to” form subgraphs.

Principles of IES Operation

The thesaurus of the model should be populated and refreshed using an automatic mode. Thus there are the certain difficulties concerning natural language text processing. To overcome these difficulties successfully, the approach offers the multilevel circuit of text processing using the relaxation method to eliminate ambiguities.

At the initial stage of text processing the syntactic analyzer (figure 2) works. It implements the syntactic analysis and decomposition of compound sentences to the simple propositions consisting of subject, predicate and object. While these operations are being accomplished, the semantics of the sentence is not taken into account. During decomposition, the text of the source documents is transformed into a set of simple statements (propositions) of three listed types (attributive proposition, proposition with relation, existential proposition) which then can be easily subjected to semantic analysis.

It is important to note that relations between concepts are not necessarily conveyed syntactically in text. They can also be conveyed by the structure of the document. There are two types of structural compositions most frequently used in documents: table structure, determines attributive relations; list structure, determines relations of various kinds, between the concept located in the list heading and concepts located in lines.

In order to assure the completeness of analysis it is necessary to allocate relations, set by structures of listed types. This task is done by the structural analyzer, who’s output, as well as for syntactic analyzer, consists of simple propositions reflecting relations between concepts. The analyzer generates them using structural information extracted from the source text as the basis.

The semantic analyzer obtains the data processed by syntactic and structural analyzers, handles them for its turn and populates the system thesaurus with prepared data. If the semantic analyzer finds any variance in source data, caused by its ambiguity or uncertainty, it can address previous level of processing – syntactic or structural analyzer – with the requirement to give another variant of the text treatment. This idea accords with principles of the relaxation method. Some missing branches of concept relations may also be evoked from the existing thesaurus knowledge base.

There is one more task assigned to the semantic analyzer – to eliminate insignificant data. In fact the final model should not be formed by the whole totality of concepts and relations, allocated in the initial documents. First of all, some concepts may just slightly touch the scope of the given problem domain. Sometimes some errors in allocation of concepts and relations may take place because of text specificity or its author's verbiage. Anyway, some mechanism is required that could free the user from dealing with a lot of insignificant details. To achieve this, the semantic analyzer uses a special self-learning filter as a part of the project thesaurus. This filter determines a list of concepts that should not be included in the thesaurus. Relations of a special type “not relevant” may also be settled between the concepts in the thesaurus in order to solve the problem more effectively.

The filter is trained by tracking actions which are user made when editing a diagram. This way we can reveal insignificant concepts and relations in the problem domain to use this knowledge later.

We need to mention that there is one more important opportunity the approach can offer: an opportunity of distribution “on a turn-key basis” of an IS designing tool assigned for usage in the context of a certain problem domain. Such a tool would possess a prepared thesaurus establishing the set of basic concepts and relations and include a trained semantic filter focused over the scope of the problem domain being aimed at. In the architecture framework of IES which is being developed according to the suggested approach, this thesaurus is represented by the separate component called “Problem domain thesaurus” (figure 2).

The project thesaurus directly delivers data needed for production of model diagrams. The structure and sense of the thesaurus content allows translation of it into the model diagram. This is in spite of the fact that there are some minor distinctions between specifications of diagrams that could be used within the approach: UML diagrams and ER diagrams.

Diagrams are displayed in some external modelling environment which is connected to IES through the special buffer module of the model image. Of course, the user may want to correct the obtained model diagram, which is initially generated by the system. But nevertheless, it continues to cooperate with the user, helping him to perform the work.

Upon the user's demand it can generate some new fragments of the model diagram, if there are any new source documents obtained, or expand the model diagram in order to restore missing relations, applying knowledge from the problem domain and the project thesauruses, etc.

The system also traces user's actions made during model diagram editing. Such a feedback mechanism is absolutely necessary for implementing the idea of self-training as applied to the problem domain thesaurus and the semantic filter. Actually, during editing of the model diagram, the user “teaches” the system, providing it with the information about concepts and relations that are of first interest to him and ones that should be excluded from consideration. In such a way, the problem domain thesaurus containing the most authentic and clean information on key concepts and typical relations between them is built. It is populated automatically during editing of the model diagram. Thus, the resulting model diagram and successive modifications made by the user are also a source of information for the IES.

The system tries to recognize semantics in the model diagrams. So a diagram which the user works with is not a senseless set of blocks and connections for a computer any more. Attention is focused on names of elements, their structure, interfacing, etc. All these aspects are analyzed by the system.

Objects and relations allocated in a problem domain, organize a model. When the diagram is built, they remain connected with texts in the source documents library. It is necessary for the user to have an opportunity to supervise the correctness of the constructed model, verifying it directly with the key information from source documents. Reverse referencing from source documents to elements of a model diagram is also needed, because documents are not something immutable. The documents library has a dynamic nature – precepts may be cancelled, or changed in some points, etc. Direct and reverse referencing between source texts and the model assure an opportunity of efficient model updating.

Examples

Now we give an example demonstrating some aspects of the approach.

Please note that the approach is being developed for use jointly with the Russian language, where the concepts’ mutual interdependence in sentences is expressed much less ambiguously than in English, at the syntax level.

Let us show how a certain expression is going to be analyzed by the system:

“Educational institutions with government accreditation grant certificates to persons who passed attestation”.

During syntactic analysis the given sentence is split into some connected simple statements which can be easily represented by the semantic network shown on fig. 3.

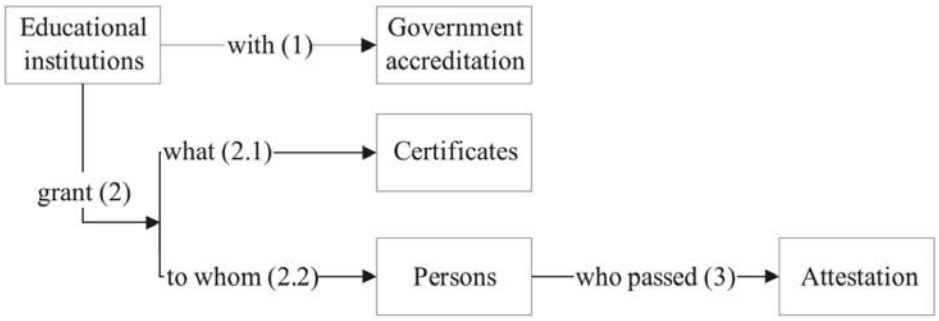


Figure 3. Semantic Network Representing Sample Sentence Structure

The semantic analyzer qualifies propositions (1, 2 and 3) such as ones with relations. Thus the verb predicate representing the action “grant” is interpreted by the semantic filter as an operation (class method). But let us assume that such interpretation is not known to semantic filter.

Simple propositions obtained which form marked section of a semantic network after the stage of semantic analysis, are directed to the problem domain thesaurus.

Propositions with relations of such a kind are displayed in the model as objects connected by the relation “referred to”; connection is directed from a subject to an object and represents the predicate (see fig. 4).

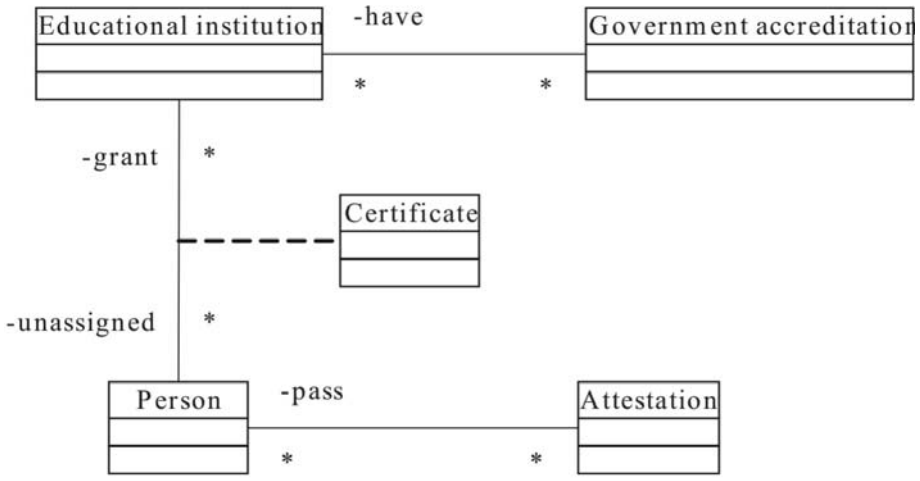


Figure 4. Model Framework Created on the Sentence

Part of the model received as a result of analysis of a given sentence, could be automatically attached to the existing model by a set of “is a” connections, revealed by the semantics comparison.

Besides that, if the problem domain thesaurus contains information about other connections between these objects and ones in the problem domain, these connections will also be restored in the model.

So, let us return to the necessity that the action “grant” should be interpreted as a method.

If it does not happen automatically, then the user manually creates the method “grant” in the object “Education institution”. After that, as a result of the semantics comparison of the operation name assigned by the user with the text of source sentence, the semantic filter is trained to interpret the verb “to grant” as the method (operation) at a later time.

Analyzing a similar text subsequently, the system should automatically add a corresponding object operation to the model. The thesaurus of the model is populated and refreshed in an automatic mode.

Bibliography

[Aiello, 2000] M. Aiello, C. Monz, L. Todoran. Combining Linguistic and Spatial Information for Document Analysis, In: Content-Based Multimedia Information Access. CID, 2000, pp. 266-275.

[Connolly, 1999] T.M. Connolly, C.E. Begg. Database Systems. A Practical Approach to Design, Implementation, and Management. Addison Wesley Longman, Inc, 1999.

[Fowler, 1997] M. Fowler, C. Scott. UML Distilled: Applying the Standard Object Modeling Language. Addison Wesley Longman, Inc, 1997.

[Johnson-Laird, 1983] P.N. Johnson-Laird. Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press, 1983.

[Katz , 1964] J.J. Katz, J.A. Fodor. The Structure of Language, Prentice-Hall, 1964.

[Larman, 2000] C. Larman. Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design. Prentice Hall, Inc, 2000.

[Miller, 1990] G. Miller. Wordnet: An on-line lexical database. In: International Journal of Lexicography, 3(4), 1990.

Authors' Information

Maxim Loginov - Perm State University, Assistant of the Department of Computer Science; PSU, 15, Bukirev St., Perm, 614990, Russia; e-mail: login@perm.ru

Alexander Mikov – Institute of Computing, Director; 40/1-28, Aksaiskaya St., Krasnodar, Russia; e-mail: alexander_mikov@mail.ru

TEXT-TO-TEXT MACHINE TRANSLATION SYSTEM (TTMT SYSTEM) – A NOVEL APPROACH FOR LANGUAGE ENGINEERING

Todorka Kovacheva, Koycho Mitev, Nikolay Dimitrov

Abstract: *The purpose of the current paper is to present the developed methodology for automatic machine translation. As a result the Text-to-Text Machine Translation System (TTMT System) is designed. The TTMT System is developed as hybrid architecture, combining different machine translation approaches. The included languages in the base version are English, Russian, German and Bulgarian. The TTMT System is designed as a universal polyglot. The architecture of the system is highly modular and allows other languages easy to be included. It uses a digital script and method for communication.*

Keywords: *machine translation (MT), natural language processing (NLP), language engineering (LE), text-to-text translation (TTT), text-to-text machine translation system (TTMT System), universal polyglot, digital language.*

Introduction

Automatic translation between human languages, also known as “machine translation” is of enormous social, political, and scientific importance. In the age of global and information based society real-time online translation on the Internet and other resources will support personal communication and information needs. There are many attempts to develop integrated translation software which can work in different scientific domains. There is already research on cross-lingual information retrieval, multilingual summarization, multilingual text generation

from databases, integration of translation with summarization, database mining, document retrieval, information extraction, etc. In machine translation research, there is much interest in exploring new techniques in neural networks, parallel processing, and particularly in corpus-based approaches: statistical text analysis (alignment, etc.), example-based machine translation, hybrid systems combining traditional linguistic rules and statistical methods, etc. [Hutchins J., 2002]

Because of the importance of the automatic translation systems in the current development phase of the society, the project for Text-to-Text Machine Translation System (TTMT System)[1] design and coding has started. In the present paper we present the TTMT System overview and hybrid architecture, based on different machine translation approaches. A digital script and method for communication[2] is also used.

Machine Translation as a Subfield of Language Engineering and Natural Language Processing

Language engineering may be defined as a discipline or act of engineering software systems that perform tasks involving processing human language [Cunningham H., 1999]. The language engineering not only collects the information and knowledge of a language among the linguistic society but also serves as a foundation on which linguistic culture and technologies can be based [Oh et. al., 1004]. It is the base for the development of natural language processing (NLP) systems. These systems use the NLP technologies, which combine algorithms and methods from artificial intelligence and linguistics. They are designed to solve the problems of automated generation and understanding of natural human languages.

Machine translation (MT) is one of the major tasks in Natural Language Processing. It investigates the use of computer software to translate text or speech in between natural languages. It consists of two major parts: decoding the meaning of the source text, and re-encoding this meaning in the target language. It is based on computational linguistics and modeling of natural languages from a computational perspective. The knowledge about the syntax, morphology, semantics and pragmatics of languages is needed. Therefore the translation process can be defined as a complex task.

Nowadays different approaches to machine translation exist. They can be divided in four main groups as follows:

1. Dictionary-based machine translation, which is based on dictionary entries (the translation is done word by word, without much correlation of meaning between them). For more information on dictionary-based machine translation, see [Ballesteros L., Croft W.B., 1996].
2. Statistical machine translation, which try to capture regularities of natural language using probability distributions of linguistic events, such as the appearance of words within a context, sentences, or whole documents. For more information about statistical language modeling and statistical machine translation see [Brown P., et. al., 1990; Casacuberta F., 1996].
3. Example-based machine translation, which is essentially a translation by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning. For more information about example-based machine translation, see [Brown R., 1996].

---

[1] The Project is currently funding by Gluon Technologies Ltd., Varna, Bulgaria,
URL: http://www.gluontechnologies.com

[2] Mitev K., BG Patent 63704 B1, Digital Script and Method for Communication on Moder Tongue

4. Rule-based machine translation. They create an intermediary symbolic representation, from which the text in the target language is generated. The approach is also known as interlingual machine translation, or transfer-based machine translation. These methods require extensive lexicons with morphological, syntatic, and semantic information, and large sets of rules.

The designed TTMT System uses a hybrid approach to machine translation, which contains the four mentioned above.

Text-to-Text Machine Translation System (TTMT System) Overview

The TTMT System is based on an in-depth analysis of the human language. The human speech can be defined by means of the digits from decimal system regardless of the speaker's mother tongue. The digital script is developed. It contains combinations of numbers from 0 to 9, which a computer device can store in its memory, and to translate finished texts from a random language into any other.

The digital script is applicable in all the world languages and dialects. Thus, the ten numbers can be turned into a universal tool for communication. The unique sequence of operations integrated in the software design to allow real time translation of a finished thought, sentence by sentence, in the respective word order. The operations are common for the phonetic speech and the speech in-writing. It allows communication in the mother tongue: voice – to voice; text – to text; voice – to text and vise-versa.

The TTMT System is designed to solve problems of one of the main tasks in the field of natural language processing – text-to-text machine translation. The methodology is developed for four base languages – English, Russian, German and Bulgarian. The translation can be done from every language to every language from those four languages. TTMT System design allow easy to include new languages. The flexible architecture of the systems makes it easy to turn into universal translator. Many of the benefits of improvement to the system flow automatically to outputs in all the languages. The system can take ready made sub-systems such as black boxes or as open source software and incorporate them to it.

The approach is simple. The whole process is divided into different tasks which are solved independently by system modules.

TTMT System Architecture

The architecture of TTMT system is highly modular. The complex problem of MT has been broken into smaller sub-problems. Every sub-problem is a task which is handled by an independent module. The modules are put together in a united system. The output of the previous module becomes the input of the following module. Because each module has a specific task, the complexity remains under control.

The modules are put together in the three main parts of the system: a ***front-end***, ***middle*** and ***back-end***. The ***front-end*** takes input in the form of text from Internet and other electronic sources. It includes also the user interface, for adjusting the text part parameters. The output is a used to fill in the database. For different languages different databases exist.

The middle part of the TTMT System is built from the ***languages databases***, ***dictionary management sub-system***, ***language analyzer and model generation tools***, ***language alignment sub-system***, ***validation and verification tools***.

The ***language database*** consists of two parts: dictionary and metadata. The dictionaries can be represented as monolingual data banks, covering the data in a specific domain, and multilingual dictionaries for different domains (medical science, law, electrical and electronic engineering, technical and computer-science domain, every-day language, etc. The available dictionaries are semantically oriented using ontology-based lexicon and text part parameters, which represent the metadata used in the system.

The data in the databases are coded digitally using the digital script. The syntax, morphology, semantics and pragmatics are also taken under consideration by coding.

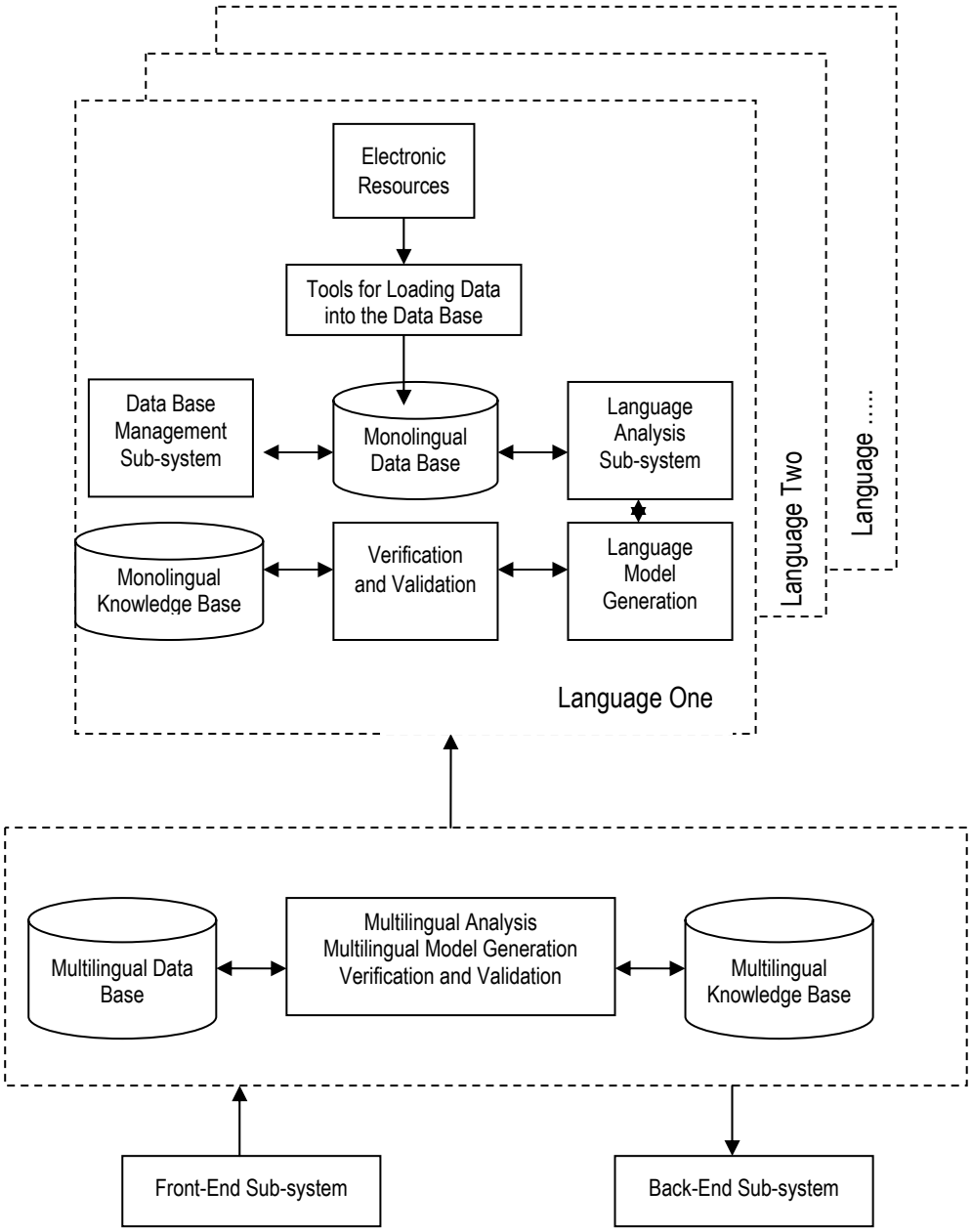


Fig.1. TTMT System Architecture

The ***dictionary management sub-system*** is used to provide customizable information and managerial functions for text data bases, and to provide the environment for dictionary development and management and merging existing dictionaries. It contains tools for extraction, indexing and searching the data. Because of the big size of each text to be stored and lots of keywords to be indexed and searched for each text, it requires special storing and managing mechanisms. This is also the need from the dictionary management sub-system.

For every language a ***language analysis package*** is designed. The ***language model*** is a result from that analysis. Morphological Analyzer is an important part of the analysis. It is extended to cover special symbols, abbreviated words, spell errors, etc.

Language alignment sub-system gathers the correspondences between representations of different languages.

Validation and verification tools are used to estimate the result, to correct the errors and learn the system to avoid this error in the future work.

The ***back-end*** outputs the synthesized translated text.

The TTMT System architecture is presented in Fig.1.

Conclusion

The developed methodology is a novel approach to natural language engineering and natural language translation. It makes possible to build the Universal Machine Translation System and to implement it in modern telecommunication systems. Future work is to include voice recognition and voice generation sub-system, and to produce speech-to-text, text-to-speech, and speech-to-speech automatic translation

Bibliography

[Hutchins J., 2002] Hutchins J., Machine translation today and tomorrow, In *Computerlinguistik: was geht, was kommt?* Festschrift für Winfried Lenders, hrsg. Gerd Willée, Bernhard Schröder, Hans-Christian Schmitz. Sankt Augustin: Gardez! Verlag, 2002, pp.159-162.

[Cunningham H., 1999] Cunningham H., A Definition and Short History of Language Engineering, Journal of Natural Language Engineering, 5(1):1-16, 1999

[Oh et. al., 1004] Oh, Gil-Rok, Choi, Key-Sun, and Park, Se-Young, Hangul Engineering, Seoul, Korea: Daeyoungsa, 1994

[Brown P., et. al., 1990] Brown P.F., Cocke J., Della Pietra S.S., Della Pietra V.J., Jelinek F., Lafferty J.D., Mercer R.L., Roosin P.S., A statistical approach to machine translation, Computational Linguistics, 16(2):79-85, 1990

[Casacuberta F., 1996] Casacuberta F., Growth transformations for probabilistic functions of stochastic grammars, International Journal of Pattern Recognition and Artificial Intelligence, 10(3):183-201, 1996

[Ballesteros L., Croft W.B., 1996] Ballesteros L., Croft W.B., Dictionary-based methods for cross-lingual information retrieval, In Proceedings of the 7th International DEXA Conference on Database and Expert Systems Applications, pp.791-801, 1996

[Brown R., 1996] Brown R., Example-Based Machine Translation in the Pangloss System, In Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp.169-174, Copenhagen, Denmark, 1996

Authors' Information

Todorka Kovacheva – Economical University of Varna, Kniaz Boris Str, Varna, Bulgaria, e-mail: todorka_kovacheva@yahoo.com, phone: +359899920659

Koycho Mitev – e-mail: patentservice@abv.bg, phone: +359899103367

Nikolay Dimitrov – Varna, Bulgaria, phone: +359896325805

Information Models

ON THE COHERENCE BETWEEN CONTINUOUS PROBABILITY AND POSSIBILITY MEASURES

Elena Castiñeira, Susana Cubillo, Enric Trillas

***Abstract**: The purpose of this paper is to study possibility and probability measures in continuous universes, taking different line to the one proposed and dealt with by other authors. We study the coherence between the probability measure and the possibility measure determined by a function that is both a possibility density and distribution function. For this purpose, we first examine functions that satisfy this condition and then we anlyse the coherence in some notable probability distributions cases.*

***Keywords**: Measure, possibility, probability, necessity.*

***ACM Classification Keywords**: I.2.3 Artificial Intelligence: Deduction and Theorem Proving (Uncertainty, "fuzzy" and probabilistic reasoning); I.2.4 Artificial Intelligence: Knowledge Representation Formalisms and Methods (Predicate logic, Representation languages).*

Introduction

Possibility distributions are, sometimes, good means for representing incomplete crisp information. It is precisely this incompleteness that often makes it impossible to determine a probability that could describe this information. Now, if the possibility distribution meets certain requirements, for example, it is either a density function or its graph "encloses" a finite area, it will always be possible to consider either the probability whose density function is this possibility distribution or an associated density function.

Fuzzy set-based possibility theory was introduced by L. Zadeh in 1978 (see [12]) and provided an alternative non-classical means, other than probability theory, of modeling and studying "uncertainty". Zadeh established in [12] the principle of consistency between possibility and probability, according to which "anything that is probable must be possible". This principle is expressed as "$\mathrm{P}(A)\leq \Pi(A)$", and the probability P could also be said to be coherent with the possibility Π. The finite case has been studied by M. Delgado and S. Moral in [4], where they characterise the probabilities that are coherent with a given possibility; also in [2] Castiñeira *et al.* deepened in that case defining a distance between possibility an probability measures, finding the closest probability to a given possibility and proving they are coherent. The case of continuous universes has been addressed by several authors, including Dubois *et al*, who, in [7], examined possibility/probability transformations taking into account the principle of insufficient reason from possibility to probability, and the principle of maximum specificity from probability to possibility. Although dealing with the same subject, the purpose of this paper is another. As density functions are to probabilities what possibility distributions are to possibility measures and, taking into account that a density function whose value is 1 at any point determines both a probability measure and a possibility measure, we set out to analyse the coherence between these probability and possibility measures.

This paper is organized as follows: After a background section, in section 2, we prove that a possibility generates a degenerated probability defined on a σ-algebra, as in the finite case where the coincident probabilities and

possibilities were degenerated. In section 3, some functions are obtained which are both possibility distributions and density functions; particularly, some classic distributions have been considered, then we address the problem of coherence between possibilities and probabilities generated by the same function. Some counterexamples show that, even in these cases, the coherence between measures cannot be guaranteed. Finally, in section 4, we deal with the coherence between some classical probability distributions and their respective possibility measures, stressing the case of the normal law, where there exists coherence.

1. Preliminaries

Let $\mathrm{F}(E)$ be the set of all fuzzy sets in a universe of discourse $E\neq \emptyset$, with the partial order $\subset$ defined by $P\subset Q$ if and only if $\mu_P(x)\leq \mu_Q(x)$ for all $x\in E$, where $\mu_P, \mu_Q\in [0,1]^E$ are the membership functions of P and Q, respectively. A $\subset$-measure in $\mathrm{F}(E)$ is any function $\mathrm{M}:\mathrm{F}(E)\rightarrow[0,1]$ such that: m_1) $\mathrm{M}(\emptyset)=0$; m_2) $\mathrm{M}(E)=1$; m_3) If $P\subset Q$, then $\mathrm{M}(P)\leq \mathrm{M}(Q)$.

Considering the standard fuzzy sets theories $(\mathrm{F}(E),\cup,\cap, ^c)$ where the operations $\cup, \cap$ are defined by the t-norm $\mathrm{T}=\mathrm{Min}$, the t-conorm $\mathrm{S}=\mathrm{Max}$ and the complement ^{c} by means of a strong negation [11], on the one hand, a possibility in $\mathrm{F}(E)$ (see [5] and [6]) is any mapping $\Pi:\mathrm{F}(E)\rightarrow[0,1]$ satisfying: p_1) $\Pi(E)=1$; p_2) $\Pi(\emptyset)=0$; and p_3)$\Pi(P\cup Q) = \mathrm{Max}(\Pi(P), \Pi(Q))$ for any $P,Q\in \mathrm{F}(E)$. On the other hand, a necessity in $\mathrm{F}(E)$ is any mapping $\mathrm{N}:\mathrm{F}(E)\rightarrow[0,1]$ satisfying: n_1) $\mathrm{N}(E)=1$; p_2) $\mathrm{N}(\emptyset)=0$; and p_3) $\mathrm{N}(P\cap Q)=\mathrm{Min}(\mathrm{N}(P), \mathrm{N}(Q))$ for any $P,Q\in \mathrm{F}(E)$.

It is easy to check that both any Π and any N verify the axiom m_3, and are, therefore, $\subset$-measures. Note that, given a possibility Π, the function $\mathrm{N}_\Pi=1-(\Pi\circ ^c)$ is a necessity measure, the bidual necessity associated with Π.

Furthermore, if $\mu\in[0,1]^E$ is such that $\sup\{\mu(x), x\in E\}=1$, the function Π_μ: $\mathrm{F}(E)\rightarrow[0,1]$ defined for all $P\in \mathrm{F}(E)$ by $\Pi_\mu(P) = \sup\{\mathrm{Min}(\mu(x),\mu_P(x)), x\in E\}$ is a possibility measure. The function μ is called possibility distribution of the Π_μ. Note that for all $A\in \mathrm{P}(E)$, where $\mathrm{P}(E)$ is the set of parts of E, the possibility measure given by the possibility distribution μ is defined by $\Pi_\mu(A)=\sup\{\mu(x), x\in A\}$.

Let M_1 and M_2 be two $\subset$-measures, M_1 is *coherent* with M_2 if $\mathrm{M}_1(A)\leq \mathrm{M}_2(A)$ for all $P\in \mathrm{F}(E)$. When $\mathrm{M}_1=\mathrm{N}$ is a necessity measure and $\mathrm{M}_2=\Pi$ is a possibility measure, it is clear that, generally, there is no coherence between N and Π, that is, neither $\mathrm{N}\leq\Pi$, nor $\Pi\leq \mathrm{N}$. Nevertheless, when $\mathrm{N}=\mathrm{N}_\Pi$ is the necessity measure associated with the possibility measure Π, $\mathrm{N}_\Pi\leq\Pi$ because $1=\Pi(P\cup P^c)=\mathrm{Max}(\Pi(P), \Pi(P^c))\leq \Pi(P)+\Pi(P^c)$, thus $\mathrm{N}_\Pi(P)=1-\Pi(P^c)\leq\Pi(P)$.

As the purpose of this paper is to compare possibility and probability measures, we will consider the possibilities as being restricted to classic sets, that is, to σ–algebras $\mathrm{A}\subseteq\mathrm{P}(E)$. Recall that A is a σ–algebra if for any $A\in\mathrm{A}$ its complement $A^c\in\mathrm{A}$, and for any countable family $\{A_n\}_{n\in\mathrm{N}}\subseteq \mathrm{A}$ it is $\bigcup_{n\in\mathrm{N}} A_n\in\mathrm{A}$. Moreover, P: $\mathrm{A}\rightarrow[0,1]$ is a probability measure if $\mathrm{P}(E)=1$ and P is σ -additive, that is, for any $\{A_n\}_{n\in\mathrm{N}}\subseteq \mathrm{A}$ such that $A_i\cap A_j=\emptyset$ if $i\neq j$, then $\mathrm{P}!\bigcup_{n\in\mathrm{N}} A_n 1=\sum_{n\in\mathrm{N}}\mathrm{P}(A_n)$ holds.

We will consider Borel's σ-algebra in R, that is, the smallest σ-algebra that contains the semi-ring $\{[a,b); a, b\in\mathrm{R}$ with $a<b\}$, or alternatively, the smallest σ-algebra that contains the open sets of R, and which is usually denoted by B. It is well known that every probability measure P: $\mathrm{B}\rightarrow[0,1]$ is univocally determined by a distribution function, F: $\mathrm{R}\rightarrow[0,1]$ ([10]), and if $F'(x)=f(x)$ exists for "almost any" point, then $\mathrm{P}([a,b])=\int_a^b f(x)dx$ (*). Generally, if f: $\mathrm{R}\rightarrow\mathrm{R}^+$ is such that $\int_{-\infty}^{\infty} f(x)dx=1$ (that is, f is a density function), f defines, as in (*), a probability measure on Borel's algebra of R. Note that, pursuant to the theorems of measure extension, every probability in (R,B) is determined by ascertaining its values in the intervals $[a,b]$.

2. Probability Generated by a Possibility Measure

Let $\mu \in [0,1]^E$ such that $\sup\{\mu(x), x\in \mathrm{R}\}=1$ and let us consider the possibility measure generated by μ on the crisp sets of R, that is, Π_μ:P(R)→[0,1] defined for each $A\in$ P(R) by $\Pi_\mu(A)=\sup\{\mu(x), x\in A\}$, Π_μ verifies: 1) $\Pi_\mu(\emptyset)=0$; 2) Monotonicity: if $A\subseteq B$ then $\Pi_\mu(A)\leq \Pi_\mu(B)$; 3) Subadditivity: $\Pi_\mu(\cup_{n\in \mathrm{N}} A_n)\leq \Sigma_{n\in \mathrm{N}} \Pi_\mu(A_n)$.

That is, Π_μ is an exterior measure in R and also verifies that $\Pi_\mu(\mathrm{R})=1$.

It is known that any exterior measure M generates a σ-additive measure on the σ-algebra of the M –measurable sets (see [9], [10]) according to:

Caratheodory's Theoreme: If M: P(E)→ R^+ is an exterior measure in a set $E\neq \emptyset$, then the family A=$\{A\in \mathrm{P}(E); \forall X\in \mathrm{P}(E), \mathrm{M}(X)= \mathrm{M}(X\cap A)+ \mathrm{M}(X\cap A^c)\}$ is a σ-algebra and the restriction of M to A is a σ-additive measure.

The Caratheodory´s method applied to the exterior measure Π_μ generates a degenerated probability as follows:

Theoreme 2.1. Let $\mu \in [0,1]^E$ such that $\sup\{\mu(x), x\in \mathrm{R}\}=1$, then the family of Π_μ-measurable sets is

$$\mathrm{A}=\{ A\in \mathrm{P}(\mathrm{R}), \mathrm{supp}(\mu)\subset A \text{ or } A\subset (\mathrm{supp}(\mu))^c\},$$

where $\mathrm{supp}(\mu)=\{x\in \mathrm{R}, \mu(x)\neq 0\}$ is the support of μ, and the possibility measure Π_μ restricted to the Π_μ -measurable sets is a degenerated probability defined for each $A\in$ A by $\Pi_\mu(A)=0$, if $A\subset (\mathrm{supp}(\mu))^c$, and $\Pi_\mu(A)=1$ if $\mathrm{supp}(\mu)\subset A$.

Proof: Let us see that A is the σ-algebra constructed by Caratheodory's method.

A is a σ-algebra trivially. The elements of A are Π_μ-measurable; indeed, if $\mathrm{supp}(\mu)\subset A$, for each $X\subset$ R,

$$\begin{aligned}\Pi_\mu(X) &= \sup\{\mu(x), x\in X\} = \sup\{\mu(x), x\in (A\cap X)\cup(A^c\cap X)\}\\ &= \mathrm{Max}\{\sup\{\mu(x), x\in A\cap X\}, \sup\{\mu(x), x\in A^c\cap X\}\}\\ &= \sup\{\mu(x), x\in A\cap X\} = \Pi_\mu(A\cap X) = \Pi_\mu(A\cap X)+\Pi_\mu(A^c\cap X)\end{aligned}$$

holds, as $\Pi_\mu(A^c\cap X)=0$. Similarly, if $A\subset(\mathrm{supp}(\mu))^c$, we could prove that A is Π_μ-measurable.

Furthermore, we will prove that the only Π_μ-measurable elements are elements of A: If $A\subset$ R is Π_μ-measurable, then, in particular, $1= \Pi_\mu(\mathrm{R})=\Pi_\mu(A)+ \Pi_\mu(A^c)$ (•) holds, and two options can be study:

1) There exists $x_0\in$ R such that $\mu(x_0)=1$. If, moreover, $x_0\in A$ it follows from (•) that $\Pi_\mu(A^c)=0$, which means that $A^c\subset (\mathrm{supp}(\mu))^c$ and, therefore, $\mathrm{supp}(\mu)\subset A$ and $A\in$ A. Similarly, if $x_0\in A^c$, we have that $A\subset (\mathrm{supp}(\mu))^c$ and $A\in$ A.

2) For all $x\in$ R, $\mu(x)<1$. In this case, μ reaches its supreme value at $+\infty$ or $-\infty$, and this point of infinity is an accumulation point of A, $x\in A$', or of A^c. Let us suppose that $x\in A$', then $\Pi_\mu(A)=1$, and it follows from (•) that $\Pi_\mu(A^c)=0$, which means that, again, $\mathrm{supp}(\mu)\subset A$ and $A\in$ A. If the point of infinity at which μ reaches the supreme is an accumulation point of A^c, it follows, similarly, that $A^c\subset (\mathrm{supp}(\mu))^c$ and $A\in$ A.

Finally, the values of Π_μ on elements of A follow from the definition of Π_μ. □

3. Possibility and Probability Measures Generated by a Density Function and Their Coherence

We will address the coherence of measures in a continuous universe when the possibility and probability are determined by the same function, that is, a possibility distribution in the first instance and a density function in the second one. For this purpose, a first section analyses how this type of functions can be derived from a given density function and, then, from a given possibility distribution. The second section deals with the coherence between a possibility and a probability both generated by a given density function.

3.1. Possibility Distributions and Density Functions

In this section, some conditions for a function to be a density function and a possibility distribution at the same time are stated; moreover the cases of some notable distributions are analysed.

Lemma 3.1. If f: R $\rightarrow$ R^+ is a bounded density function, then the function μ_f : R $\rightarrow$ R defined for each $x\in$ R by $\mu_f(x) = kf(kx),$ where k=1/sup $\{f(x), x\in \mathsf{R}\}$, is a density function and a possibility distribution function.

Additionally, if f is continuous, then there exists $y_0\in$ R such that $\mu_f(y_0) = 1.$

***Proof*:** μ_f is a density function. Indeed, $\int_{-\infty}^{\infty}\mu_f(x)dx = \int_{-\infty}^{\infty} f(kx)d(kx) = 1$ It is also a possibility distribution, since $0 \le \mu_f(x) \le \sup\{\mu_f(x),\ x\in \mathsf{R}\} = k \sup\{f(kx),\ x\in \mathsf{R}\} = 1.$

Finally, if f is continuous, there exists $x_0\in$ R such that $f(x_0) = \sup\{f(x),\ x\in \mathsf{R}\} = 1/k$; hence, it suffices to consider $y_0 = x_0/k$, since then $\mu_f(x_0/k) = 1.$ □ μ_f will be said to be the possibility distribution associated with f.

Some examples

The possibility distributions associated with some well-known probability distributions are listed below (for more details about these distributions, see [3]).

(a) Normal distribution of parameters α,σ, $N(\alpha,\sigma)$: Its density function is $f(x) = (1/\sqrt{2\pi})e^{-(x-\alpha)^2/2\sigma^2}$ with maximum $f(\alpha) = 1/(\sigma\sqrt{2\pi}),$ then $\mu_f(x) = \sigma\sqrt{2\pi}\, f(\sigma\sqrt{2\pi}x) = e^{-(\sigma\sqrt{2\pi}x-\alpha)^2/2\sigma^2}.$ In particular, when $\sigma = 1/\sqrt{2\pi}$, $\mu_f(x) = f(x) = e^{-\pi(x-\alpha)^2}$ which is a density function for the normal distribution $N(\alpha,1/\sqrt{2\pi})$.

(b) Cauchy distribution with parameters a,b: Its density function is $f(x) = \dfrac{a}{\pi(a^2+(x-b)^2)}$ whose maximum, reached in $b,$ is $f(b) = 1/(a\pi)$; hence, its associated possibility distribution is

$$\mu_f(x) = a\pi f(a\pi x) = \frac{a^2}{a^2+(a\pi x-b)^2}$$

If $b = 0,$ then $\mu_f(x) = 1/(1+\pi^2x^2)$, and its probability distribution is a Cauchy distribution with $a = 1.$

(c) Gamma distribution of parameters $p>0, a>0, \Gamma(p,a)$: Its density function is $f(x) = \dfrac{a^p}{\Gamma(p)}x^{p-1}e^{-ax}$ if $x>0,$ and $f(x) = 0$ if $x\le 0,$ where $\Gamma(p) = \int_0^{+\infty} e^{-x}x^{p-1}dx$ is the second-class Euler's function.

Note, firstly, that if $p\in(0,1),$ then f is not bounded and, therefore, there is no associated possibility distribution. When $p = 1,$ it is also a particular case of the exponential distribution that will be dealt with in the following example. If $p>1,$ the function is bounded, reaching its maximum value in $x = \dfrac{p-1}{a},$ and its associated

possibility distribution can be ascertained. It will be calculated for two particular cases so as to avoid tedious calculations.

If $p = 2,$ then $f(x) = a^2 x e^{-ax}$ if $x > 0,$ and $f(x) = 0$ if $x \le 0,$ and its maximum is $f(1/a) = a/e;$ therefore, the associated possibility distribution is $\mu_f(x) = e^2 x e^{-ex}$ if $x > 0,$ and $\mu_f(x) = 0,$ if $x \le 0,$ which is also a density function for the distribution $\Gamma(2, e).$ If $p = 3,$ we get the law $\Gamma(3, e^2/2).$

(d) Exponential Distribution of parameter θ: Its density function is $f(x) = \theta e^{-\theta x}$ if $x \ge 0,$ and $f(x) = 0$ if $x < 0,$ whose maximum is $f(0) = \theta.$ Hence, the associated possibility distribution is $\mu_f(x) = e^{-x}$ if $x \ge 0,$ and $\mu_f(x) = 0,$ if $x < 0,$ which is a density function for the exponential distribution with $\theta = 1$ or also for $\Gamma(1,1).$ ◁

The inverse problem of getting a density function that is also a possibility distribution from another possibility distribution is easily solved if this distribution "encloses" a finite area, as shown in the following result.

Lemma 3.2. Let $\mu \in [0,1]^{\mathsf{R}}$ such that $\int_{\mathsf{R}} \mu(x)dx = A < +\infty$ and let us suppose that there exists $x_0 \in \mathsf{R}$ such that $\mu(x_0) = 1,$ then the function defined for each $x \in \mathsf{R}$ by $f_\mu(x) = \mu(A(x - x_0) + x_0)$ is a density function and also a possibility distribution.

Proof: Let $\alpha(x) = A(x - x_0) + x_0,$ then $\int_{-\infty}^{+\infty} f_\mu(x)dx = \int_{-\infty}^{+\infty} \mu(\alpha(x))dx = \frac{1}{A}\int_{-\infty}^{+\infty} \mu(\alpha(x))d(\alpha(x)) = 1$ holds.

Therefore, μ is a density function. Additionally, $f_\mu(x_0) = \mu(x_0) = 1,$ which means that f_μ is also a possibility distribution. □ f_μ will be said to be a density function associated with μ.

Note that there are many density functions associated with a function μ under the above conditions. Indeed, $f_\mu(x) = Ax$ and all its translations would also be density functions. The fact that we considered the translation to x_0 is really a practical matter, as if $\mu.$ reaches the value 1 at a single point $x_0,$ then the graph of f_μ is obtained by "squashing" the graph of $\mu.$ and leaving the fixed point $(x_0, 1),$ which would mean that it would be "most like" the original μ.

Example: The function $\mu(x) = e^{-|x|},$ with $x \in \mathsf{R},$ is a possibility distribution, since $\mu \in [0,1]^{\mathsf{R}}$ and $\mu(0) = 1,$ but it is not a density function, as $\int_{-\infty}^{+\infty} e^{-|x|}dx = 2.$ However, associated density functions can indeed be found: $f_\mu(x) = e^{-2|x|}$ and its translations.

3.2 Coherence Between Possibility and Probability

Let $\mu \in [0,1]^{\mathsf{R}}$ such that $\int_{\mathsf{R}} \mu(x)dx = 1$ and $\sup_{x \in \mathsf{R}} \mu(x) = 1.$ Let Π_μ be the generated possibility by μ and P_μ the probability with density function $\mu.$ Our aim is to study when $\mathrm{P}_\mu \le \Pi_\mu,$ that is, when P_μ is coherent with $\Pi_\mu.$ The following result shows that there is "local coherence" with the possibility for "small" subsets.

Theorem 3.3. Let $A \in \mathsf{B}$ such that $\mathsf{L}^1(A) \le 1,$ where L^1 designates the Lebesgue measure in R; then for any $\mu \in [0,1]^{\mathsf{R}}$ such that $\int_{\mathsf{R}} \mu(x)dx = 1$ and $\sup_{x \in \mathsf{R}} \mu(x) = 1,$ it is $\mathrm{P}_\mu(A) \le \Pi_\mu(A).$

Proof: $\mathrm{P}_\mu(A) = \int_A \mu(x)dx \le \sup_{x \in A} \mu(x) \cdot \mathsf{L}^1(A) \le \Pi_\mu(A).$ □

Generally, it cannot be guaranteed that $\mathrm{P}_\mu(A) \le \Pi_\mu(A)$ for any $A \in \mathsf{B}$, as shown by the following examples.

Pareto distribution of parameters a, x_0 : Its density function is $f(x) = \frac{a}{x_0}\left(\frac{x_0}{x}\right)^{a+1}$ if $x \ge x_0$, and $f(x) = 0$ if $x < x_0$, and its associated possibility function taking $x_0 = a$ is $\mu_f(x) = (a/x)^{a+1}$ if $x > a$, and $\mu_f(x) = 0$ if $x < a$. Then, for each $b > a$, $\mathrm{P}_{\mu_f}((b,+\infty]) = \int_b^{+\infty}\left(\frac{a}{x}\right)^{a+1} dx = \left(\frac{a}{b}\right)^a > \left(\frac{a}{b}\right)^{a+1} = \Pi_{\mu_f}((b,+\infty])$.

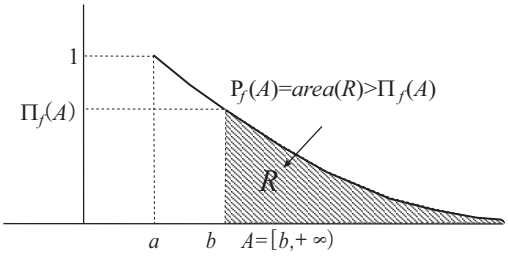


Figure 1: Density function of the Pareto distribution.

Cauchy distribution: As discussed previously, in the family of Cauchy density functions, $\mu(x) = \frac{1}{1+\pi^2 x^2}$ is also a possibility distribution. If $A = (-\infty, \sqrt{3}/\pi] \cup [\sqrt{3}/\pi, \infty)$, $\mathsf{L}^1(A) > 1$, and $\mathrm{P}_\mu(A) = 2\int_{\sqrt{3}/\pi}^{+\infty} \frac{1}{1+\pi^2 x^2} dx = \frac{1}{3}$; however $\Pi_\mu(A) = \sup\left\{\frac{1}{1+\pi^2 x^2}; x \in \left(-\infty, -\sqrt{3}/\pi\right] \cup \left[\sqrt{3}/\pi, +\infty\right)\right\} = \mu\left(\frac{\sqrt{3}}{\pi}\right) = \frac{1}{4}$. Thus $\mathrm{P}_\mu(A) > \Pi_\mu(A)$.

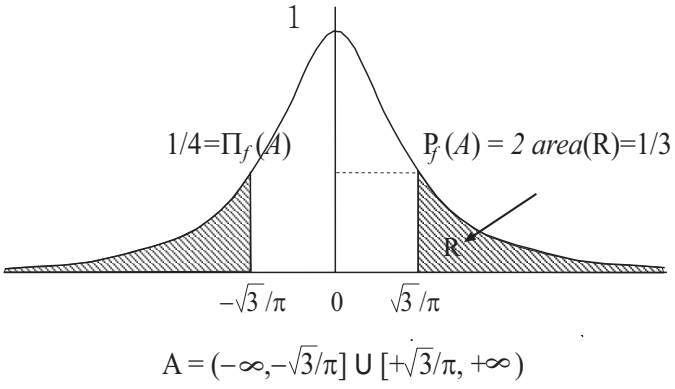


Figure 2: Density function of the Cauchy distribution.

4. A Survey of the Coherence in Some Notable Distributions Cases

In this section, we deal with the coherence between some notable distributions and the possibility measures generated by the density functions of the above distributions.

4.1. Coherence and Normal Distribution

Bearing in mind how important the normal distribution is, this section is given over to studying the coherence between the probability and possibility generated by its density function.

As discussed in section 3.1, it holds that the density functions of the distributions $N(\alpha,1/\sqrt{2\pi})$, with $\alpha \in \mathrm{R}$, are also possibility distributions; furthermore, they are the only ones within the normal family, as it should hold that

$$\sup_{x\in\mathrm{R}} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\alpha)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} = 1$$

then necessarily has to be $\sigma = 1/\sqrt{2\pi}$.

Theorem 4.1. Let f be the density function of the normal distribution $N(\alpha,1/\sqrt{2\pi})$, if Π_f and P_f are, respectively, the possibility and probability measures generated by f, then $\mathrm{P}_f(A) \le \Pi_f(A)$ for all $A \in \mathsf{B}$.

***Proof*:** It can be proven, without loss of generality, for $f(x) = e^{-\pi x^2}$ which corresponds to $N(0,1/\sqrt{2\pi})$, since any of the others is a translation of this one, and the relationship between probability and possibility will be the same. Firstly, we will check that $\mathrm{P}_f((-\infty,-a]\cup[a,+\infty)) \le \Pi_f((-\infty,-a]\cup[a,+\infty))$ for all $a \ge 0$. Indeed, if $a \ge 1$, it is $e^{-\pi x^2} \le x e^{-\pi x^2}$ for any $x \ge a$, then

$$\mathrm{P}_f((-\infty,-a]\cup[a,+\infty)) = 2\int_a^{+\infty} e^{-\pi x^2} dx \le 2\int_a^{+\infty} x e^{-\pi x^2} dx = \frac{e^{-\pi a^2}}{\pi} < f(a) = \Pi_f((-\infty,-a]\cup[a,+\infty)).$$

If $a \in [0,1)$, the function $G(a) = f(a) - \mathrm{P}_f((-\infty,-a]\cup[a,+\infty)) = e^{-\pi a^2} - 2\int_a^{+\infty} e^{-\pi x^2} dx$ is non-negative. Indeed, from $G'(a) = -2a\pi e^{-\pi a^2} - 2\frac{d}{da}\left(\int_0^{+\infty} e^{-\pi x^2} dx - \int_0^a e^{-\pi x^2} dx\right) = 2e^{-\pi a^2}(-a\pi + 1)$ it follows that G is increasing in $[0,1/\pi)$ and decreasing in $(1/\pi,1]$, moreover as $G(0) = 0$ and

$G(1) = e^{-\pi} - 2\int_1^{+\infty} e^{-\pi x^2} dx \ge e^{-\pi} - 2\int_1^{+\infty} e^{-\pi x} dx = e^{-\pi}\left(1 - \frac{2}{\pi}\right) > 0$, then $G(a) \ge 0$ for all $a \in [0,1]$.

Finally, let us see that $\mathrm{P}_f(A) \le \Pi_f(A)$ for any $A \in \mathsf{B}$. If 0 is an accumulation point of A, then $\mathrm{P}_f(A) \le 1 = f(0) = \Pi_f(A)$. If 0 is not an accumulation point of A, then there exists $a > 0$ such that $A \subset (-\infty,-a]\cup[a,+\infty)$ and a or $-a$ is either an element of A or an accumulation point of A. Therefore,

$$\mathrm{P}_f(A) \le \mathrm{P}_f((-\infty,-a]\cup[a,+\infty)) \le \Pi_f((-\infty,-a]\cup[a,+\infty)) = f(a) = \Pi_f(A). \square$$

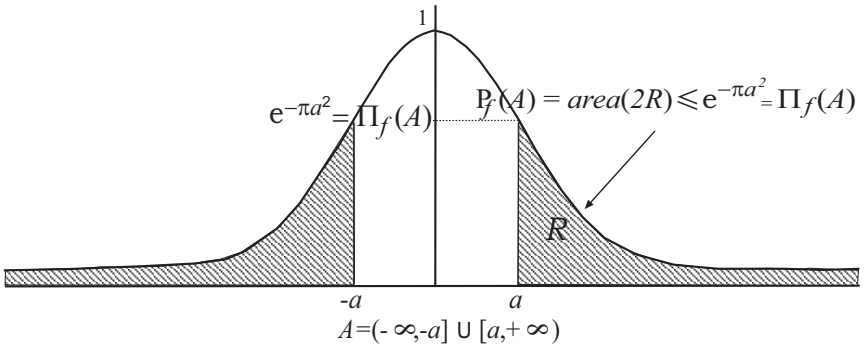


Figure 3: Density function of the normal distribution.

4.2. Coherence and Other Distributions

Even though important distributions, like the Cauchy distribution, do not generate coherent probabilities and possibilities as they are considered here, we can find other common distributions, apart from the important case of the normal distribution, which also generate coherent probabilities and possibilities. Let us take a look at some of these.

1. Uniform distribution, with density function $f(x)=1$ if $|x-a|\le 1/2$ and $f(x)=0$ if $|x-a|>1/2$. Trivially, $\mathrm{P}_f(A)\le\Pi_f(A)$ is satisfied for any $A\in\mathsf{B}$, since $\int_A f(x)dx=\mathsf{L}^1(A\cap[a-1/2,a+1/2])$.

2. Simpson's distribution, with density function $f(x)=1-|x-a|$ if $|x-a|\le 1$ and $f(x)=0$ if $|x-a|>1$. Let $A\in\mathsf{B}$, if a is an accumulation point of A, then $\Pi_f(A)=1\ge\mathrm{P}_f(A)$. If a is not an accumulation point of A, there exists $\varepsilon\in(0,1)$ such that $A\subset(-\infty,a-\varepsilon]\cup[a+\varepsilon,+\infty)$ and $\Pi_f(A)=f(a+\varepsilon)=f(a-\varepsilon)=1-\varepsilon$. Therefore,

$$\mathrm{P}_f(A)\le\mathrm{P}_f\big((-\infty,a-\varepsilon]\cup[a+\varepsilon,+\infty)\big)=(1-\varepsilon)^2<1-\varepsilon=\Pi_f(A).$$

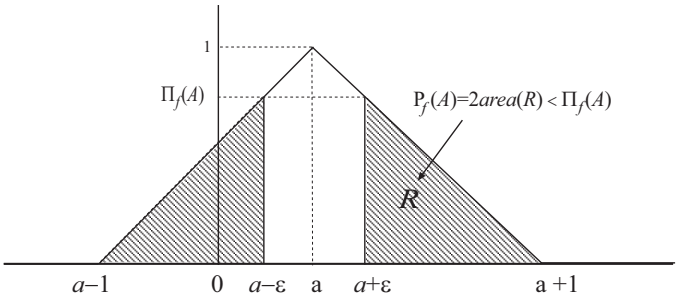


Figure 4: Density function of Simpson's distribution.

3. Exponential distribution, with density function $f(x)=e^{-x}$ if $x\ge 0$, and $f(x)=0$ if $x<0$. For each $a\in\mathsf{R}$:

- If $a\ge 0$, it is $\mathrm{P}_f([a,+\infty))=\int_a^{+\infty}e^{-x}dx=e^{-a}=\Pi_f([a,+\infty))$.
- If $a<0$, it is $\mathrm{P}_f([a,+\infty))=\int_a^{+\infty}e^{-x}dx=1=\Pi_f([a,+\infty))$.

For each $A\in\mathsf{B}$, there exists $a\in\mathsf{R}$ such that $A\subset[a,+\infty)$ and $a\in A$ or a is an accumulation point of A; thus, $\mathrm{P}_f(A)\le\mathrm{P}_f([a,+\infty))=\Pi_f([a,+\infty))=\Pi_f(A)$.

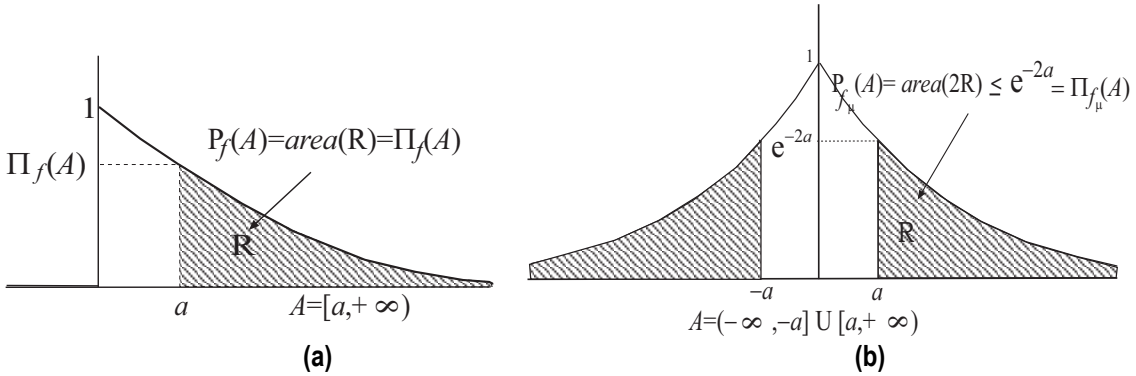


Figure 5: (a) Density function of the exponential distribution, and (b) density function $f_\mu(x)=e^{-2|x|}$

4. Finally, going back to the example in section 3.1, let $f_\mu(x)=e^{-2|x|}$ be the density function associated with the possibility distribution $\mu(x)=e^{-|x|}$. The probability and possibility measures generated by f_μ are also coherent. Indeed, for all $a\ge 0$,

$$\mathrm{P}_{f_\mu}\big((-\infty,a]\cup[a,+\infty)\big)=2\int_a^{+\infty}e^{-2x}dx=e^{-2a}=f_\mu(a)=\Pi_{f_\mu}\big((-\infty,a]\cup[a,+\infty)\big),$$

from which we can deduce, just as we did for the normal law, that for all $A\in\mathsf{B}$, $\mathrm{P}_{f_\mu}A)\le\Pi_{f_\mu}(A)$.

Conclusions and Further Works

In this paper, we have discussed the topic of the coherence between probability and possibility measures in the continuous case, that is, when these measures are defined on σ-algebras in the set R of real numbers. For this purpose, we have firstly found functions that are density functions and possibility distributions at the same time and, then we have studied the coherence between probability and possibility measures generated by the same density function. Moreover, the case of some significant distributions has been analysed.

The problem of finding the closest probability to a given possibility is an interesting open problem, technically more complex than in the finite case, in which it was successfully accomplished in [2].

Acknowledgements

This paper is supported by CICYT (Spain) under Project TIN 2005-08943-C02-01.

Bibliography

[1] G. Birkhoff. Lattice Theory. American Mathematical Society, Providence, 1973.

[2] E. Castiñeira, S. Cubillo and E. Trillas. On Possibility and Probability Measures in finite Boolean algebras. Soft-Computing, 7 (2), 89-96, 2002.

[3] H. Cramer. Métodos matemáticos de estadística. Aguilar, Madrid. (in Spanish), 1970.

[4] M. Delgado and S. Moral. On the concept of Possibility-Probability Consistency. Fuzzy Sets and Systems, 21, 311-318, 1987.

[5] D. Dubois and H. Prade. Théorie des possibilités. Applications à la représentation des connaissances en informatique. Masson, Paris. (in French), 1988.

[6] D. Dubois, H. T. Nguyen and H. Prade, Possibility Theory, Probability and Fuzzy Sets: Misunderstandings, Bridges and Gaps" in Fundamentals of Fuzzy Sets, D.Dubois and H. Prade Eds. Kluwer Academic Publishers, 2000.

[7] D. Dubois, H. Prade and S. Sandri. On possibility/probability transformations. Proceedings of 4th IFSA Conference, (Brussels), 50-53, 1991.

[8] J. A. Drakopoulus. Probabilities, possibilities and fuzzy sets. Fuzzy Sets and Systems, 75, 1-15, 1995.

[9] P. R. Halmos. Measure Theory. Springer-Verlag, New York, 1988.

[10] M. Loève. Probability theory I. Springer-Verlag, New York, 1977.

[11] E. Trillas. Sobre funciones de negación en la teoría de conjuntos difusos". Stochastica, 1 (3), 47-60 (in Spanish), 1979. Reprinted (English version) in Avances in Fuzzy Logic, (eds. S. Barro et altri) Universidad de Santiago de Compostela, 31-43, 1998.

[12] L.A. Zadeh. Fuzzy sets as a basis for a Theory of Possibility. Fuzzy Sets and Systems, 1, 3-28, 1978.

Authors' Information

Elena Castiñeira – Dept. Applied Mathematic. Computer Science School of Technical University of Madrid. Campus Montegancedo. 28660 Boadilla del Monte (Madrid). Spain; e-mail: ecastineira@fi.upm.es

Susana Cubillo – Dept. Applied Mathematic. Computer Science School of Technical University of Madrid. Campus Montegancedo. 28660 Boadilla del Monte (Madrid). Spain; e-mail: scubillo@fi.upm.es

Enric Trillas – Dept. Artificial Intelligence. Computer Science School of Technical University of Madrid. Campus Montegancedo. 28660 Boadilla del Monte (Madrid). Spain; e-mail: etrillas@fi.upm.es

RELATIONSHIP BETWEEN SELFCONTRADICTION AND CONTRADICTION IN FUZZY LOGIC*

Carmen Torres, Susana Cubillo, Elena Castineira

***Abstract**: This paper focuses on the study of self-contradiction as a particular case of contradiction between two fuzzy sets, and so, some self-contradiction degrees are defined from the contradiction degrees between two fuzzy sets. Furthermore, the definitions of measures of contradiction must be consistent with the idea that a disjunction with non-contradictory information remains non-contradictory, and a conjunction with contradictory information must remain contradictory; in this sense, some results are attained. Finally, contradiction in the compositional rule of inference is studied.*

***Keywords**: fuzzy sets, t-norm, t-conorm, strong fuzzy negations, contradiction, measures of contradiction, fuzzy relation, compositional rule of inference.*

Introduction and Preliminary Definitions

The study about contradiction was initiated by Trillas et al. in [7] and [8]. They introduced the concepts of both self-contradictory fuzzy set and contradiction between two fuzzy sets. Moreover, the need to study not only contradiction but also the degree of such contradiction is pointed out in [1] and [2], suggesting some measures for this purpose. In [5] new ways to measure the contradiction degree are obtained dealing with the problem from a geometrical point of view.

This paper begins, as a previous step, with a study on the relation between self-contradiction and contradiction between two fuzzy sets. Then, taking into account that the self-contradiction of a fuzzy set could be understand as the contradiction with itself, remembering some contradiction degrees defined in [5], the corresponding self-contradiction degrees for a fuzzy set, will be proposed, firstly, depending on a given strong negation, and later, without depending on any fixed negation.

In the following section, the problem of consistency with connectives will be managed. In fact, it is necessary to obtain non-contradictory knowledge, when the premises of non-contradictory information are relaxed. And, in a similar way, the information obtained adding contradictory premises, must also be contradictory.

Finally, last section will be devoted to study how contradictoriness is transmitted in the reasoning throughout the Compositional Rule of Inference.

Previously, we will remember some definitions and properties necessary throughout this article.

Definition 1.1 ([9]) A fuzzy set (FS) P, in the universe $X \neq \emptyset$, is a set given as $P=\{(x,\mu(x)): x \in X\}$ such that, for all $x \in X$, $\mu(x) \in [0,1]$, and where the function $\mu \in [0,1]^X$ is called membership function. We denote $\mathcal{F}(X)$ the set of all fuzzy sets on X.

Definition 1.2 $P \in \mathcal{F}(X)$ with membership function $\mu \in [0,1]^X$ is said to be a normal fuzzy set if $\text{Sup}\{\mu(x) : x \in X\}=1$.

Definition 1.3 A fuzzy negation (FN) is a non-increasing function $N: [0,1] \rightarrow [0,1]$ with $N(0)=1$ and $N(1)=0$. Moreover, N is a strong fuzzy negation if the equality $N(N(y))=y$ holds for all $y \in [0,1]$.

* This work is supported by cicyt (Spain) under project tin 2005-08943-c02-001.

N is a strong negation if and only if, there is an order automorphism g in the unit interval (that is, g:[0,1]→ [0,1] is an increasing continuous function with g(0)=0 and g(1)=1) such that $N(y)=g^{-1}(1-g(y))$ for all $y \in [0,1]$ (see [5]); from now on, let us denote $N_g=g^{-1}(1-g)$. Furthermore, the only fixed point of N_g is $n_g=g^{-1}(1/2)$.

Definition 1.4 ([4]) A function T: [0,1] x [0,1] → [0,1] is said to be a *t-norm* if it is a commutative, associative and non-decreasing in both variables function verifying T(y,1)=y for all $y \in [0,1]$.

Definition 1.5 ([4]) A function S: [0,1] X [0,1] → [0,1] is said to be a *t-conorm* if it is a commutative, associative and non-decreasing in both variables function verifying S(y,0)=y for all $y \in [0,1]$.

Definition 1.6 ([7]) Given $\mu, \sigma \in [0,1]^X$ and a strong negation N_g, then μ and σ are N_g-contradictory if and only if $\mu(x) \leq N_g(\sigma(x))$, for all $x \in X$. This inequality is equivalent to $g(\mu(x))+g(\sigma(x)) \leq 1$, for all $x \in X$.

Definition 1.7 ([7]) Given $\mu \in [0,1]^X$ and a strong negation N_g, μ is said to be N_g-self-contradictory if and only if $\mu(x) \leq N_g(\mu(x))$, for all $x \in X$. This inequality is equivalent to $g(\mu(x)) \leq 1/2$, for all $x \in X$.

Therefore, the definition of N_g-self-contradictory fuzzy set is a particular case from that of N_g-contradictory fuzzy sets, where the two sets are the same.

Definition 1.8 $\mu, \sigma \in [0,1]^X$ are contradictory if they are N_g-contradictory regarding some strong FN N_g. And μ is self-contradictory if it is N_g-self-contradictory for some strong FN N_g. This condition is equivalent to the fact that μ is not a normal fuzzy set ($Sup\{\mu(x) : x \in X\}<1$). Again, the definition of self-contradiction is a particular case from that of contradiction.

Self-contradiction and Contradiction between Two FS

The goal of this section is study if there exists some direct relation between the self-contradiction of two fuzzy sets and the contradiction between them. In fact, we have the following properties.

Proposition 2.1 Given $\mu, \sigma \in [0,1]^X$, if μ and σ are N_g–self-contradictory, for some strong fuzzy negation N_g, then μ, σ are N_g–contradictory.

The following example shows that reciprocal is not true.

Example 2.2 Let us consider the set X={x,y} and $\mu, \sigma \in [0,1]^X$ such that $\mu(x)=3/4$, $\mu(y)=0$ and $\sigma(x)=0$, $\sigma(y)=3/4$; and the standard negation $N_s=1-id$. Then $\mu(x)+\sigma(x)=3/4$ and $\mu(y)+\sigma(y)=3/4$ and so μ, σ are N_s-contradictory between them. Nevertheless μ and σ are not N_s-self-contradictory ($\mu(x)>1/2$ and $\sigma(y)>1/2$).

Proposition 2.3 Given $\mu, \sigma \in [0,1]^X$, if μ and σ are self-contradictory, then μ, σ are contradictory.

Proof: As $\mu, \sigma \in [0,1]^X$ are self-contradictory there exist order automorphisms g and g' on [0,1], such that $g(\mu(x)) \leq 1/2$ and $g'(\sigma(x)) \leq 1/2$ for all $x \in X$. Let us take the following function on [0,1], g''=Min{g,g'}. This function is continuous because g and g' are continuous; g''(0)=0, g''(1)=1. Let us see that g'' is increasing: let $y_1, y_2 \in [0,1]$ be such that $y_1<y_2$, then $g''(y_1)=Min\{g(y_1),g'(y_1)\} \leq g(y_1)<g(y_2)$ and $g''(y_1)=Min\{g(y_1),g'(y_1)\} \leq g'(y_1)<g'(y_2)$. Therefore $g''(y_1)<Min\{g(y_2),g'(y_2)\}=g''(y_2)$. So g'' is an order automorphism in the unit interval, and moreover $g''(\mu(x))+g''(\sigma(x)) \leq g(\mu(x))+g'(\sigma(x)) \leq ½+½ =1$ for all $x \in X$. Therefore, μ, σ are $N_{g''}$-contradictory and so μ, σ are contradictory.

Newly, reciprocal is not true as the following example shows.

Example 2.4 Let us consider the set $X = \{x_n\}_{n \in IN} \cup \{y_n\}_{n \in IN}$ and $\mu, \sigma \in [0,1]^X$ such that $\mu(x_n)=n/(n+1)$, $\mu(y_n)=1/(n+2)$ and $\sigma(x_n)=1/(n+2)$, $\sigma(y_n)=n/(n+1)$. Then $\mu(x_n)+\sigma(x_n)=(n^2+3n+1)/(n^2+3n+2)<1$ and $\mu(y_n)+\sigma(y_n)=(n^2+3n+1)/(n^2+3n+2)<1$ and it follows that μ, σ are N_s-contradictory between them. Nevertheless μ and σ are not self-contradictory ($Sup\mu(x)=1$ and $Sup\sigma(x)=1$).

N_g-self-contradiction and Self-contradiction Degrees

Clearly, self-contradiction of a fuzzy set could be viewed as contradiction of the set with itself. Taking this into account, the degrees of contradiction defined in above papers, provide us the respective degrees of self-contradiction, as in this section is shown.

In [5], some functions were defined as a model to determine different degrees of N_g-contradiction between two fuzzy sets.

Definition 3.1 Given $\mu, \sigma \in [0,1]^X$ and N_g a strong FN , we define the following contradiction measure functions:

i) $C_1^{N_g}(\mu,\sigma) = \text{Max}\left(0, \inf_{x\in X}\left(N_g(\sigma(x)) - \mu(x)\right)\right)$

ii) $C_2^{N_g}(\mu,\sigma) = \text{Max}\left(0, \inf_{x\in X}\left(N_g(\mu(x)) - \sigma(x)\right)\right)$

iii) $C_3^{N_g}(\mu,\sigma) = \text{Max}\left(0, 1 - \sup_{x\in X}\left(g(\mu(x)) + g(\sigma(x))\right)\right)$

iv) $C_4^{N_g}(\mu,\sigma) = \dfrac{d\left(X_{\mu\sigma}, R_{N_g}\right)}{d\left((0,0), R_{N_g}\right)}$, where *d* is the Euclidean distance, $X_{\mu\sigma} = \{(\mu(x),\sigma(x)) : x \in X\}$ and

$R_{N_g} = \left\{(y_1, y_2) \in [0,1]^2 : N_g(y_1) < y_2\right\}$ is the region free of contradiction. Therefore,

$d\left(X_{\mu\sigma}, R_{N_g}\right) = \text{Inf}\left\{d\left((\mu(x),\sigma(x)), (y_1,y_2)\right) : x \in X, (y_1,y_2) \in R_{N_g}\right\}$ and

$d\left((0,0), R_{N_g}\right) = \text{Inf}\left\{d\left((0,0),(y_1,y_2)\right) : (y_1,y_2) \in R_{N_g}\right\}$.

Another new function could serve as definition of contradiction degree:

v) $C_5^{N_g}(\mu,\sigma) = N_g\left(1 - \dfrac{d\left(X_{\mu\sigma}, R_{N_g}\right)}{d\left((0,0), R_{N_g}\right)}\right) = N_g\left(1 - C_4^{N_g}(\mu,\sigma)\right)$

For the standard negation $N_s(y)=1-y$ the equality $C_5^{N_s}(\mu,\sigma) = C_i^{N_s}(\mu,\sigma)$ is verified, for all i=1,2,3,4.

And for N_g with $g(y)=y^2$ is $C_5^{N_g}(\mu,\sigma) = \sqrt{1 - \left(1 - C_4^{N_g}(\mu,\sigma)\right)^2} = \sqrt{C_3^{N_g}(\mu,\sigma)}$.

Considering N_g-self-contradiction as a particular case of N_g-contradiction between two fuzzy sets with $\mu=\sigma$, the N_g-contradiction degrees given in 3.1 are turned into the following N_g-self-contradiction degrees:

i) $C_{s1}^{N_g}(\mu) = C_1^{N_g}(\mu,\mu) = \text{Max}\left(0, \inf_{x\in X}\left(N_g(\mu(x)) - \mu(x)\right)\right) = C_2^{N_g}(\mu,\mu) = C_{s2}^{N_g}(\mu)$

ii) $C_{s3}^{N_g}(\mu) = C_3^{N_g}(\mu,\mu) = \text{Max}\left(0, 1 - 2\sup_{x\in X}\left(g(\mu(x))\right)\right)$

This measure of N_g-self-contradiction, $C_{s3}^{N_g}(\mu)$, was also defined in [3].

iii) $C_{s4}^{N_g}(\mu) = C_4^{N_g}(\mu,\mu) = \dfrac{d\left(X_{\mu\mu}, R_{N_g}\right)}{d\left((0,0), R_{N_g}\right)} = \dfrac{d\left(\left(\sup_{x\in X}\mu(x), \sup_{x\in X}\mu(x)\right), R_{N_g}\right)}{d\left((0,0), R_{N_g}\right)}$

iv) $C_{s5}^{N_g}(\mu) = C_5^{N_g}(\mu,\mu) = N_g\left(1 - \dfrac{d\left(X_{\mu\mu}, R_{N_g}\right)}{d\left((0,0), R_{N_g}\right)}\right) = N_g\left(1 - \dfrac{d\left(\left(\sup_{x\in X}\mu(x), \sup_{x\in X}\mu(x)\right), R_{N_g}\right)}{d\left((0,0), R_{N_g}\right)}\right)$

Proposition 3.2 Let N_g be a Yager strong negation, $N_g(y)=\left(1-y^r\right)^{\frac{1}{r}}$, with 0<r≤2 or $N_\lambda(y)=\frac{1-y}{1+\lambda y}$ with λ>0 a Sugeno strong negation, then for all $\mu \in [0,1]^X$ it is:

$$d\left(X_{\mu\mu}, R_{N_g}\right)=\begin{cases} 0 \quad \text{if} \quad \sup_{x\in X}\mu(x)\geq n_g \\ d\left(\left(\sup_{x\in X}\mu(x), \sup_{x\in X}\mu(x)\right),\left(n_g,n_g\right)\right) \text{ in other case} \end{cases}$$

and $d\left((0,0),R_{N_g}\right)=d\left((0,0),\left(n_g,n_g\right)\right)=\sqrt{2n_g^2}=\sqrt{2}n_g$

Consequently,

$$C_{s4}^{N_g}(\mu)=\text{Max}\left(0,1-\frac{\sup_{x\in X}(\mu(x))}{n_g}\right) \quad \text{and} \quad C_{s5}^{N_g}(\mu)=N_g\left(1-\text{Max}\left(0,1-\frac{\sup_{x\in X}(\mu(x))}{n_g}\right)\right)=N_g\left(\text{Min}\left(1,\frac{\sup_{x\in X}(\mu(x))}{n_g}\right)\right)$$

this last measure of N_g-self-contradiction, $C_{s5}^{N_g}(\mu)$, was introduced in [1].

However, a similar result is not true for all strong fuzzy negation, as the following example shows.

Example 3.3 Let N_g be with $g(y)=y^3$; in this case $d\left((0,0),R_{N_g}\right)=1$ and $d\left((0,0),(n_g,n_g)\right)=2^{\frac{1}{6}}>1$ $\left(n_g=\left(\tfrac{1}{2}\right)^{\frac{1}{3}}\right)$.

Given $\mu\in[0,1]^X$ such that $X_{\mu\mu}=\left\{\left(\frac{1}{10},\frac{1}{10}\right)\right\}$ it is $C_{s4}^{N_g}(\mu)=\frac{d\left(\left(\frac{1}{10},\frac{1}{10}\right),R_{N_g}\right)}{d\left((0,0),R_{N_g}\right)}=0.9$ and however $Max\left(0,1-\frac{Sup\mu(x)}{n_g}\right)=1-\frac{\frac{1}{10}}{n_g}=0.874$. Moreover, $C_{s5}^{N_g}(\mu)=N_g\left(1-C_{s4}^{N_g}(\mu)\right)=N_g(0.1)\neq N_g\left(\frac{\frac{1}{10}}{n_g}\right)$.

Until now, we have managed contradiction depending on a fixed strong negation. We continue studying contradiction without depending on any fixed negation.

The following degrees of contradiction, between two fuzzy sets, were given in [5].

Definition 3.4 Given $\mu, \sigma \in [0,1]^X$, we have the following contradiction measure functions:

i) $C_1(\mu,\sigma)=\text{Min}\left(d\left(X_{\mu\sigma},L_1\right),d\left(X_{\mu\sigma},L_2\right)\right)$, denoting L_1 the line $y_1=1$ and L_2 the line $y_2=1$.

ii) $C_2(\mu,\sigma)=0$ if there exists $\{x_n\}_{n\in N}\subset X$ such that $\lim_{n\to\infty}\{\mu(x_n)\}=1$ or $\lim_{n\to\infty}\{\sigma(x_n)\}=1$ and, in other case

$$C_2(\mu,\sigma)=1-\frac{\sup_{x\in X}\left(\mu(x)+\sigma(x)\right)}{2}=\frac{d_1\left(X_{\mu\sigma},(1,1)\right)}{d_1((0,0),(1,1))}$$, being d_1 the reticular distance.

iii) $C_3(\mu,\sigma)=0$ if there exists $\{x_n\}_{n\in N}\subset X$ such that $\lim_{n\to\infty}\{\mu(x_n)\}=1$ or $\lim_{n\to\infty}\{\sigma(x_n)\}=1$, and, in other case

$$C_3(\mu,\sigma)=\frac{d\left(X_{\mu\sigma},(1,1)\right)}{d((0,0),(1,1))}.$$

Newly, considering self-contradiction as a particular case of contradiction between two fuzzy sets with $\mu=\sigma$, the contradiction degrees given in 3.4 are turned into the following self-contradiction degrees:

i) $C_{s1}(\mu)=C_1(\mu,\mu)=\inf_{x\in X}(1-\mu(x))=1-\sup_{x\in X}(\mu(x))=C_2(\mu,\mu)=C_{s2}(\mu)$, the measure of self-contradiction $1-\sup_{x\in X}(\mu(x))$ was also introduced in [1].

ii) $$C_{s3}(\mu)=C_3(\mu,\mu)=\frac{d\left(X_{\mu\mu},(1,1)\right)}{\sqrt{2}}=\frac{d\left(\left(\underset{x\in X}{Sup}\,\mu(x),\underset{x\in X}{Sup}\,\mu(x)\right),(1,1)\right)}{\sqrt{2}}=\frac{\sqrt{2}\left(1-\underset{x\in X}{Sup}\,\mu(x)\right)}{\sqrt{2}}=C_{s1}(\mu)$$

Contradiction Degrees and Connectives

In this section, the problem of consistency with connectives, will be managed. In fact, if we have non-contradictory premises, and these ones are relaxed (by an OR connective, that is, by means of a t-conorm) , then the new information must also be non-contradictory. And, in a similar way, if we have contradictory premises, and we add new information (by an AND connective, of a t-norm), the information must also be contradictory.

The following results handle this subject.

Proposition 4.1 Given $\mu \in [0,1]^X$, if μ is not N_g–self-contradictory, for a strong fuzzy negation N_g, then $S(\mu,\sigma)$ is not N_g–self-contradictory, for all S t-conorm and for all $\sigma \in [0,1]^X$.

In particular, if μ, $\sigma \in [0,1]^X$ are not N_g-contradictory then μ or σ is not N_g-self-contradictory and subsequently $S(\mu,\sigma)$ is not N_g-self-contradictory, for all S t-conorm.

Proposition 4.2 Given $\mu \in [0,1]^X$, if μ is not self-contradictory ($Sup\{\mu(x): x \in X\}=1$), then $S(\mu,\sigma)$ is not self-contradictory for all S t-conorm and for all $\sigma \in [0,1]^X$ ($Sup\{S(\mu(x),\sigma(x)): x \in X\}=1$).

In particular, if μ, $\sigma \in [0,1]^X$ are not contradictory then μ or σ is not self-contradictory and subsequently $S(\mu,\sigma)$ is not self-contradictory, for all S t-conorm.

Then, it is obtained that the disjunction with non-contradictory information provides non-self-contradictory information.

In addition, the definitions of measures of contradiction also must be consistent with the idea that a disjunction with non-contradictory information remains non-contradictory. Indeed, we have the following result:

Proposition 4.3 Given $C_i^{N_g}$ with i=1,2,3,4,5 (or C_i with i=1,2,3), and $\mu,\sigma \in [0,1]^X$, if $C_i^{N_g}(\mu,\sigma)=0$ (or $C_i(\mu,\sigma)=0$), then for any t-conorm S it holds that $C_{si}^{N_g}(S(\mu,\sigma))=0$ (or $C_{si}(S(\mu,\sigma))=0$).

In general, for all weak measure of self-contradiction (that is, $C:[0,1]^X \to [0,1]$ such that $C(\mu_\emptyset)=1$, $C(\mu)=0$ if μ normal and C anti-monotonic, as defined in [3]) it is verified that: if $C(\mu)=0$ then $C(S(\mu,\sigma))=0$ for all S t-conorm and $\sigma \in [0,1]^X$. Furthermore, for all weak measure of contradiction (that is, $C:[0,1]^X \times [0,1]^X \to [0,1]$ such that $C(\mu_\emptyset,\mu_\emptyset)=1$, $C(\mu,\mu)=0$ if μ normal and C symmetric and anti-monotonic [3]) it is verified that: if $C(\mu,\sigma)=0$ then $C(S(\mu,\sigma),S(\mu,\sigma))=0$ for all S t-conorm.

Proposition 4.4 Given $\mu \in [0,1]^X$, if μ is N_g–self-contradictory, for some strong fuzzy negation N_g, then $T(\mu,\sigma)$ is N_g–self-contradictory, for all t-norm T and for all $\sigma \in [0,1]^X$.

Moreover, if μ, $\sigma \in [0,1]^X$ are N_g-contradictory then $T(\mu,\sigma)$ is N_g-self-contradictory, for all t-norm T.

Proposition 4.5 Given $\mu \in [0,1]^X$, if μ is self-contradictory ($Sup\{\mu(x): x \in X\}<1$), then $T(\mu,\sigma)$ is self-contradictory for all t-norm T and for all $\sigma \in [0,1]^X$ ($Sup\{T(\mu(x),\sigma(x)): x \in X\}<1$).

Moreover, if μ, $\sigma \in [0,1]^X$ are contradictory then they are N_g-contradictory, for some strong fuzzy negation N_g, and consequently $T(\mu,\sigma)$ is N_g-self-contradictory, and therefore $T(\mu,\sigma)$ self-contradictory, for all t-norm T.

Then, it is obtained that the conjunction with contradictory information provides self-contradictory results.

Similarly, definitions of measures of contradiction also must be consistent with the idea that a conjunction with contradictory information must remain contradictory. Indeed, we have the following result:

Proposition 4.6 Given $C_i^{N_g}$ with i=1,2,3,4,5 (C_i with i=1,2,3), and $\mu,\sigma \in [0,1]^X$, if $C_i^{N_g}(\mu,\sigma)>0$ (or $C_i(\mu,\sigma)>0$), then for any t-norm T it holds that $C_{si}^{N_g}(T(\mu,\sigma))>0$ (or $C_{si}(T(\mu,\sigma))>0$).

In particular, if T is a t-norm of the Lukasiewicz's family, that is, $T = g^{-1} \circ W \circ (g \times g)$, with W(x,y)=Max(0,x+y-1), where *g* is an order automorphism in the unit interval, it holds that if $C_i^{N_g}(\mu,\sigma) > 0$ then $C_{si}^{N_g}(T(\mu,\sigma)) = 1$, or equivalently, $T(\mu,\sigma)=\mu_{\emptyset}$.

In general, for all weak measure of self-contradiction it is verified that: if $C(\mu)>0$ then $C(T(\mu,\sigma))>0$ for all t-norm T and $\sigma \in [0,1]^X$. In a similar way, for all weak measure of contradiction it is verified that: if $C(\mu,\sigma)>0$ then $C(T(\mu,\sigma),T(\mu,\sigma))>0$ for all t-norm T.

Contradiction in Inference

For inference purposes in both classical and fuzzy logic, neither the information itself should be contradictory, nor should any of the items of available information contradict each other. In order to avoid these troubles in fuzzy logic, it is necessary to study self-contradiction and contradiction in the fuzzy inference systems.

The Compositional Rule of Inference ([4]) is based on the Zadeh's Logical Transform:

$$T_J(\mu)(y) = \underset{x \in X}{Sup}\, T(\mu(x), J(x,y))$$

Where $J : X \times X \to [0,1]$ is a given fuzzy relation, T a t-norm and $\mu \in [0,1]^X$ any fuzzy set. We aim to study the relationship between the contradiction in the input μ and the contradiction in the output $T_J(\mu)$. Also, we want to research the relationship between the degrees of contradiction of the input μ and the degrees of contradiction of the output $T_J(\mu)$.

Proposition 5.1 Given $\mu \in [0,1]^X$, if μ is N_g–self-contradictory (or self-contradictory), then $T_J(\mu)$ is N_g–self-contradictory (or self-contradictory), for all t-norm T and all fuzzy relation J.

Reciprocals are not true, as the following example shows.

Example 5.2 Let us consider the set $X = [0,1]$, $\mu \in [0,1]^X$ such that $\mu(x)=1-x$, $J(x,y)=Min(x,y)$ and $T(x,y)=Min(x,y)$ for all x, y $\in$ [0,1]. Therefore, $T_J(\mu)(y) = Min\left(\frac{1}{2}, y\right)$ and thus $\underset{y \in [0,1]}{Sup}\, T_J(\mu)(y) = \frac{1}{2}$. Then, $T_J(\mu)$ is N_s-self-contradictory and self-contradictory but μ is neither N_s-self-contradictory nor self-contradictory ($\underset{x \in [0,1]}{Sup}\, \mu(x) = 1$).

Moreover, if μ is N_g-self-contradictory (or self-contradictory) then, from proposition, 5.1 and 2.1 (or 2.3), it is obtained that μ and $T_J(\mu)$ are N_g-contradictory (or contradictory) between them, for all t-norm T.

Proposition 5.3 Given $\mu \in [0,1]^X$ and a reflexive fuzzy relation J, (that is, $J(x,x)=1\ \forall x \in X$), μ is N_g–self-contradictory (or self-contradictory) if and only if $T_J(\mu)$ is N_g–self-contradictory (or self-contradictory), for all t-norm T.

In addition, if J is a reflexive fuzzy relation, then μ is N_g–self-contradictory (or self-contradictory) if and only if μ and $T_J(\mu)$ are N_g–contradictory (or contradictory) between them, for all t-norm T.

Now, let us study if there is some relationship between the contradiction measures of the input μ and those of the inference output $T_J(\mu)$.

Proposition 5.4 Given a reflexive fuzzy relation J and $\mu \in [0,1]^X$ such that $C(\mu)=0$ then $C(T_J(\mu))=0$, for all C weak contradiction measure.

If J is not reflexive the last proposition is not true, in general, as the following example shows.

Example 5.5 Let us consider X, μ, T and J as in the example 5.2; J(x,x)=Min(x,x)=x. Then, J is not reflexive. Moreover, μ (x)=1-x is a normal fuzzy set, so C(μ)=0 for all C weak contradiction measure (in particular for C_{si}), however $C_{si}(T_J(\mu)) = 1 - \underset{y\in[0,1]}{Sup}\, T_J(\mu)(y) = \frac{1}{2}$

Also, if J is a reflexive fuzzy relation it is $\mu \le T_J(\mu)$ and therefore $C(\mu) \ge C(T_J(\mu))$, for all weak contradiction measure C.

Finally, let us see that for the N_g-self-contradiction and self-contradiction degrees considered in this paper, the equality between the contradiction degree of the input μ and the contradiction degree of the output $T_J(\mu)$ is verified; being of interest, for it, to consider a previous proposition.

Proposition 5.6 Given $\mu\in[0,1]^X$, for all J fuzzy relation and all t-norm T, the inequality $\underset{x\in X}{Sup}\, T_J(\mu)(x) \le \underset{x\in X}{Sup}\, \mu(x)$ holds.

Consequently, if J is reflexive, it is $\underset{x\in X}{Sup}\, T_J(\mu)(x) = \underset{x\in X}{Sup}\, \mu(x)$.

Corollary 5.7 Given $\mu\in[0,1]^X$, if J is a reflexive fuzzy relation it is $C_{si}^{N_g}(\mu) = C_{si}^{N_g}(T_J(\mu))$ and $C_{si}(\mu) = C_{si}(T_J(\mu))$ for all i and for all t-norm T, being $C_{si}^{N_g}(\mu)$ and $C_{si}(\mu)$ the N_g-self-contradiction and self-contradiction degrees given in definition 3.1 and 3.4.

Conclusion

This paper deepens on the study of contradictoriness in fuzzy sets. New self-contradiction measures have been obtained by means of contradiction measures between two fuzzy sets when the two sets are the same.

Furthermore, some results about the propagation of contradictoriness throughout connectives (t-norms and t-conorms) have been attained. As it was expected, these results are coherent with the human intuition.

Finally, the compositional rule of inference, commonly used in reasoning processes, is studied from the point of view of the contradiction. Results prove non-contradictoriness of input, assure the same property in the output.

Bibliography

[1] E. Castiñeira, S. Cubillo and S. Bellido. Degrees of Contradiction in Fuzzy Sets Theory. Proceedings IPMU'02, 171-176. Annecy (France), 2002.

[2] E. Castiñeira, S. Cubillo. and S. Bellido. Contradicción entre dos conjuntos. Actas ESTYLF'02, 379-383. León (Spain), 2002, (in Spanish).

[3] S. Cubillo and E. Castiñeira. Measuring contradiction in fuzzy logic. International Journal of General Systems, Vol. 34, Nº1, 39-59, 2005.

[4] H. T. Nguyen and E. A. Walker. A first course in fuzzy logic. CRC Press, 1997.

[5] C. Torres, E. Castiñeira S Cubillo and V. Zarzosa. A geometrical interpretation to define contradiction degrees between two fuzzy sets. International Journal "Information Theories and Applications". 2005.

[6] E. Trillas. Sobre funciones de negación en la teoría de conjuntos difusos. Stochastica III/1, 47-60, 1979 (in Spanish). Reprinted (English version) (1998) in Avances of Fuzzy Logic. Eds. S. Barro et altr, 31-43.

[7] E. Trillas, C. Alsina and J. Jacas. On Contradiction in Fuzzy Logic. Soft Computing, 3(4), 197-199, 1999.

[8] E. Trillas and S. Cubillo. On Non-Contradictory Input/Output Couples in Zadeh's CRI. Proceedings NAFIPS, 28-32. New York, 1999.

[9] L. A. Zadeh. Fuzzy Sets. Inf. Control, volume 20, 301-312, 1965.

Authors' Information

Carmen Torres – Dept. Applied Mathematic. Computer Science School of University Politécnica of Madrid. Campus Montegancedo. 28660 Boadilla del Monte (Madrid). Spain; e-mail: ctorres@fi.upm.es

Elena Castineira – Dept. Applied Mathematic. Computer Science School of University Politécnica of Madrid. Campus Montegancedo. 28660 Boadilla del Monte (Madrid). Spain; e-mail: ecastineira@fi.upm.es

Susana Cubillo – Dept. Applied Mathematic. Computer Science School of University Politécnica of Madrid. Campus Montegancedo. 28660 Boadilla del Monte (Madrid). Spain; e-mail: scubillo@fi.upm.es

RECURSIVE MATRICES FOR AN INFORMATION RETRIEVAL PROBLEM

Adriana Toni, Juan Castellanos, Jose Joaquin Erviti

Abstract *Let $v_1..v_n$ be variables storing values from an arbitrary commutative semigroup S. We are interested in the study and design of data structures for implementing the following operations. The operation update(i,x) increments the value of v_i by x, and the operation retrieve(i,j) returns $v_i + \ldots + v_j$. Our interest centers upon improving the average complexity of the operations. We define matrices representing a solution for the problem inside a matricial model of computation. We achieve a constant average complexity for the set of update and retrieve operations.*

Keywords*: data structures, models of computation, analysis of algorithms and problem complexity*

Introduction

Let $v_1..v_n$ be variables storing values from an arbitrary commutative semigroup S. We desire to execute the following operations on these variables:

a) retrieve(i,j) returns $v_i + \ldots + v_j \qquad \forall\ 1 \le i \le j \le n,$

b) update(i,x): $v_i := v_i + x \qquad \forall\ 1 \le i \le n, \quad x \in S$

This problem is known as the range query problem of size n.

We can organize the variables as an array V of length n, and implement the operations as above. In this case, the complexity of executing an update operation is constant meanwhile the worst case complexity of a retrieve is linear on n.

Our interest centers upon improving the average complexity of the operations assuming that each one of them is selected with the same probability.

We can use different data structures involving a different number of variables storing values in the semigroup, and provide the corresponding algorithms to implement the update and retrieve operations, and still be solving the same computational problem.

A matricial model for the study of the range query problem has been defined, relative to which computational complexity is assesed (see [6])

The model comprises all programs verifying:

a) A set of variables $Z=\{z_1,z_2..z_m\}$ is maintained.
b) Retrieve(i,j) is performed by adding up a subset of these variables.
c) Update(j,x) is performed by incrementing a subset of these variables by amounts which depend linearly on x.

The model defined in [6] consists of triples $<R,U,Z>$ where

$Z = \{z_1 \ldots z_m\}$ is a set of variables storing values on an arbitrary semigroup S, $R=(r_i,j)$ is a zero-one matrix of dimension $\frac{n(n+1)}{2} \times m$ and $U=(u_i,j)$ is a zero-one matrix of dimension $m \times n$. Each row of R describes the subset of variables of Z which have to be added to execute one of the retrieve operations, and the i-*th* column of U describes the subset of such variables which have to incremented to execute an $\mathrm{update}(i,x)$. So, a pair of R and U matrices describes a solution for the range query problem of size n (m, the number of required program variables, may change although if has to be greater or equal n). Associated with a triple $<R,U,Z>$, the programs implementing the operations are defined as follows.

Definition 1

Given a triplet <Z,R,U> within the matrix model for the range query problem of size n, with $Z = \{z_1 \ldots z_m\}$, *then the update and retrieve operations must be implemented through the following programs:*

- *update(j,x): for l:=1 to m do* $[z_1 \leftarrow z_l + u_{l,j}\, x]$
- *Retrieve(i,j): output* $\sum_{l=1}^{m} r_{k,l} z_l$, *where* $k = \sum_{s=0}^{i-2} (n-s) + (j-i+1)$

The following proposition establishes a condition on *R, U* that entails reworking the programs defined above.

Proposition 2

Let H be the matrix of dimensions $\frac{n(n+1)}{2} \times n$ *defined by:*

$$H_{ij} = \begin{cases} 1 & l \le j \le l + (i - w_{l-1} - 1) \\ 0 & otherwise \end{cases}$$

where

$w_k = \sum_{l=0}^{k-1} (n-l)$, $k=0 \ldots n$

$i \in \{(w_{l-1}+1) \ldots w_l\}$, $l=1 \ldots n$

Then the programs given in Definition 1 represent a solution for the range query problem of size n if and only if $R \times U = H$.

Remark 3

Note that if T^n *is the triangular matrix of dimensions* $n \times n$ *consisting of 0s above the main diagonal and 1s along and below the main diagonal,*

$$T=(t_{ij})_{i,j=1..n} \quad \text{with} \quad t_{ij} = \begin{cases} 1 & i \ge j \\ 0 & i \prec j \end{cases}$$

then the following statement holds for all k = 0…(n-1),

$$H_{[w_k+1..w_{k+1}],[k+1..n]} = T^{n-k}$$
$$H_{[w_k+1..w_{k+1}],[1..k]} = 0$$

where $w_i = \sum_{s=0}^{i-1}(n-s)$ *for i = 0…n, and* $H_{[i\ldots j],[r\ldots s]}$ *is the box of matrix H composed of the rows in the interval [i…j] and the columns in the interval [r…s].*

Hence, for any problem size n, the corresponding matrix H^n *can be expressed as a function of the matrix T of different dimensions as:*

$$H = \begin{pmatrix} T^n \\ \overrightarrow{T^{n-1}} \\ \vdots \\ \overrightarrow{T^{n-(n-1)}} \end{pmatrix}$$

where

$$\overrightarrow{T^{n-i}} = \begin{pmatrix} 0 & \ldots & \ldots & 0 & \\ \vdots & & & \vdots & \\ \vdots & & & \vdots & \left| T^{n-i} \right. \\ 0 & \ldots & \ldots & 0 & \end{pmatrix} \qquad \text{i=1..(n-1)}$$

Example 4

The matrix H for the range query problem of sizes n=2 and n=4, denoted H^2 *and* H^4*, respectively, is shown below. Lines have been added to highlight the logical division into boxes*

$$H^2 = \left(\begin{array}{cc} 1 & 0 \\ \hline 1 & 1 \\ \hline 0 & 1 \end{array}\right), \qquad H^4 = \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ \hline 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ \hline 0 & 1 & 1 & 1 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 1 & 1 \\ \hline 0 & 0 & 0 & 1 \end{array}\right)$$

In the following we define the complexity associated with the operations within the matrix model.

Definition 5

Given a triplet <Z,R,U> that solves the range query problem of size n within the matrix model, with Z = $(z_1\ldots z_m)$*,* $R = (r_{i,j})_{i=1\ldots\frac{n(n+1)}{2},\, j=1\ldots m}$, $U = (u_{i,j})_{i=1\ldots m,\, j=1\ldots n}$*, let the complexity associated with the Retrieve (i, j) operation be defined as:*

$$|\{r_{kl} / r_{kl} \neq 0 \wedge (1 \leq l \leq m)\}| \quad \text{where} \quad k = (j-i+1) + \sum_{s=0}^{i-2}(n-s)$$

and the complexity associated with the Update (j, x) operation be defined as:

$$|\{u_{lj} / u_{lj} \neq 0 \wedge (1 \leq l \leq m)\}|$$

Let *m* be the number of columns of *R* and of rows of *U.* The average complexity of *Update* operations is given by

$$p = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}}{n}$$

and the average complexity of the *Retrieve* operations by

$$t = \frac{\sum_{i=1}^{n(n+1)/2}\sum_{j=1}^{m} r_{ij}}{n(n+1)/2}$$

It is known for any data structure solving the *range query problem of s*ize *n* that $\rho + t = \Omega(\log n)$

Range Query Problem-Solving Matrices

In the following we will define pairs of matrices of 0s and 1s whose product is the matrix *H* —that is, matrices that represent solutions to the *range query problem* – in an attempt to minimize the total number of 1s present in both matrices and, therefore, the average complexity of the operations. A recursive definition will be given later (see Definition 13). In particular, let the matrices $R = (r_i, j)$, $U = (u_i, j)$ be defined, whose product for any dimension *n* of the form 2^k with $k \in N^+$ is H^n.

Remember that the average complexity is calculated by dividing the total number of 1s by the number of different operations. Hence, if we are dealing with the problem of size *n* and let

$$\psi(n) = \sum_{i=1}^{n(n+1)/2}\sum_{j=1}^{m} r_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} u_{ij}$$

where *m* is the number of variables $z_1 \ldots z_m$ used to implement the solution (*m* may vary, although it necessarily has to be greater than or equal to *n*), then the average complexity is given by $\frac{\psi(n)}{n + \frac{n(n+1)}{2}}$ (*n* is the number of different *Update(j, x)* operations as a function of the first argument and $\frac{n(n+1)}{2}$ is the number of different possible arguments for a *Retrieve(i, j)* operation).

We will prove that our matrices hold for

$$\psi(n) = \frac{3}{2}n^2 - \frac{3}{2}n\log_2 n + \frac{9}{2}n - 2\log_2 n - 4$$

and this implies an average complexity that has a constant order of complexity.

Remember that a superindex is used to specify the size of the problem corresponding to the matrix *H*. Hence H^n denotes the matrix for the *range query problem of si*ze n, although the size of H^n is $\frac{n(n+1)}{2} \times n$.

Concepts and Notation

Let $I^m_{i,j}$ denote the matrix resulting from permuting the i^{th} and j^{th} rows in the identity matrix of dimensions m × m, denoted I^m. For any matrix *M* of dimensions *m* × *n*, $I^m_{i,j} \times M$ returns the matrix *M* in which rows *i, j* have switched position.

Generally, if $I^m_\sigma = I^m_{i_1,j_1} \times I^m_{i_2,j_2} \times \ldots \times I^m_{i_k,j_k}$, the effect of the multiplication $I^m_\sigma \times M$ is to switch the position of rows i_k and j_k of *M*, *then* do the same thing with rows i_{k-1} and j_{k-1}, *then* with rows i_{k-2} and j_{k-2}… until finally rows i_1, j_1 have been switched.

For any $n = 2^k$, any row rearrangement of matrix H^n associated with the range query problem of size *n* can be achieved by multiplying H^n by a certain identity transform I_σ, where $\sigma \in Permutations\left(\left(\frac{n+1}{2}\right)\right)$ and $Permutations(k) = \{ f : \{1 \ldots k\} \to \{1 \ldots k\} / \text{ f one to one-onto}, k \in N^+ \}$. This is due to a known algebraic result, stating that any permutation that is a member of *Permutations(k)* can be expressed as a composition of a certain number of permutations of that set whereby all the elements of {1…*k*} save two are held fixed.

The matrix H^n should ultimately be rearranged as the matrix S^n, which is defined below.

Definition 6 *For all n of the form 2^l with $l \in N^+$, let*

$$S_n = \begin{pmatrix} & \\ H^{\frac{n}{2}} & 0 \\ M_1^{\frac{n}{2}} & T^{\frac{n}{2}} \\ M_2^{\frac{n}{2}} & T^{\frac{n}{2}} \\ \vdots & \vdots \\ M_{\frac{n}{2}}^{\frac{n}{2}} & T^{\frac{n}{2}} \\ & \\ 0 & H^{\frac{n}{2}} \end{pmatrix}$$

where for any $m \in N^+$, k=1…m, the matrices M^m_k are square matrices of dimensions m defined as

$$(M^m_k)_{i,j} = \begin{cases} 1 & k \le j \\ 0 & otherwise \end{cases}$$

Hereinafter let $I_{H^n \to S^n}$ denote the matrix I_σ that leads to the transformation of H^n into S^n.

Hence, $I_{H^n \to S^n} \times H^n = S^n$.

Let us look at a couple of simple examples from which the specific expression of the matrix $I_{H^n \to S^n}$ can be easily deduced.

Example 7 *If* n=2, *then* $H^2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} = S^2$ *and hence* $I_{H^2 \to S^2} = I^2$

Let us also look at the example of the transformation of H^4 into S^4, illustrating the operations that need to be performed and the expression corresponding to the matrix $I_{H^4 \to S^4}$.

Here

$$H^4 = \left(\begin{array}{cccc} 1&0&0&0\\ 1&1&0&0\\ 1&1&1&0\\ \hline 1&1&1&1\\ 0&1&0&0\\ \hline 0&1&1&0\\ 0&1&1&1\\ \hline 0&0&1&0\\ \hline 0&0&1&1\\ \hline 0&0&0&1 \end{array}\right) \xrightarrow{I_{3,5}} \left(\begin{array}{cccc} 1&0&0&0\\ 1&1&0&0\\ 0&1&0&0\\ \hline 1&1&1&1\\ 1&1&1&0\\ \hline 0&1&1&0\\ 0&1&1&1\\ \hline 0&0&1&0\\ \hline 0&0&1&1\\ \hline 0&0&0&1 \end{array}\right) \xrightarrow{I_{4,5}} \left(\begin{array}{cccc} 1&0&0&0\\ 1&1&0&0\\ 0&1&0&0\\ \hline 1&1&1&0\\ 1&1&1&1\\ \hline 0&1&1&0\\ 0&1&1&1\\ \hline 0&0&1&0\\ \hline 0&0&1&1\\ \hline 0&0&0&1 \end{array}\right) = S^4$$

Hence, $I_{H^4 \to S^4} = I_\sigma = I_{4,5} \times I_{3,5}$

Where $\sigma : \{1..10\} \to \{1..10\}, \quad \sigma(3) = 4,\ \sigma(4) = 5,\ \sigma(5) = 3$

Remark 8

How does this type of transformations *affect the matrix approach to the range query problem? We know that it involves studying pairs of integer matrices R^n, U^n such that $R^n \times U^n = H^n$. But if $R^n \times U^n = H^n$, then $I_{H^n \to S^n}$ x $R^n \times U^n = S^n$. Hence the problem can be reformulated equivalently as entailing the study of matrix pairs whose product is the matrix S^n. In this case the algorithm that implements the Retrieve operations given in Definition 1 has to be modified, and the definition of the programs associated with a triplet <Z,R,U> is now as follows:*

Definition 9

Given a triplet <Z,R,U>, with Z = ($z_1 \ldots z_m$), R, U matrices of dimensions $n \times m$ and $m \times n$, respectively, let us define the following algorithms to implement the Update and Retrieve operations:

1. *Update(j,x): for l: 1 to m do* $z_l \leftarrow z_l + u_{l,j}x$
2. *Retrieve(l,j): output* $\sum_{l=1}^{m} r_{k,l} z_l$ *where k is given by:*

   a) $k = \sum_{s=0}^{i-2}\left(\frac{n}{2} - s\right) + (j - i + 1), \qquad 1 \le i \le j \le \frac{n}{2}$

   b) $k = \sum_{s=0}^{i-2}(n - s) + (j - i + 1), \qquad \frac{n}{2} < i \le j \le n$

   c) $k = \frac{n}{2}(i-1) + \left(j - \frac{n}{2}\right) + \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)}{2}, \qquad i \le \frac{n}{2}, \qquad j \ge \frac{n}{2}$

The intuitive idea is that the row number (k) of the matrix R^n associated with a Retrieve(i,j) operation is different now, as some rows have switched position.

The change of approach has no bearing on the complexity study of the operations, as, remember, the effect of multiplying any matrix by a certain I_σ does not alter the number of non-null matrix elements, but only switches the position of certain rows.

Recursive Definition of Our Problem-Solving Matrices

In this section we will give a recursive definition of our matrix pairs R^n, U^n as a function of the problem size n. The matrices hold for $R^n \times U^n = H^n$.

As mentioned already, the definition is valid for values of the form $n = 2^k$.

Let us refer to *blocks* of consecutive rows of the matrix R^n, which we consider to be divided into n horizontal blocks, the first formed by the first n rows, the second by the next (n-1) rows, the third by the (n-2) rows… up to the $(n-1)^{th}$ block, which is composed of two rows and the n^{th} block which consists of just the last row. Let R_i^n denote the i^{th} block of R^n and $R_i^n(j)$, the j^{th} row of this block such that the matrix R^n is given by

$$R^n = \begin{pmatrix} R_1^n \\ R_2^n \\ \vdots \\ R_n^n \end{pmatrix}$$

(note that if the dimensions of R^n are $\frac{n(n+1)}{2} \times m$, then the dimensions of each block R_i^n are $(n-i+1) \times m$).

R^n, U^n pairs are constructed by applying a function called *Refinement*. This function can be viewed as a two-stage process: the first stage involves executing a sequence of extension steps, and the second rearranging the rows of R^n by multiplying by a given identity transform matrix I_σ.

The following Definition 11 and Lemma 11 are needed to define the extension step concept.

Definition 10

Given a matrix M of 0s and 1s, two columns i, j are said to be disjoint if the set of rows $\{k/m_{k,i} = 1 = m_{k,j}\}$ is empty. Similarly, two rows i, j of M are said to be disjoint if the set of columns $\{k/m_{i,k} = 1 = m_{j,k}\}$ is empty.

Lemma 11

Let A, B be two matrices of 0s and 1s such that $A \times B = S^n$, of dimensions $\frac{n(n+1)}{2} \times m$ and $m \times n$, respectively. Assume that there are two columns i, j of A that are not disjoint and let $\{l_1 \ldots l_q\}$ be the set of rows of A for which $a_{l_k,i} = 1 = a_{l_k,j}$, $k=1 \ldots q$ holds. Then the rows i, j of B are disjoint.

Definition 12

Let A, B be two matrices of 0s and 1s such that $A \times B = S^n$, of dimensions $\frac{n(n+1)}{2} \times m$ and $m \times n$, respectively.

Assume that there is a set of columns $C = \{c_1 \ldots c_l\}$, $l \geq 2$, for which there is a non-empty maximal *set —including all the rows that meet the following condition— of rows $F = \{f_1, \ldots, f_q\}$ such that $a_{f_i,c_j} = 1 \quad \forall c_j \in C, \forall f_i \in F$*

We define the extension step associated with the sets C, F as the execution of the following actions on A and B:

1. *Insert a new column z_0 in A such that* $a_{i,z_0} = 1 \Leftrightarrow i \in F \quad \forall i = 1..\frac{n(n+1)}{2}$
2. *Add a new row z_0 in B such that* $b_{z_0,j} = \sum \{b_{k,j} / k \in C\} \qquad \forall j = 1..m$
3. *Columns $c_1 \ldots c_l$ of A are modified as follows* $a_{f_i,c_k} := 0 \quad \forall c_k \in C, \quad \forall f_i \in F$

It can be easily demonstrated that if an extension step is applied to a matrix pair whose product is the matrix S^n, the product of the resulting matrices is the very same matrix S^n.

We are now able to give a recursive definition of our matrix pairs.

Definition 13

Let us recursively define matrix pairs R^n, U^n of dimensions $\frac{n(n+1)}{2} \times m$ *and $m \times n$, respectively, with n* of the form 2^k with $k \in N^+$, *as follows*

1. *$n=2$:* $\langle R^2, U^2 \rangle = \langle H^2, I^2 \rangle$
2. *$n=2^k$:* $\langle R^2, U^2 \rangle = Refinement(\hat{R}^{2\hat{n}}, \hat{U}^{2\hat{n}})$ *where*

a) $$\hat{R}^{2n} = \begin{pmatrix} R^n & 0 \\ f_n^m(R_1^n(n)) & R_1^n \\ f_n^m(R_2^n(n-1)) & R_1^n \\ \vdots & \vdots \\ f_n^m(R_n^n(1)) & R_1^n \\ 0 & R^n \end{pmatrix}, \quad \hat{U}^{2n} = \begin{pmatrix} U^n & 0 \\ 0 & U^n \end{pmatrix}$$ *(m is the number of columns of R^n)*

b) $f_k^m : \{0,1\}^m \to [R^m \to R^k]$, *that is to say* $f_k^m(v_1..v_m)$, *returns a matrix —a linear mapping— of dimensions $k \times m$.* f_k^m *is defined such that the k rows are precisely the argument vector ($v_1 \ldots v_m$):*

$$f_k^m(v_1..v_m) = \begin{pmatrix} v_1 & \cdots & \cdots & v_m \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ v_1 & \cdots & \cdots & v_m \end{pmatrix}$$

c) Refinement($\hat{R}^{2n}, \hat{U}^{2n}$) is a two-phase process.

i. Extension steps: as many consecutive extension steps are executed on the matrix $\hat{R}^{2\hat{n}}$ *as necessary to assure that each row of the blocks* $f_n^m\left(R_i^n(n-i+1)\right), i = 1..n$, *and the blocks* R_1^n *have just one 1. The extension steps should be bound to sets of columns C that include either columns of the left-hand blocks* only*—blocks of the form* $f_n^m\left(R_i^n(n-i+1)\right)$*— or columns of the right-hand blocks —of the form* R_1^n.

Let $\hat{R}'^{2n}$, $\hat{U}'^{2n}$ *denote the matrices resulting from executing these extension steps.*

The matrix $\hat{U}'^{2n}$ *is* ***actually the final matrix*** U^{2n} *that we aim to define, as it is unaffected by the second phase of the Refinement process.*

Note: We have proved that the number of extension steps needed to construct R^{2n}, U^{2n} *is exactly* $2\left(\frac{n}{2}+\left(\frac{n}{2^2}-1\right)+\left(\frac{n}{2^3}-1\right)+\ldots+1\right)$.

ii. Rearrangement: the product of the matrices $\hat{R}^{2n}, \hat{U}^{2n}$ *is a matrix* S^{2n} *from which it likewise follows that the product of the matrices* $\hat{R}'^{2n}$, $\hat{U}'^{2n}$ *is* S^{2n}*, as it has already been demonstrated that the extension steps do not affect the product of the matrices. Assuming this result, this phase involves rearranging the matrix* $\hat{R}'^{2n}$ *by means of the multiplication* $I_{S^{2n}\to H^{2n}} \times \hat{R}'^{2n}$, *where* $I_{S^{2n}\to H^{2n}}$ *is the matrix that holds for* $I_{S^{2n}\to H^{2n}} \times S^{2n} = H^{2n}$. *Finally, the matrix* R_1^n *that we are trying to define is precisely* $R^{2n} = I_{S^{2n}\to H^{2n}} \times \hat{R}'^{2n}$.

Note: the existence of the matrix $I_{S^{2n}\to H^{2n}}$ *is straightforwardly deduced from the existence of the matrix* $I_{H^{2n}\to S^{2n}}$, *since if the expression corresponding to* $I_{H^{2n}\to S^{2n}}$ *is* $I^c_{i1,j1} \times I^c_{i2,j2} \times \ldots \times I^c_{ik,jk}$, *where c is the number of rows of* H^{2n}*, then* $I_{S^{2n}\to H^{2n}} = I^c_{ik,jk} \times \ldots \times I^c_{i2,j2} \times I^c_{i1,j1}$.

Remarks 14

a) *From the definitions of the matrices* T^n, S^{2n}, H^{2n} $\hat{R}^{2n}, \hat{U}^{2n}$ *it follows that* $\hat{R}^{2n}$ *x* $\hat{U}^{2n} = S^{2n}$

b) *As a consequence, and by definition of* R^n, U^n, *it holds that* $R^n \times U^n = H^n$.

c) *The maximum number of 1s present in each row of* R^n *is 2, whatever the value of n.*

d) *Let* $n = 2^{k+1}$ *for a certain natural number k. The number of extension steps that are executed in the Refinement phase of the matrices construction process is* $2\left(\frac{n}{2}+\left(\frac{n}{2^2}-1\right)+\left(\frac{n}{2^3}-1\right)+\ldots+1\right)$.

e) *Let A, B be two matrices of 0s and 1s, such that A* $\times$ *B =* S^n, *of dimensions* $\frac{n(n+1)}{2} \times m$ *and m* $\times$ *n, respectively.*

The execution of an extension step associated with the column and row sets C= $\{c_1 \ldots c_l\}$, *F=*$\{f_1 \ldots f_q\}$ *of A, respectively, leads to a change in the total number of 1s present in the two matrices according to the following expression:*

$\sum_{i\in C, j\in\{1..n\}} b_{i,j} + q - (l \times q)$. *If the value of* $\sum_{i\in C, j\in\{1\ldots n\}} b_{i,j} + q - (l \times q)$ *is greater than 0 then the total number of 1s in the matrices increases; if the value is equal to 0 then the number of 1s is unchanged, and if the value is less than 0 the total number of 1s decreases.*

As a consequence, each extension step executed in the Refinement phase of the process of constructing our matrices given in Definition 13 decreases the total sum of the number of 1s present in the two matrices.

Theorem 15

Let R^n, U^n be matrices of dimensions $\frac{n(n+1)}{2} \times m$ and $m \times n$, respectively, with n of the form 2^k with $k \in N^+$, *as defined in Definition 13.*

Let $\#R^n$, $\#U^n$ denote the number of 1s in the matrices R^n and U^n respectively.

It holds that

$$\#R^n + \#U^n = \frac{3}{2}n^2 - \frac{3}{2}n\log_2 n + \frac{9}{2}n - 2\log_2 n - 4$$

This represents a constant average complexity for the set of $\frac{n(n+1)}{2} + n$ Retrieve and Update operations.

As regards the number of variables $z_1 \ldots z_m$ required by the solution defined by our matrices as a function of the problem size *n*. Let *Var(n)* denote this number of variables, which, as we know, is the same as the number of columns and rows of R^n *and* U^n respectively. *It holds that*

$$m = n\log_2 n - 2n + 2\log_2 n + 2$$

Proof

For the proof of these results, see [5].

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 el *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

Adriana Toni – e-mail: atoni@fi.upm.es

J.B. Castellanos – e-mail: jcastellanos@fi.upm.es

Jose Joaquin Erviti – e-mail: jerviti@fi.upm.es

Facultad de Informatica, Universidad Politecnica de Madrid, Spain

DESCRIPTION REDUCTION FOR RESTRICTED SETS OF (0,1) MATRICES [1]

Hasmik Sahakyan

Abstract: *Any set system can be represented as an n-cube vertices set. Restricted sets of n-cube weighted subsets are considered. The problem considered is in simple description of all set of partitioning characteristic vectors. A smaller generating sets are known as "boundary" and "steepest" sets and finally we prove that the intersection of these two sets is also generating for the partitioning characteristic vectors.*

1. Introduction

In recent years, the processing of data flows has become a topic of active research in several fields of computer science. Continuous arrival of data items in rapid, potentially unbounded flows raises new challenges and research problems. The study of known combinatorial algorithms and their computational complexity for data flow conditions become an important issue.

Consider a $(0,1)$-matrix A of size $m \times n$. Let $R=(r_1,\cdots,r_m)$ and $S=(s_1,\cdots,s_n)$ denote the row and column sums of A respectively, and let $U(R,S)$ be the set of all $(0,1)$-matrices with row sums R and column sums S.

It was found by Gale and Ryser [R,1966] a necessary and sufficient condition for the existence of a $(0,1)$ matrix of the class $U(R,S)$. This result has found a recent revival in the field of discrete tomography [H, 1997]. In discrete tomography the problem is to reconstruct a discrete valued function f from knowledge of weighted sums of function values over subsets of the domain. A much studied special case is $m \times n$ $(0,1)$-matrices with known row and column sums, precisely matrices in the class $U(R,S)$.

As the number of matrices in this class may be high, it is of interest to study the reconstruction problem where with additional constraints on the $(0,1)$-matrices, which could either lead to a unique realization, or reduce the number of alternative solutions. The restrictions may be of different nature: requirements on rows of reconstructed matrices – to be different, some geometrical requirements such as convexity and connectivity, etc. It is proven ([B,1996], [W,2001], [D,1999] that the existence problems of horizontal and vertical convex matrices and the existence problem for connected) matrices (polyominoes are NP-complete; and the reconstruction problem for horizontal and vertical convex polyominoes can be solved in polynomial time. At the same time the complexity of the existence problem for matrices with different rows is still an open problem [BL,1988].

We assume now that we consider the last mentioned problem for data flow conditions and the coordinates of column sum vector S might varied slowly by the data flow. Then - which are the allowable values for coordinates of S to correspond to column sum vectors?

Complete description of the set of all integer-value vectors, which serve as column sum vectors for $(0,1)$-matrices with different rows, is given through its boundary elements [S,1997]. An alternative description of this set is known through its special elements - "steepest" vectors. The main result of this research states: the description might be given by the common (intersecting) elements of these sets - of upper boundary and steepest vectors, which minimizes the descriptor set size.

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

2. Problem Description

Let consider the problem of existence of a $(0,1)$-matrix by the given column sums vector S and with different rows. Let assume that the coordinates of vector S is varying slightly by data flow, and then the problem is in description of all integer vectors, which serve as column sums vectors for $(0,1)$-matrices of fixed size and with different rows. .

This problem has an equivalent formulation in terms of unit cube E^n.

Let $M \subseteq E^n$ be a vertex subset of fixed size $|M| = m$, $0 \le m \le 2^n$. An integer, nonnegative vector $S = (s_1, s_2, \cdots, s_n)$ is called the **characteristic vector of partitions** of set M, if its coordinates equal to the partition-subsets sizes of M by coordinates $x_1, x_2, \cdots, x_n$ - the Boolean variables composing E^n. s_i equals the size of one of the partition-subsets of M by the i-th direction and then $m - s_i$ is the size of the complementary part of partition. To make this notation precise we will later assume that s_i is the size of the partition subset with $x_i = 1$. Then the problem is in description of all integer-coordinate vectors, which serve as characteristic vectors of partitions for vertex subsets of size m.

3. Description through the Boundary Elements

Let Ξ^n_{m+1} denotes the set of all vertices of n dimensional, $m+1$ valued discrete cube, i.e. the set of all integer-vectors $S = (s_1, s_2, \cdots, s_n)$ with $0 \le s_i \le m$, $i = 1, \cdots, n$. The vertices are distributed schematically on the $m \cdot n + 1$ layers of Ξ^n_{m+1} according to their weights – sums of all coordinates. The L-th layer contains all vectors $S = (s_1, s_2, \cdots, s_n)$ with $L = \sum_{i=1}^{n} s_i$.

Let ψ_m denotes the set of all characteristic vectors of partitions of m-subsets of E^n. It is evident, that - $\psi_m \subseteq \Xi^n_{m+1}$. Let $\hat{\psi}_m$ and $\check{\psi}_m$ are subsets of ψ_m, consisting of all its upper and lower boundary vectors, correspondingly: $\hat{\psi}_m$ ($\check{\psi}_m$) is the set of all "upper" ("lower") vectors $S \in \psi_m$, for which for all $R \in \Xi^n_{m+1}$ greater than S (less than S), $R \notin \psi_m$.

These sets of all "upper" and "lower" boundary vectors have symmetric structures - for each upper vector there exists a corresponding (opposite) lower vector, and vice versa; so that also the numbers of these vectors are equal:

$$\hat{\psi}_m = \{\hat{S}_1, \cdots, \hat{S}_r\} \text{ and } \check{\psi}_m = \{\check{S}_1, \cdots, \check{S}_r\}.$$

Let $\hat{S}_j$ and $\check{S}_j$ be an arbitrary pair of opposite vectors from $\hat{\psi}_m$ and $\check{\psi}_m$ $\check{\psi}_m$ correspondingly. $I(\hat{S}_j)$ (equivalently $I(\check{S}_j)$) will denote the minimal sub-cube of Ξ^n_{m+1}, passing through this pair of vectors. Then,

$$I(\hat{S}_j) = \{Q \in \Xi^n_{m+1} \mid \hat{S}_j \le Q \le \check{S}_j\} \text{ (the coordinate-wise comparison is used).}$$

The following theorem states that the minimal sub-cubes passing the pairs of corresponding opposite vectors of the boundary subsets are continuously and exactly filling the vector area ψ_m.

Theorem 1 [S,1997]: $\psi_m = \bigcup_{j=1}^{r} I(\hat{S}_j)$.

It follows that the description of ψ_m is provided through the set of upper boundary vectors $\hat{\psi}_m = \{\hat{S}_1, \cdots, \hat{S}_r\}$ (correspondingly, the set of lower boundary vectors $\check{\psi}_m = \{\check{S}_1, \cdots, \check{S}_r\}$). Let assume that the upper boundary vectors are distributed between the layers L_{min} and L_{max} of Ξ^n_{m+1}. Then for each layer L, $L_{min} \le L \le L_{max}$, it is sufficient to have all upper boundary vectors situated on that layer..

4. Description through the "Steepest" elements

Let introduce a concept of "steepest" vectors, defined for each layer.

Definition 1 [B,1988]

Let $S = (s_1, s_2, \cdots, s_n)$ and $S' = (s'_1, s'_2, \cdots, s'_n)$ be two vectors of length n with integer, nonnegative components, and let $s_1 \ge s_2 \ge \cdots \ge s_n$ and $s'_1 \ge s'_2 \ge \cdots \ge s'_n$. S' is an **elementary flattening** of S if and only if S' can be obtained from S by:

(1) finding i, j such that $s_i \ge s_j + 2$; and then

(2) transferring 1 from the larger to the smaller; that is, $s'_i = s_i - 1$ and $s'_j = s_j + 1$; and then

(3) re-ordering the resulting sequence so that it is decreasing.

We mention that operation of elementary flattening doesn't move vector from one layer of Ξ^n_{m+1} to another.

Definition 2 [B,1988]

Let $S = (s_1, s_2, \cdots, s_n)$ and $S' = (s'_1, s'_2, \cdots, s'_n)$ be two vectors of length n with integer, nonnegative components, and let $s_1 \ge s_2 \ge \cdots \ge s_n$ and $s'_1 \ge s'_2 \ge \cdots \ge s'_n$. S' is **flatter** than S, and S is **steeper** than S', if and only if S' can be obtained from S by a non-empty sequence of elementary flattening.

S is a **steepest** vector if and only if there is no vector in ψ_m, which is steeper than S.

The following theorem is an extension of similar result [B, 1988], which is in terms of hypergraphs and degree sequences:

Theorem 2. If S belongs to ψ_m then all vectors flatter than S also belong to ψ_m.

Proof:

Let $S = (s_1, s_2, \cdots, s_n) \in \psi_m$ and $S' = (s'_1, s'_2, \cdots, s'_n)$ is flatter than S. It follows that there exists a sequence of elementary flattening, which transfers S to S'. We prove that after each elementary flattening, the obtained vector belongs to ψ_m. Let $s_i \ge s_j + 2$ and after an elementary flattening we receive the vector $(s_1, \cdots, s_i - 1, \cdots, s_j + 1, \cdots, s_n)$.

Consider the partitioning of E^n by i th and j th directions. Let $E^{n-2}_{x_i=1, x_j=1}$, $E^{n-2}_{x_i=1, x_j=0}$, $E^{n-2}_{x_i=0, x_j=1}$, $E^{n-2}_{x_i=0, x_j=0}$ be the corresponding sub-cubes, and $M_{x_i=1, x_j=1}$, $M_{x_i=1, x_j=0}$, $M_{x_i=0, x_j=1}$, $M_{x_i=0, x_j=0}$ - the corresponding subsets of M, belonging to these sub-cubes.

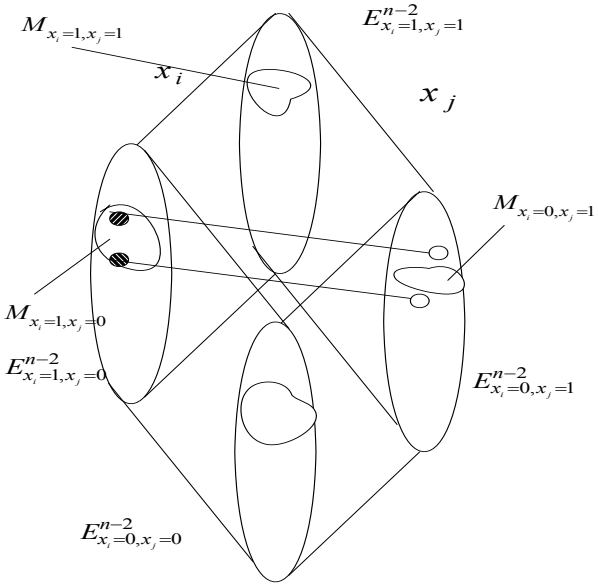
Then we have: $\left|M_{x_i=1,x_j=1}\right| + \left|M_{x_i=1,x_j=0}\right| = s_i$

$\left|M_{x_i=1,x_j=1}\right| + \left|M_{x_i=0,x_j=1}\right| = s_j$

Hence $\left|M_{x_i=1,x_j=0}\right| - \left|M_{x_i=0,x_j=1}\right| \geq 2$

Therefore there exist two vertices in $M_{x_i=1,x_j=0}$ such that the corresponding vertices in $E^{n-2}_{x_i=0,x_j=1}$ don't belong to $M_{x_i=0,x_j=1}$. Moving one of them from $M_{x_i=1,x_j=0}$ to $M_{x_i=0,x_j=1}$, will provide the necessary $s_i - 1$ and $s_j + 1$ values.

The geometrical visualisation is through the following picture:



It follows from the above theorem that the steepest vectors of each layer L, $L_{min} \leq L \leq L_{max}$ of Ξ^n_{m+1} provide the description of all vectors from ψ_m belonging to that layer.

5. Description through the Boundary Steepest Elements

On one hand, ψ_m can be described through the set of all upper boundary vectors, and on the other hand - through the set of all "steepest" vectors. Below we prove that ψ_m can be described having only the intersection of these two sets – which is the set of all "boundary steepest" vectors.

The theorem below states that if some layer of Ξ^n_{m+1} contains more than one upper boundary vector, then only the steepest ones of them are necessary for the description of ψ_m, or the same – if among the steepest vectors are both boundary and non-boundary, then only the boundary ones are necessary to describe the whole set of partitioning characteristic vectors.

Theorem 2. If a layer of ψ_m contains a boundary vector, then it can be obtained by operations of flattening from only an other boundary vector.

The theorem has been proved by contradiction, considering all possible cases.

Conclusion

Any set system can be represented as a subset of n-cube vertices set. For a given subset it is important to know the partition sizes, - the coordinates of partitioning characteristic vectors. Smaller generating sets are known as "boundary" and "steepest" sets and finally we prove that the intersection of these two sets is also generating for the partitioning characteristic vectors.

Bibliography

[S, 1997] H. Sahakyan. On a class of (0,1)-matrices connected to the subsets partitioning of E^n, Doklady NAS Armenia, v. 97, 2, 1997, pp. 12-16.

[B, 1988] Billington D., Conditions for degree sequences to be realisable by 3-uniform hypergraphs". The Journal of Combinatorial Mathematics and Combinatorial Computing". 3, 1988, pp. 71-91.

[D, 1999] Durr Ch., Chrobak M., Reconstructing hv-convex polyominoes from orthogonal orojections. Information Processing Letters 69 (1999) pp. 283-291.

[R, 1966] H. J. Ryser. Combinatorial Mathematics, 1966.

[H, 1997] G.T. Herman and A. Kuba, editors. Discrete Tomography: Foundations, Algorithms and Applications. Birkhauser, 1999.

[B, 1996] E. Barcucci, A. Del Lungo, M. Nivat, and R. Pinzani. Reconstructing convex polyominoes from horizontal and vertical projections. Theoret. Comput. Sci., 155:321{347, 1996.

[W, 2001] G.J. Woeginger. The reconstruction of polyominoes from their orthogonal projections. Inform. Process. Lett., 77:225{229, 2001.

Author's Information

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

A HW CIRCUIT FOR THE APPLICATION OF ACTIVE RULES IN A TRANSITION P-SYSTEM REGION

Victor J. Martinez, Luis Fernandez, Fernando Arroyo, Abraham Gutierrez

***Abstract**: P systems or Membrane Computing are a type of a distributed, massively parallel and non deterministic system based on biological membranes. They are inspired in the way cells process chemical compounds, energy and information. These systems perform a computation through transition between two consecutive configurations. As it is well known in membrane computing, a configuration consists in a m-tuple of multisets present at any moment in the existing m regions of the system at that moment time. Transitions between two configurations are performed by using evolution rules which are in each region of the system in a non-deterministic maximally parallel manner.*

This work is part of an exhaustive investigation line. The final objective is to implement a HW system that evolves as it makes a transition P-system. To achieve this objective, it has been carried out a division of this generic system in several stages, each of them with concrete matters. Within this division, in this paper the stage is developed by obtaining the part of the system that is in charge of the application of the active rules. These ones have been reached in a previous circuit [Fernández, 2005b] and they become the input to this circuit.

***Keywords**: Membrane Computing, Evolution Rules, Circuit design, Digital systems, Transition P System.*

Introduction

The Membrane Computing or P Systems (created by [Păun, 1998]) are computation systems based on the biomolecular processes of living cells. According to this, the investigations are based on the idea that the imitation of the procedures that take place in Nature and their application to machines, can lead to discover and to develop new computation models that will give place to a new generation of intelligent computers. These systems perform a computation through transition between two consecutive configurations. Transitions between two configurations are performed by using evolution rules present in each region. If the system reaches a configuration in which there are no applicable rules at any membrane, it is said that the system reaches a halting configuration and, hence, the computation is successful.

There are many papers about software tools implementing different P-system variants [Arroyo 2003], [Arroyo 2004b] and [Fernandez, 2005a]. However, they are very interesting in order to define hardware implementation of these kinds of systems. Moreover, evolution of transition P- systems is very complicate to translate into hardware devices due mainly to the membrane dissolving or membrane division capabilities of rules. Besides that, the non-deterministic maximally parallel manner in which rules are applied inside membranes is much appropriated to be implemented in digital hardware devices. In the case of P-systems hardware implementations only can find a few of papers: a Hardware Circuit for Selecting Active Rules [Fernandez 2005b], connectivity arrays for membrane processors [Arroyo 2004a], a multisets and evolution rules representation in membrane processors [Arroyo 2004b] or even a hardware membrane system description using VHDL [Petreska 2003] .

This work is a part of a very ambitious project: to find and carry out a Hardware System able to simulate P systems evolution. This generic system has been divided into several stages. The first stage one develops a circuit able to determine active rules in a determined configuration for the membrane [Fernández 2005b]. In this document, the second stage is developed: as a circuit for the application of the active rules obtained in the first stage.

In order to proceed in an appropriate way, it is needed to define a data structure containing information about the initial membrane state, that is, the initial multiset of objects, the set of evolution rules and the corresponding priority relation among them. The circuit takes these data inputs and produce, as output, a set of evolution rules, active rules, which are able to produce the needed changes into the system in order to make evolve it to into the next configuration. Obviously, there are some needed constrains for the hardware design, and they are the following:

a. The cardinality $O = \{a,b,c,d,e,f,g,h,i,j\}$ of the alphabet is limited to 10.
b. The multiset of objects associated to the membrane i, ω_i . It is represented in a hardware register of 4-bits words of length 10.
c. The finite set of evolution rules *Ri* associated to the membrane *I* (10 per membrane).
d. The priority relation ρ_i among the rules *Ri*.
e. Extra information to determine applicability of rules: existence register, for determining whether or not there exist children membranes for the next evolution step.

Therefore, the previous Circuit for Selecting Active Rules determine which of the evolution rules present inside a membrane are active (in binary positive logic) accordingly to the multiset of objects present in the membrane.

The objective of the work that now shows up will consist in obtaining a HW circuit that indicates the rules to be applied to a Transition P System. The inputs are the following registers: multiset of objects associated to the membrane *i*; Evolution Rules antecedents and Initial Active Rules. The output will be a register that contains the number of times that each rule (See Fig.1) should be applied. These values associated to each rule will serve to carry out, in a later process, the communication of elements among regions.

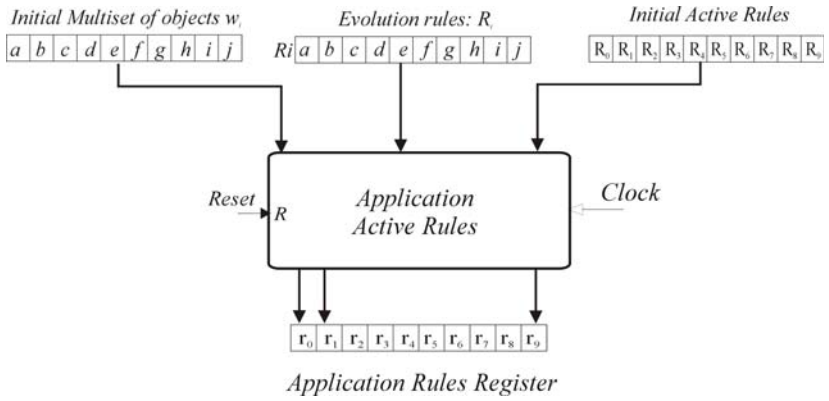


Fig. 1: General structure of the circuit for Application of Active Rules.

Application of the Active Rules

The application of the active evolution rules is a repetitive process that can be carried out in different ways. The paper [Fernandez, 2006] allows us to obtain circuits or systems optimized as for level of complexity and resources utilization. A first option to apply the rules could consist on the step-by-step application of the group of active rules. In this case, we choose in an aleatory way a single time each rule until draining all the possible applications.

However, it can be proven that one obtains the same result if we apply an evolution rule a certain number of times. The maximum number of times a rule can be applied into a multiset ω_i of state of the cell in an evolution step will be called "value MAX ." This value is obtained considering that you apply only this rule, that is to say, without keeping in mind the other rules.

The following example illustrates the concept of value *MAX*: If we have a multiset of objects $\omega_i=a^{10}b^5c^7$ and the following active rules $R_1=a^1$, $R_2=b^2c$ and $R_3=bc^2$;el value MAX of R_1 is 10, the value MAX of R_2 is 2 and the value MAX of R_3 is 3.

The fact of applying a rule a value bigger than one implies that it consumes, in an evolution step, a bigger number of elements from the multiset ω . Hence, the whole process requires a minor number of steps and, therefore, the end of the process will be reached in a shorter length of time. The following pseudo-code sequence illustrates the necessary process to obtain the number of times that each active rule should be applied in a region:

(1) $R \leftarrow InitialActiveRules$
(2) **BEGIN**
(3) **DO**
(4) $r_i \leftarrow Aleatory(R)$
(5) $MAX \leftarrow Applicability(r_i, \omega)$
(6) $N \leftarrow Aleatory(1, MAX)$
(7) $\omega \leftarrow (\omega - (N * Antecedent(r_i)))$
(8) $counts(N, r_i)$
(9) $R \leftarrow ActiveRules$
(10) **WHILE** $|R| \geq 1$
(11) **END**

Explanation of the algorithm:

- (1) The process uses the group of active available rules initially R .
- (4) At each iteration, one of the rules r_i of this group will be applied. Such rule will be randomly obtained.
- (5) On the selected rule, the value of applicability *MAX* is obtained and it is applied with an aleatory multiplicity *N* between 1 and the value *MAX* (6)
- (7) The application of selected rule r_i will consist in subtracting from the starting multiset ω , the values of the antecedent elements multiplied by the value *N* of rule application. In turn, we will increase *N* times the particular accountant that counts the number of times that rule has been applied (8).
- (9) On the new obtained multiset ω it has been proved again the applicability of the available rules.
- (10) Every time the group of applicable rules is upgraded, the end of the process has been controlled. The stop condition is obtained when the number of applicable rules is zero. While R is bigger or the same as the unit, it executes, once more, a new iteration of the process.

Basic Functional Units (F.U.)

This section defines the previous step to design the complete circuit of active rules application to the evolution of a transition P system. It will consist on obtaining certain basic operative functional units. These functional units will solve each one of the simple tasks needed to obtain the complete application. The design of the final circuit will be based on the assembling of these modules together with the corresponding combinational and sequential logic which allows their integrated operation.

Applicability MAX F.U.:

In this case, we will obtain the design of the circuit that obtains the *MAX* applicability of a rule. This value supposes the largest number of times that a rule can be applied independently of the other ones. To do so, we will only keep in mind the multiplicities of their elements and the multiplicities of the elements of the multiset ω .

Therefore, the inputs into this circuit will be two registers: one with the content of the values of the multiset ω of state of the membrane and another with the antecedent of the rule we want to get their value *MAX*. The output will be the value *MAX* of this rule.

To obtain this value we should carry out the division among each couple of elements similar of the multiset ω_i with r_i . From each division, we will participated the maximum that will represent the largest number of times that mentioned element could be consumed in an evolution step without bearing in mind the other elements. We will take the smallest value out of all the partial results of these divisions.

1 Aleatory Active Rule F.U.

This circuit will randomly select a rule from the ActiveRules register to be applied. The ActiveRules register contains binary values indicating, with positive logic, the active rules from existing rules of the system. The output of this circuit will be a register in which only one rule will be randomly selected. One input Enable *"E"* will activate the starting of the clock that attacks the random generator. This generator will select decimal numbers aleatory in a serial way until getting some one that corresponds with the number of the active rule. When this value is obtained, the aleatory number generation should stop.

The main part of the circuit will be a 1 bit Decimal Multiplexer. The selection inputs of the Multiplexer correspond with the outputs from the Decimal aleatory generator. This generator will stop when some of the outputs of the Multiplexer are set to 1. The position that indicates this output corresponds with the randomly selected active rule.

When only one position of the ActiveRules register is present, it means that only one active rule exists. In this case it is not necessary to carry out the process of aleatory selection. To avoid a loss of efficiency that could happen in this case, the circuit has been endowed with a special operation condition. The solution consists in detecting this condition previously and to provide the Multiplexer with a specific input *"ALL"* that allows to get the content of all the multiplex inputs (See Fig.2).

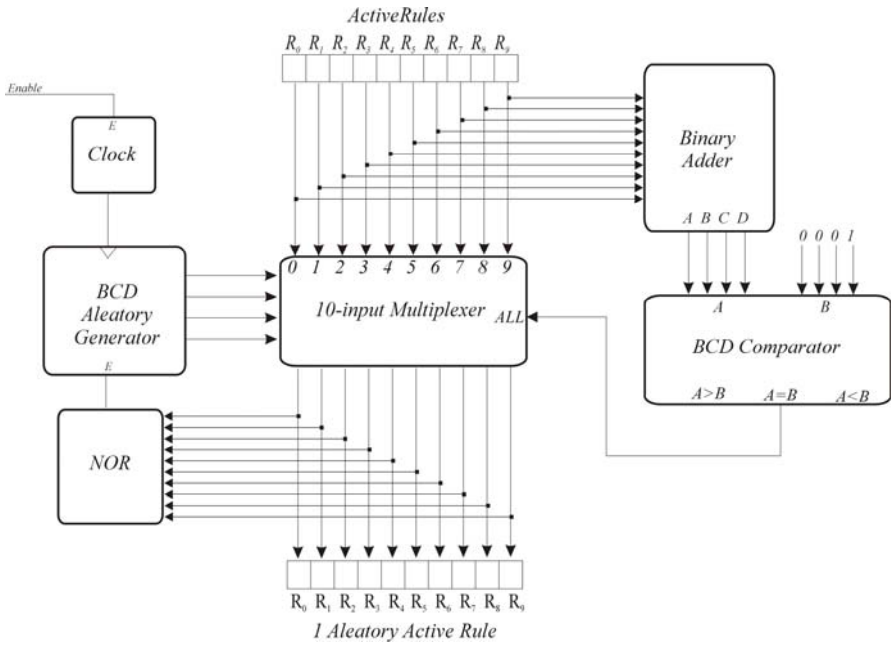


Fig. 2: Internal structure of the circuit for 1 Aleatory Active Rule.

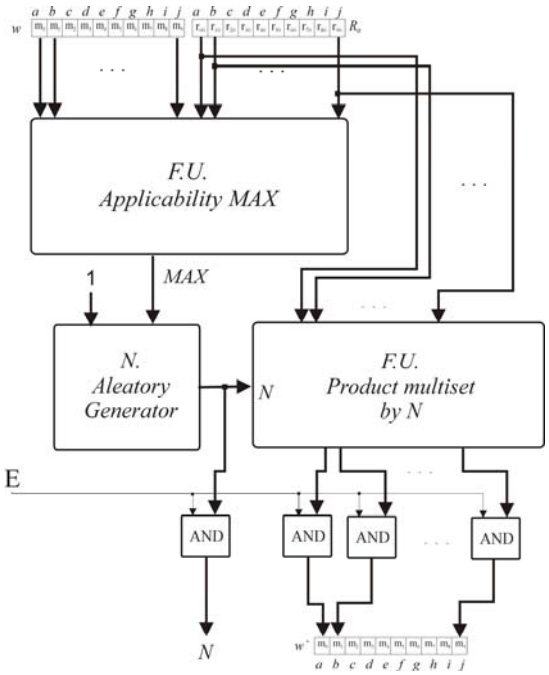


Fig. 3: Internal structure of the circuit for random Aplication r_i 1-*MAX*.

Application r_i Aleatory 1-*MAX* F.U.

This circuit obtains an output register associated to the application of the rule r_i. This register contains the multiplicity of those elements that they should be subtracted from the multiset associated to the membrane region. This result is reached in the case the rule r_i is applied between a value 1 and its *MAX* value, elected in an aleatory way. To count the number of times each rule is being applied, we should also extract this aleatory value of application *N*.

The inputs are, therefore, the register that contains the multiset of objects associated to the region ω and the antecedent r_i. Also, we will endow the circuit with an input Enable *"E"* that allows selecting witch rules will act in each evolution step. The output will be stored in another register ω. Later on, these values will be subtracted from the state multiset in order to obtain the resulting new values.

Internally (See Fig.3), the circuit is formed by an "F.U. Applicability MAX of a rule" (to obtain the value MAX) and an "F.U. Product multiset by N". Also, a generating circuit of aleatory numbers will select one random value between 1 and *MAX*. This number will be the value for which we will multiply by the antecedent of the rule. Finally, a series of AND gates allows to enable or to disable the output in function of the sign Enable (*E*) :

General Structure of the Circuit

The general circuit is the result of the assembling of the different Functional Units, together with the sequential logic of control. The sequential logic determines the evolution of the internal steps the circuit should travel though until reaching the stop condition. This condition will be given when the register *R* of active rules is empty.

The sequence of events that the *sequential controller* should activate is based on the use of a 2 bits counter that allows reaching 4 states. Each state determines an evolution event. The counting continues in a recurrent way until the stop condition is reached, and then the accountant will be stopped (See Fig. 4).

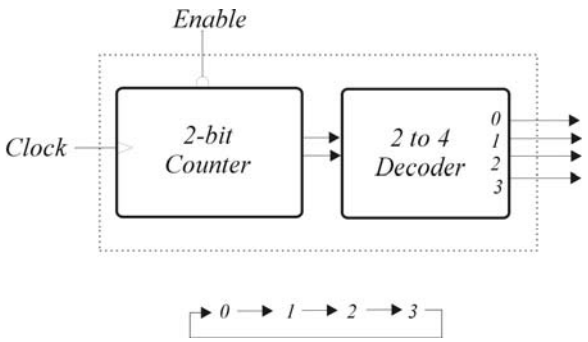


Fig.4. Internal structure of the Sequential controller and of their states sequence.

The general structure of the circuit is based on the following sequence of states:

a. **Initialization (Reset) phase**: The sign *R = 1* load the register ω with the values of the initial multiset of objects. On the other hand, the Initial Active Rules Register has been obtained according to the paper [Fernández, 2005b].
b. **State 0**: It proceeds to load the register ω and to calculate active rules. The Active Rules Register is obtained by checking in each evolution step the condition of applicability for the rules active initials and for the new multisets of objects. Applicable rules are those rules accomplishing that their antecedent is included in the multiset of objects found inside the membrane.
c. **State 1**: The Active Rules Register is loaded and begins the process to obtain 1 Aleatory Active Rule.

d. **State 2**: The 1 Aleatory Active Rule register is loaded. The application of the selected rule begins in order to obtain the new multiset of objects ω and the calculation of the number of times that such rule will be applied.
e. **State 3**: The Multiset of Objects Register ω and the Application Rules Register are loaded. The Application Rules Register will store the number of times each rule has been applied to. This register will be the output result we want to obtain.
f. **Turn to the state 0**: to load the register ω and starting a new calculation cycle with the new values.

The Fig. 5 shows the circuit to determine the times the active rules should be applied.

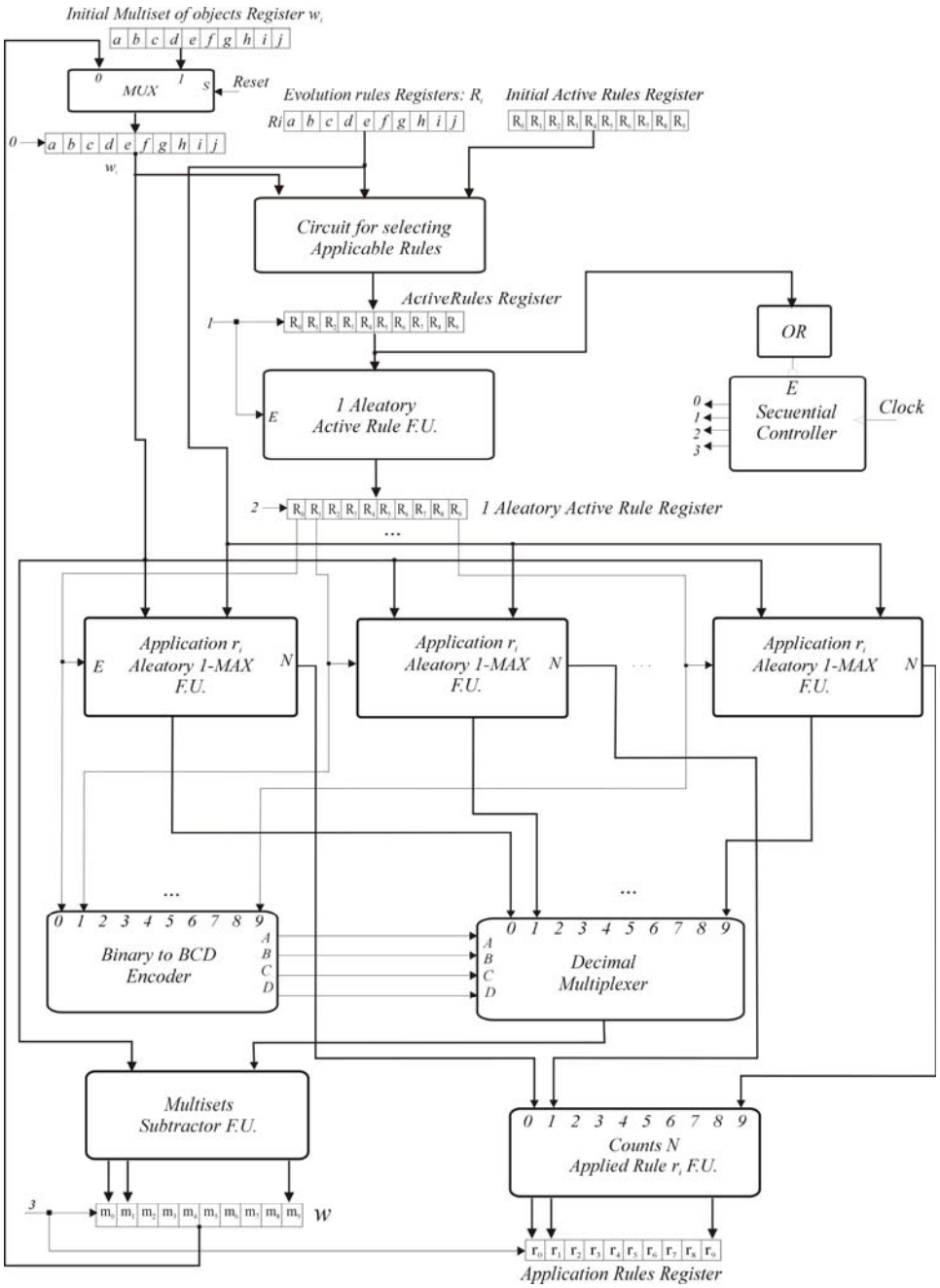


Fig. 5: Internal structure of the circuit to determine the times that the active rules should be applied.

Conclusions

This paper presents a direct way to design a circuit able to obtain the number of active rules application inside the membrane in a non deterministic and massively parallel application. The final objective is implementing a hardware circuit accomplishing the outlined initial requirement. That is, given an initial multiset of objects ω, a finite set of evolution rules and an initial Active Rules, the circuit provides the number set of application of rules to a membrane. The hardware implementation is founded on basic components like registers, logical gates, multiplexer and sequential elements. The development of the digital system can be carried out using hardware-software architectures like Handel C or VHDL. The physical implementation can be accomplished on hardware programmable devices like FPGA's. The next step in the process for the developing of a membrane processor is to design a circuit able to communicate elements from one region to another.

Bibliography

[Arroyo 2003] F. Arroyo, C. Luengo, A.V. Baranda, L.F. de Mingo, A software simulation of transition P systems in Haskell, Pre-Proceedings of Second Workshop on Membrane Computing, Curtea de Arges, Romania, August 2002 and Proc. of WMC02, Curtea de Arges, Romania, (Gheorghe Paun, Grzegorz Rozenberg, Arto Saloma, Claudio Zandron Eds.) Lecture Notes in Computer Science 2597, Springer-Verlag, 2003, 19-32.

[Arroyo 2004a] F. Arroyo, C. Luengo, J. Castellanos, L.F. de Mingo: A binary data structure for membrane processors: Connectivity arrays, Artiom Alhazov, Carlos Marín-Vide and Gheorghe Păun (eds.): Preproceedings of the Workshop on Membrane Computing, Tarragona, July 17-22 2003, 41-52 and in (Carlos Marín-Vide, Giancarlo Mauri, Gheorghe Păun, Grzegrorz Rozenberg, Arto Saloma Eds.) Lecture Notes in Computer Science, 2933, Springer Verlag, 2004, 19-30.

[Arroyo 2004b] F.Arroyo, C.Luengo, L.Fernandez, L.F.Mingo, J.Castellanos. Simulating membrane systems in digital computers. International Journal Information Theories and Applications, 11, 1 (2004), 29-34.

[Arroyo 2004c] F. Arroyo, C. Luengo, J. Castellanos, L.F. de Mingo, Representing Multisets and Evolution Rules in Membrane Processors, Pre-proceedings of the Fifth Workshop on Membrane Computing (WMC5), Milano, Italy, June 2004, 126-137.

[Fernández, 2005a] L.Fernández, F.Arroyo, J.Castellanos, V.J.Martinez, Software Tools / P Systems Simulators Interoperability, Pre-proceedings of the 6th Workshop on Membrane Computing, Vienna - Austria, July 2005.

[Fernández, 2005b] L.Fernández, V.J.Martínez, F.Arroyo, L.F.Mingo, A Hardware Circuit for Selecting Active Rules in Transition P Systems, Workshop on Theory and Applications of P Systems. Timisoara (Rumanía), september, 2005.

[Fernández, 2006] L.Fernández, F.Arroyo, J.Castellanos, J.A.Tejedor, I.García, New Algorithms for Application of Evolution Rules based on Applicability Benchmarks, BIOCOMP06 International Conference on Bioinformatics and Computational Biology, Las Vegas (USA), july, 2006 (submitted).

[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.

[Petreska 2003] B.Petreska, C.Teuscher, A hardware membrane system. Preproceedings of the Workshop on Membrane Computing (A.Alhazov, C.Martín-Vide and Gh.Paun, eds) Tarragona, July 17-22 2003, 343-355.

Authors' Information

Victor J. Martinez Hernando – Dpto. Arquitectura y Tecnologia de Computadores de la Escuela Universitaria de Informatica de la Universidad Politecnica de Madrid, Ctra. Valencia, km. 7, 28031 Madrid (Spain); e-mail: victormh@eui.upm.es

Luis Fernandez Munoz – Dpto. Lenguajes, Proyectos y Sistemas Informaticos de la Escuela Universitaria de Informatica de la Universidad Politecnica de Madrid; Ctra. Valencia, km. 7, 28031 Madrid (Spain); e-mail: setillo@eui.upm.es

Fernando Arroyo Montoro – Dpto. Lenguajes, Proyectos y Sistemas Informaticos de la Escuela Universitaria de Informatica de la Universidad Politecnica de Madrid, Ctra. Valencia, km. 7, 28031 Madrid (Spain); e-mail: farroyo@eui.upm.es

Abraham Gutierrez – Dpto. Informatica Aplicada de la Escuela Universitaria de Informatica de la Universidad Politecnica de Madrid, Ctra. Valencia, km. 7, 28031 Madrid (Spain); e-mail: abraham@eui.upm.es

ПРОЦЕДУРА НЕЧЕТКОЙ ФОРМАЛИЗАЦИИ ПОКАЗАТЕЛЕЙ В ОЦЕНКЕ УСТОЙЧИВОГО ФУНКЦИОНИРОВАНИЯ БАНКОВ

Александр Я. Кузёмин, Вячеслав В. Ляшенко

Аннотация. *Рассмотрены вопросы формального описания таких показателей устойчивого функционирования банка как ликвидность и прибыльность. Предложена нечеткая интерпретация оценки эффективности управления банком. Проанализирована возможность формализации оценки функционирования банка на основе иерархии уровней соответствующего нечеткого множества.*

Ключевые слова: *нечеткие числа, функция принадлежности, ликвидность, прибыльность, банк, управление.*

Введение.

Устойчивое и планомерное развитие банковского сектора играет очень важную роль в воспроизводственной структуре экономики, так как посредством банковской деятельности организуется движение и перераспределение денежных и капитальных ресурсов. Вместе с тем, реально возникающие трудности на различных этапах развития или преобразования социально-экономических систем, требуют должного обеспечения базисного условия устойчивой деятельности банков, которое, в общем случае, выражается через установление приемлемого соотношения ликвидности и прибыльности. Иными словами ликвидность и прибыльность необходимо рассматривать в качестве основных компонентов единой системы оценки финансовой устойчивости и надежности всей системы управления банком. При этом одной из ключевых задач, которая требует решения, является адекватное описание взаимосвязи ликвидности и прибыльности, что, в конечном счете, и определяет актуальность данного направления исследования.

Обоснование цели исследования.

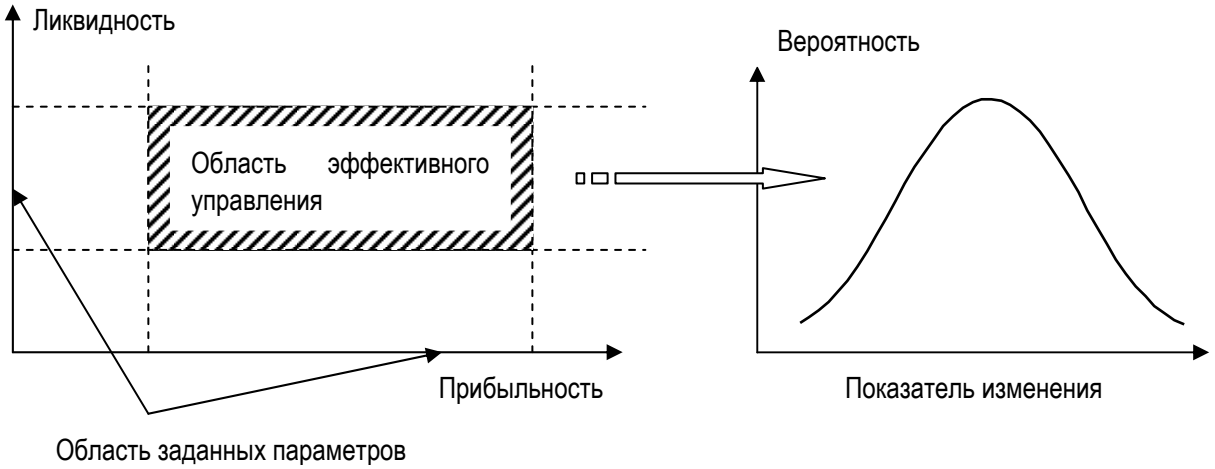
Основу решения выбранного направления исследований, как правило, базируются на статистических выводах либо моделях, в основу которых положены подходы теории игр. Однако и в том и в другом случае математическую базу исследований составляют вероятностные методы анализа данных.

Примером такого рассмотрения деятельности банка следует назвать работы E. Berglof, G. Roland, G.J. Mailath, L.J. Mester, T. Hellmann, K. C. Murdock, J.E. Stiglitz [1, 2, 3]. Тем не менее, основная проблема, которая возникает при построении адекватной системы управления экономическим процессом или объектом связана с тем, что законы экономического развития предполагают наличие такого взаимодействия между разными субъектами рынка и учета влияния на это взаимодействие разных проявлений среды окружения, которые не владеют определенной статистической природой в классическом понимании. Поэтому, построение системы управления некоторым экономическим процессом или объектом требует определенной формализации, учитывающей не только имеющуюся статистическую неопределенность, а и субъективную вероятность, которая объективно присутствует при принятии экономических решений. Некоторым образом решение данного вопроса достигается посредством ввода в рассмотрение различных аспектов информационной насыщенности, рассматриваемых показателей деятельности банка. Именно это направление выбрано в качестве основного в работе М. Малютиной и С. Париловой [4]. Тем не менее, данное направление не в полной мере способствует решению поставленной задачи, так как возникает другая проблема, которая связана с необходимостью рассмотрения ранжирования различных проявлений информационной насыщенности

тех или иных показателей банковской деятельности. Таким образом, естественным является использование формальных подходов теории нечетких множеств, которые и позволяют описать возникающие субъективные вероятности в исследовании экономических процессов в целом, и банковской деятельности в частности. В тоже время, несмотря на достаточно значительное количество работ в обозначенной предметной области исследования [5, 6, 7, 8], открытым остается вопрос, касающийся выбора процедур построения и обоснования вида функций принадлежности нечетких переменных, которые в дальнейшем используются в соответствующих моделях. Основная причина открытости данного вопроса, прежде всего, связана с многовекторностью направлений использования методов теории нечетких множеств, большинство из которых находятся на начальных этапах своего развития. Исходя из этого, в качестве основной цели данного исследования рассматривается формализация этапов процедуры нечеткого описания взаимосвязи экономических параметров, которая характеризует параметры ликвидности и прибыльности банковской деятельности. Обоснованность выбора данной цели исследования также связано с тем, что в общем случае схемы или диаграммы, объединяющие данные о финансово-экономических показателях деятельности какого-либо субъекта хозяйствования, содержат противоречивую информацию. Поэтому дополнительная обработка таких данных, в первую очередь, должна быть направлена на преобразование финансовых данных в информацию, которая будет полезна в процессе принятия решений, выявлении и интерпретации скрытых тенденций.

Вероятностная интерпретация оценки эффективного управления банком с точки зрения взаимосвязи его ликвидности и прибыльности.

В общепринятом понимании взаимосвязь между ликвидностью и прибыльностью банка может быть выражена в виде обратно пропорциональной зависимости. Этот факт имеет очень простое экономическое объяснение. Так с ростом степени ликвидности банковских активов снижается вероятность получения более высоких доходов от таких активов, и, наоборот, менее ликвидные активы банка способны априори приносить более высокие доходы. Классическим примером такого проявления взаимосвязи ликвидности и прибыльности в банковской деятельности есть то, что более рискованные кредитные операции могут принести и более высокие доходы.



а) б)

Рис. 1. Сущность оценки эффективности управления банком

В тоже время параметры ликвидности в целом являются регулируемыми из вне (с точки зрения конкретного банка) и устанавливаются в целом по банковской системе. При этом прибыльность банка во многом связана с проведением активно-пассивных операций банка и, в общем, может быть охарактеризована спрэдом между его кредитными и депозитными ставками. Данные ставки в свою очередь, с одной стороны, подчинены классическому закону соответствия спроса и предложения, а с другой подвержены конкурентному влиянию с боку других банков.

Таким образом, суть оценки эффективности управления банком с точки зрения его параметров ликвидности и прибыльности может быть интерпретирована как вероятность нахождения в некоторой заданной области, которая и определяется соответствующими показателями рассматриваемых параметров (рис. 1а).

При этом, задавая различные значения изменения предполагаемых параметров допустимых значений ликвидности и прибыльности можно получить кривую (рис. 1б), которая в целом будет характеризовать эффективность управления банком с точки зрения возможного изменения анализируемых параметров.

Нечеткая формализация эффективности управления банком с точки зрения взаимосвязи его ликвидности и прибыльности.

Вместе с тем такая вероятностная интерпретация эффективного управления банком может быть рассмотрена и в терминах теории нечетких множеств, что, как было указано выше, является более целесообразным. Данный переход возможен путем введения в рассмотрение функции принадлежности некоторого набора показателей ликвидности и прибыльности банка соответствующему подмножеству эффективных управляющих воздействий данных показателей. которые находятся в центре области эффективного управления (рис. 2).

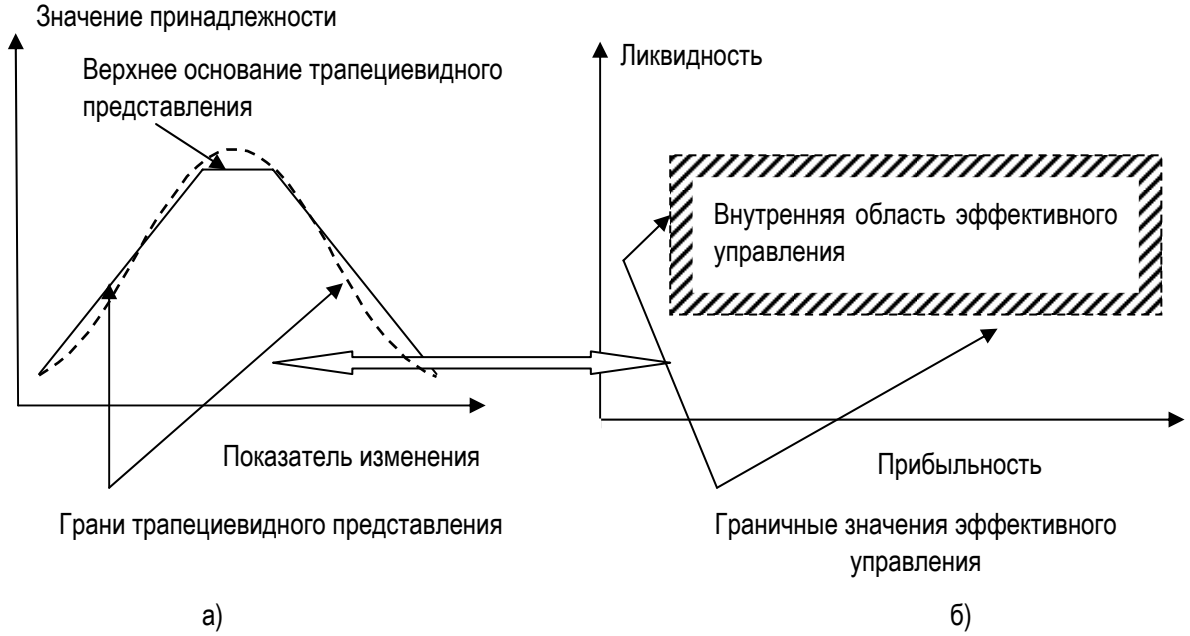


Рис. 2. Взаимосвязь трапециевидного представления функции принадлежности показателей банковской деятельности и области эффективного управления банком

Тогда нечеткая интерпретация эффективности управления банком в заданном фазовом пространстве сводится к построению и оценке соответствующих функций принадлежности, которые характеризуют степень достижения эффективного управления банком на заданных интервалах изменения анализируемых показателей банковской деятельности. При этом в качестве формального описания таких функций имеет смысл выбрать нечеткую интерпретацию изменений предполагаемых параметров в пределах допустимых значений показателей ликвидности и прибыльности, которые в вероятностной модели представлены соответствующей вероятностной кривой (см. рис. 1б). Целесообразность такого перехода обусловлена тем, что нечеткая формализация соответствующей вероятностной кривой возможна на основе понятия нечеткого числа L-R типа [9], которое в данном случае можно рассматривать как трапециевидное нечеткое число (рис. 2а).

Такая интерпретация функции принадлежности позволяет не только формально описать исследуемые процессы, но и учесть существующие экономические аспекты в их развитии. Так в данном случае грани трапециевидного представления функции принадлежности оценки эффективности банковской деятельности характеризуют управление ликвидностью и прибыльностью банка с точки зрения их граничных значений. В тоже время верхнее основание трапециевидного представления функции принадлежности можно рассматривать с точки зрения таких значений ликвидности и прибыльности,

Уровни нечеткого представления эффективного управления банком.

Данный подход был апробирован на реальных данных анализа взаимосвязи показателей ликвидности и прибыльности банковской деятельности для банковской системы Украины в целом. В результате такого анализа были построены различные функции принадлежности, которые характеризуют эффективность банковской деятельности с учетом изменения допустимых значений ликвидности для различных интервалов спрэда между кредитными и депозитными ставками (рис. 3, в данном случае анализировались значения текущей ликвидности).

Как видно из рис. 3, предложенное представление формализации показателей устойчивого функционирования банков, позволяет не формально проанализировать различные комбинации рассматриваемых параметров и обосновать наиболее приемлемые. В данном случае наиболее приемлемым, с точки зрения эффективного функционирования банковской системы Украины в целом, можно считать варьирование спрэда между кредитными и депозитными ставками в переделах 4–12% и достаточного уровня текущей ликвидности в пределах 55–65%.

В тоже время, рассмотренный пример наталкивает на мысль о необходимости рассмотрения различных функциональных представлений, характеризующих степень достижения эффективного управления банком для определенных значений одного из исследуемых параметров в зависимости от интервала изменения другого. Решение такой задачи возможно на основе введения в рассмотрение уровней нечеткого множества области эффективного управления. В данном примере в качестве таких уровней можно рассматривать функциональные зависимости функций принадлежности значений ликвидности для определенных интервалов изменения спрэда. Тогда оценка эффективности управления банком может быть определена на основании рассмотрения различных условий обобщения соответствующих уровневых подмножеств. При этом сущностная сторона формализации такого процесса определяется конкретными условиями функционирования банка на различных временных этапах его деятельности, что может быть представлено в виде отдельных операций над нечеткими множествами и нечеткими числами.

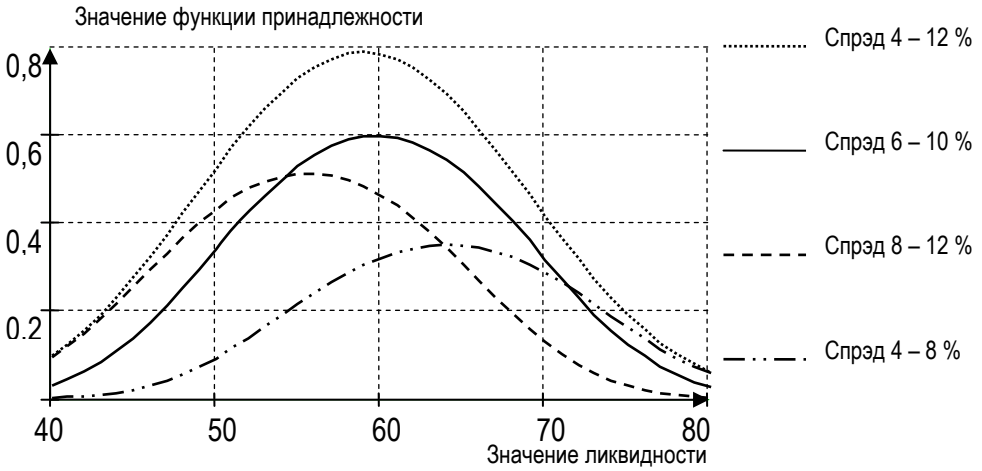


Рис. 3. Функции принадлежности, характеризующие степень достижения эффективного управления банковской системой на заданных интервалах изменения анализируемых показателей банковской деятельности

Заключение

Таким образом, в работе рассмотрена процедура перехода от вероятностной интерпретации оценки эффективного управления банком к ее нечеткой модели. Вместе с тем соблюдена сущностная характеристика рассматриваемого вопроса. Это позволяет проводить анализ эффективности управления банком с учетом возможного изменения различных его параметров, которые, и определяют, устойчивость функционирования банка.

Литература

Berglof , E., Roland, G. «Soft Budget Constraints and Banking in Transition Economies» J. of Comp. Econ., 26, 1998.

Mailath, G.J., Mester, L.J., «A positive Analysis of Bank Closure», J. of Financ. Intermed., 3, 1994.

Hellmann T., Kevin C. Murdock, Joseph E. Stiglitz 'Liberalization, «Moral Hazard in banking, and Prudential regulation: Are Capital Requirements Enough?» 1998, forthcoming in AER.

Малютина М., Парилова С. Поведение банков в условиях переходной экономики: причины чрезмерных рисков. – М.: EERC, 2001. – 52 с.

Rocha-Mier Luis, Villareal Francisco and Sheremetov Leonid. Agent-based Collective Intelligence for Inter-bank Payment Systems // FSSCEF 2004. – Vol. 2. – P. 321–322.

Canfora Gerardo, D'Alessandro Vincenzo and Troiano Luigi. Opportunities Within the New Basel Capital Accord for Assessing Banking Risk by Means of Soft Computing // FSSCEF 2004. – Vol. 2. – P. 457–465.

Дончеко В.С. Нечеткие множества: Аксиома абстракции, статистическая интерпретация, наблюдения нечетких множеств // KDS 2005. – Vol. 1. – P. 218–222.

Недосекин А.О. Применение нечетких моделей в управлении финансами банков // http://sedok.narod.ru.

Ахрамейко А.А., Железко Б.А., Ксеневич Д.В., Морозевич А.Н. Методика многоуровневой агрегированной оценки и прогнозирования финансового состояния предприятий // http://sedok.narod.ru/scgroup.html.

Информация об авторах

Александр Я. Кузёмин – профессор; e-mail: kuzy@kture.kharkov.ua

Вячеслав В. Ляшенко – ст.н.с.

Харьковский Национальный Университет по Радио Электроники; Харьков, Украина

ВЕРОЯТНОСТНЫЙ ПОДХОД СРАВНИТЕЛЬНОЙ ОЦЕНКИ ФУНКЦИОНИРОВАНИЯ БАНКОВСКОЙ СИСТЕМЫ

Александр Я. Кузёмин, Вячеслав В. Ляшенко

Аннотация. *Рассматривается целесообразность анализа ликвидности и прибыльности банков в качестве ключевого фактора в системе их экономической безопасности. Обосновывается вероятностная модель интерпретации устойчивости функционирования банковской системы. Приводится вероятностная сравнительная оценка функционирования различных банковских систем.*

Ключевые слова: *ликвидность, прибыльность, банковская система, вероятность.*

Введение.

Анализ финансовых потоков как банковской системы в целом, так и отдельных банков в частности является одной из ключевых составляющих построения адекватной системы экономической безопасности любого субъекта хозяйствования действующего в рыночной экономике. Это связано с тем, что именно благодаря банкам и их деятельности осуществляется движение и перераспределение денежных и капитальных ресурсов. Поэтому рассмотрение различных вопросов функционирования и развития банковской системы постоянно находится в центре внимания, что и делает данное направление весьма актуальным. При этом особое внимание заслуживает сравнительная оценка различных банковских систем. В целом это способствует не только выявлению приемлемых подходов к решению различных проблемных вопросов, а и возможности упреждающей оценки в принятии решений, касающихся соответствующего развития банковской системы.

Наиболее распространенными подходами, сравнительного анализа различных экономических систем, как правило, является

- либо относительное обобщение динамики соответствующих макро показателей, которое основывается на описательных статистических данных [1, 2],
- либо построение кластерных моделей, которые позволяют ранжировать степень развития сравниваемых систем. Примером такого подхода можно назвать исследования А.М. Карминского, А.А. Пересецкого, С.В. Голованя, А.В.Копылова [3, 4], В. Снитюк [5].

Тем не менее, использование обозначенных выше подходов, прежде всего, так или иначе, предполагает выбор определенных показателей, которые используются в дальнейшем сравнительном анализе. При этом необходимо определять значимые факторы, провести их согласующую ранжировку и лишь затем осуществлять сравнительный анализ, что само по себе является довольно-таки сложной задачей. Поэтому, на наш взгляд, в качестве предварительного анализа целесообразно использовать несколько иной подход, который базируется на вероятностной сравнительной оценке определенного показателя.

Такая интерпретация позволяет не только избежать процедуры согласованной ранжировки данных, но и произвести соответствующий сравнительный анализ, который значительно может дополнить классические подходы.

Обоснование параметров и интерпретация сравнительной оценки.

Прежде всего, следует отметить, что существенную роль при анализе устойчивости банковской системы занимают вопросы управления банковской ликвидностью, так как именно уровень ликвидности с точки зрения ее достаточности отождествляется с возможностью осуществления как обычных, так и непредвиденных обязательств банка перед своими клиентами. В то же время уровень банковской

ликвидности значительной мерой взаимосвязан с определенным уровнем прибыльности как отдельного банка, так и банковской системы в целом. Таким образом, в качестве одних из параметров сравнительной оценки развития банковской системы можно выбрать ее уровень ликвидности и прибыльности.

Если рассматривать вероятностную интерпретацию управления банковской деятельностью с учетом определенного уровня ликвидности, следует учитывать, что банк стремится поддерживать объем ликвидных средств на уровне, достаточном для обеспечения выполнения взятых обязательств. Вместе с тем банк определяет вероятность того, что ему будут необходимы заемные ресурсы для выполнения своих обязательств [6]. При этом также следует учитывать, – уровень ликвидности довольно таки сильно взаимосвязан с уровнем прибыльности банка, что в общепринятом понимании может быть выражено в виде обратно пропорциональной зависимости.

Тогда интерпретация развития банковской системы на основе анализа ликвидности может быть рассмотрена как вероятность попадания случайной двумерной величины в некоторую заданную область, где в качестве границ такой области выступают приемлемые и допустимые параметра уровней ликвидности и прибыльности. Именно данная модель и применена для последующего анализа различных банковских систем.

Сравнительная оценка функционирования банковских систем Украины и России.

В качестве примера использования вероятностной оценки сравнения банковских систем на основе учета взаимосвязи уровней ликвидности и прибыльности рассмотрены соответствующие показатели аналогичных периодов банковских систем Украины и России.

При этом анализировались соответствующие вероятностные характеристики исследуемых параметров в допущении гипотезы о нормальном распределении представленных данных. Так в первом случае рассматривалась вероятность оптимально возможного уровня прибыльности банковских систем в зависимости от возможного интервала варьирования уровня текущей ликвидности (рис. 1), где в качестве прибыльности выступает величина спрэда между кредитными и депозитными ставками.

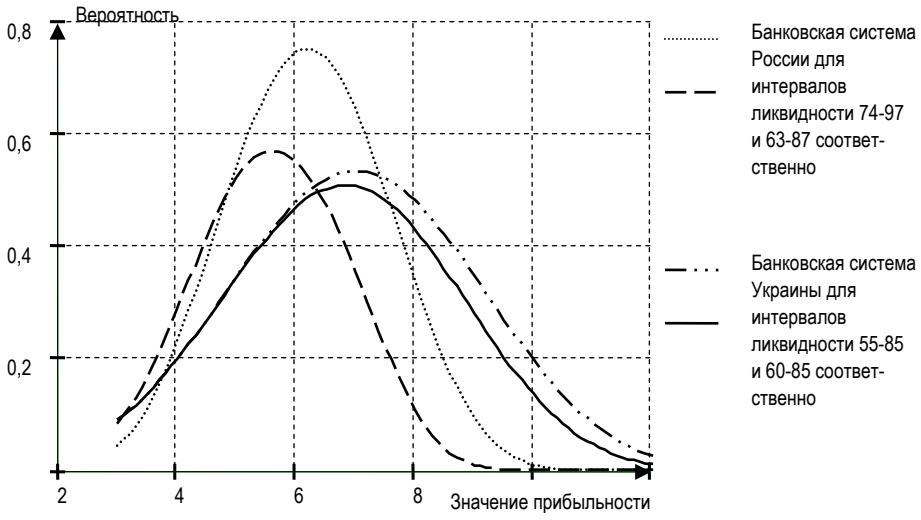


Рис. 1. Вероятностна оценка прибыльности анализируемых банковских систем для различных возможно допустимых интервалов текущей ликвидности

Иными словами, в данном случае важным является установление приемлемого соотношения между уровнем различных процентных ставок, что также может быть индикатором устойчивого развития банковской системы.

Как видно из рис. 1 с точки зрения возможно допустимого интервала текущей ликвидности наиболее устойчивое состояние развития демонстрирует банковская система России. Это следует как из больших значений соответствующей вероятности, так и меньшего значения спрэда между уровнем различных процентных ставок. При этом в данном случае принимается во внимание тот факт, что прибыльность по банковской системе в целом определяется не столько максимальной величиной спрэда процентных ставок, сколько возможностью получения более доступных ресурсов и соответственно объемом оборота средств, проходящих через банковскую систему.

Во втором случае на рис. 2 представлена вероятность того, насколько целесообразным является увеличение спрэда между депозитными и кредитными ставками с учетом возможно допустимых изменений интервалов текущей ликвидности.

Как видно из рис. 2 соответствующие граничные значения спрэдов для различных банковских систем коррелируют с данными рис. 1. В то же время вероятностная оценка целесообразности увеличения спрэда для банковской системы Украины в отдельном случае является больше чем для банковской системы России. Этот факт можно интерпретировать как большую склонность банковской системы Украины к увеличению спрэда между кредитными и депозитными ставками. Иными словами, в данном аспекте, можно говорить о менее устойчивом развитии банковской системы, который связан с риском либо формирования, либо размещения соответствующей ресурсной базы банковской деятельности.

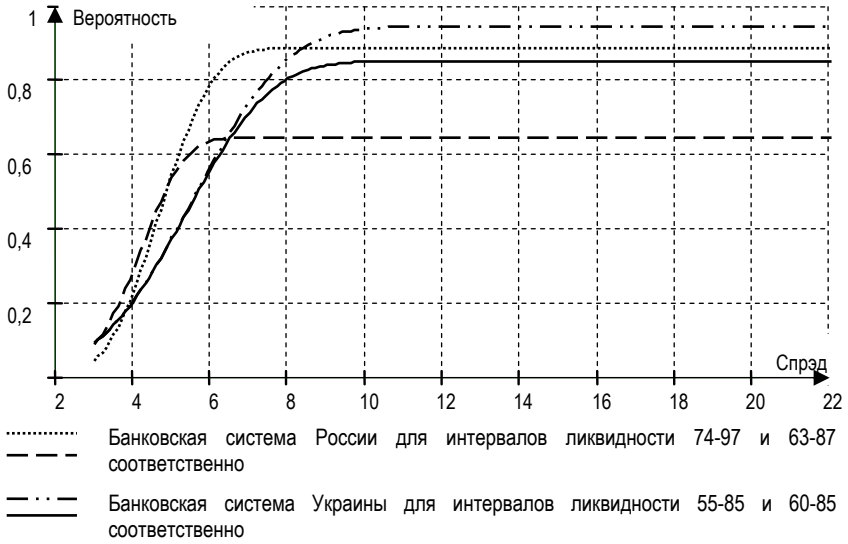


Рис. 2. Вероятностная оценка целесообразности увеличения спрэда между депозитными и кредитными ставками с учетом возможно допустимых интервалов текущей ликвидности

Заключение.

Таким образом, рассмотренный вероятностный подход сравнительной оценки функционирования банковской системы позволяет не только проанализировать относительное функционирование различных банковских систем, но и выявить определенные особенности их развития, что является довольно таки существенным с точки зрения построения адекватной системы экономической безопасности.

Литература.

1. Медведева Е.В. Сравнительный анализ банковских систем России и Чешской Республики в условиях переходного периода в экономике // Материалы V Международной научно-практической конференции «Страны с переходной экономикой в условиях глобализации». – М.: РУДН, 2006.
2. Дробышевский С., Козловская А., Левченко Д., Пономаренко С., Трунин П., Четвериков С. Сравнительный анализ денежно-кредитной политики в переходных экономиках // Научные труды ИЭПП. – М.: ИЭПП, 2003. – № 58Р.
3. Карминский А.М., Пересецкий А.А., Головань С.В. Модели рейтингов российских банков. Построение, анализ динамики и сравнение // Препринт #WP 2004 XXP. – М.: РЭШ, 2004. – 56 с.
4. Головань С.В., Карминский А.М., А.В. Копылов, Пересецкий А.А. Модели вероятности дефолта российских банков. Предварительное разбиение банков на кластеры // Препринт # 2003 XXX. – М.: РЭШ, 2003. – 49 с.
5. Снитюк В. Эволюционная кластеризация сложных объектов и процессов // XI-th International Conference «Knowledge-Dialogue-Solution» – Varna, 2005. – Vol. 1. – P. 232–237.
6. Shaffer S. A. Test of Competition in Canadian Banking // Journal of Money, Credit and Banking. – 1993. – Vol. 25, № 1. – P. 37–56.

Информация об авторах

Александр Я. Кузёмин – профессор; Харьковский Национальный Университет по Радио Электроники; Харьков, Украина; e-mail: kuzy@kture.kharkov.ua

Вячеслав В. Ляшенко – ст.н.с.; Харьковский Национальный Университет по Радио Электроники; Харьков, Украина

АЛГОРИТМ ДЛЯ ВЫЧИСЛЕНИЯ ДИФРАКЦИИ ФРЕНЕЛЯ, ОСНОВАННЫЙ НА ДРОБНОЕ ФУРЬЕ ПРЕОБРАЗОВАНИЕ

Георги Стоилов

***Аннотация**: Для решения дифракционного интеграла в оптике использовано дробное Фурье преобразование (ДрФП). Предложено использование подхода со сканированием для нахождения порядка ДрФП. Таким способом ускоряется вычислительный процесс дифракции. Показан базовый алгоритм и промежуточные результаты вычисления на каждом этапе.*

***Ключевые слова**: дифракция Френеля, дробное Фурье преобразование*

Введение

Анализ множества оптических систем и устройств ведет к вычислению дифракции в разных условиях. Применение современных способов оптической и цифровой обработки изображений предоставляет возможность решение этой задачи. Точное вычисление дифракционной картине получено в результате освещении комплексно пропускающих объектов или отражающих поверхностей представляет собой задачу, требующую большие вычислительные ресурсы. Поэтому явна необходимость введения быстрых вычислительных алгоритмов и уменьшение вычислений упрощением решения волнового уравнения [1]

$$\nabla^2 \upsilon - \frac{1}{c^2}\frac{\partial^2 \upsilon}{\partial t^2} = -s\,, \qquad (1)$$

где *c* - скорость света, υ - скалярная величина, описывающая волну в произвольной точке пространства, *s(x,y,z,t)* – известная функция излучающей поверхности.

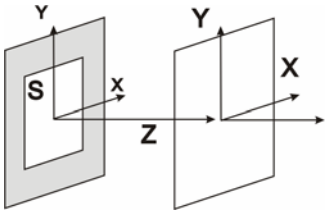


Рис.1. Излучающая и регистрирующая поверхности

В ряде случаев излучающая и регистрирующая поверхностей (Рис.1) можно представить как параллельные плоскости. В основе большей части используемых приближений стоит решение уравнения (1) при помощи интеграла Кирхгофа [2]:

$$A(X) \approx \int_s \frac{1}{r} P(x) e^{ik(x-X)} \cos\left(\frac{\sqrt{x-X}}{r}\right) dx \qquad (2)$$

$$k = \frac{2\pi}{\lambda}, r = \sqrt{(x-X)^2 + \Delta Z^2} \ ,$$

где *P(x)* - функция излучения поверхности *S*; *r* - радиус-вектор; λ - длина волны; причем решение упрощается использованием разных предположений. Для простоты рассмотрим решение в одномерном случае. Самое неточное приближение дифракционного интеграла, применяемое в оптике - это эго замена с Фурье преобразованием (ФП) [2].

$$A(X) \approx \frac{1}{r} \int_s P(x) e^{ik(x-X)} dx \qquad (3)$$

Необходимое условие для применения этого подхода – чтобы апертура излучающего объекта и размеры дифракционной картины были бы намного меньше, чем расстояние между ними. В оптических систем обычно рассматривается дифракция света в вакууме или ее прохождение через сферических оптических элементов (линзы), введение которых вносит квадратический фазовый множитель в подинтегральной величине интеграла Кирхгофа. Для решения этой разновидности дифракционного интеграла успешно применяется ДрФП, введенное Виктором Немиасом в 1980 году [3]. Оно имеет разные определения, которые доказано эквивалентные и используются в зависимости от области применения. Одно из них это дефиниция:

$$f_a(X) \equiv \int_{-\infty}^{+\infty} \sqrt{1 - i\cot(\alpha)}\, e^{i\pi\left(\cot\alpha . X^2 - 2\csc\alpha . x . X + \cot\alpha . x^2\right)} f(x) dx \ , \qquad (4)$$

где a - порядок ДрФП и $\alpha = \frac{a\pi}{2}$.

Строгое (точное) доказательство и условия эго применения в различных оптических дифракционных задачах разработано Озактасом, Залевским и Кутаем [4]. Дифракционный интеграл для оптической системы линз и переход волной через вакуум описываются следующим образом:

$$\hat{h}_{lens}(x,X) = \delta(X-x)e^{\frac{-i\pi X^2}{\lambda f}}$$
$$\hat{h}_{space}(x,X) = e^{\frac{i2\pi d}{\lambda}} e^{\frac{-i\pi}{4}} \sqrt{\frac{1}{\lambda d}} e^{\frac{i\pi (X-x)^2}{\lambda d}} \quad , \qquad (5)$$
$$\hat{g}(X) = \int_s P(x)\hat{h}\ (x,X)dx$$

где $\hat{g}(X)$ - комплексная амплитуда волнового поля в плоскости дифракции на расстояние *d*, $\hat{h}_{lens}(x,X)$ - ядро, использовано в случае применения тонких линз с фокусном расстояние *f* и $\hat{h}_{space}(x,X)$ - распространение света через вакуум.

Подобно ФП, ДрФП может быть представлено в виде суммы вместо в виде интеграла. Основной цели этого перехода к использованием дискретных функциях вместо непрерывных - это эго приложение в компьютерной обработки цифровых изображений. ДрФП можно представить посредством несколько последовательных операций, одна из которых - ФП [4,6]. Естественное развитие подхода - искать быстрые алгоритмы вычисления аналогичные быстрому Фурье преобразованию (БФП). Для вычисления ДрФП использован алгоритм и программа быстрого преобразования цитированных выше авторов.

Задача

В ряде случаев, условие применения ДрФП не может быть исполнено из-за большой апертурой объекта по сравнению с расстоянием, на котором регистрируется дифрагированная волна. Аналитическое решение интеграла в этом случае не существует. Известно, что, в дальней области, функция дифрагированной волны можно описать посредством ФП. С ДрФП представляется поведение функции в промежуточных состояний при переходом функции к ее Фурье образ. На основе этого можно искать решение интеграла Киргова посредством ДрФП и найти этой величиной ряда ДрФП, при которой получается самая лучшая аппроксимация.

Алгоритм вычисления

Выбирается такая строка в изображении, которая проходит через некоторой сложной части, т. е. строка состоит из элементов разной и возможно большей амплитудой. Таким образом, приближение будет иметь более выделенный максимум при перемене параметра аппроксимации. Для него вычисляется интеграл Кирхгофа, учитывая уравнение (2) и БДрФП (6). Ищется лучшее совпадение двух решений посредством изменения порядка ДрФП. С найденном таким образом параметром выбранной строки в изображении вычисляется ДрФП для целого изображения.

Контроль приближения может осуществиться посредством выбора несколько строк и колонок, для которых вычисляется параметр и находится их среднее значение.

Другой вариант алгоритма представляет вычисление интеграла Кирхгофа и после чего к полученными данными применяется обратное преобразование посредством ДрФП. Тогда восстановленный и оригинальный образ можно сравнить легче, потому что обычно в оригинальном образе не имеет комплексной составляющей, а в восстановленном образе она присутствует только в результате неточного приближения и вычисления.

Результаты

Для проверки алгоритма выбран простой объект – кольцо (рис. 2) с постоянной величиной интенсивности освещенных зон и ноль для фона. Для простоты рассмотрена зависимость только горизонтальной составной дифракцией. Таким образом, изменения изображения в концу освещенных зон более ясно видны. Выбрана квадратная апертура с размером 102.4 мкм, шаг дискретизации 100 нм. Вычисление интеграла Кирхгофа для зоны Френеля сделано при расстоянии 10 нм. Выбрана длина волны 533 нм (Рис.3.).

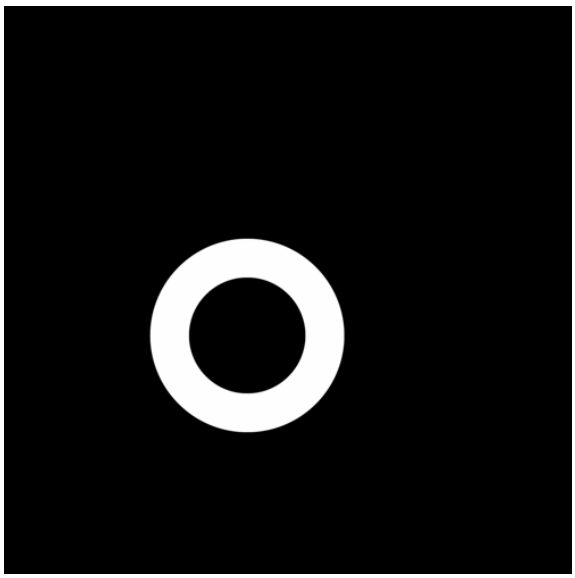


Рис. 2. Оригинал тестового изображения

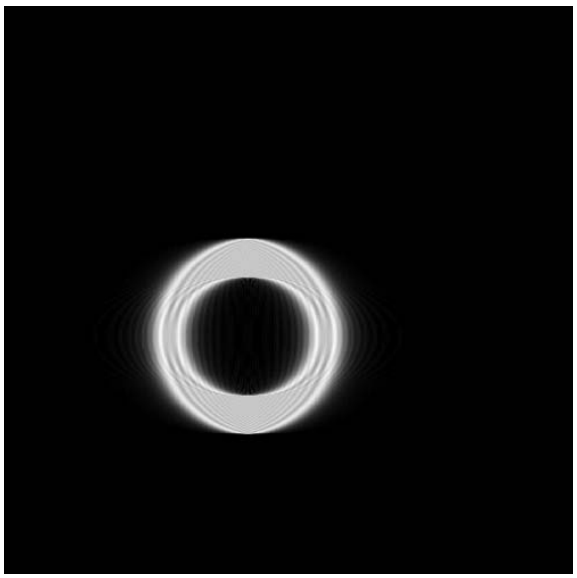


Рис. 3. Дифракционная картина в зоне Френеля

Обратное вычисление реализуется посредством поиска решения через ДрФП. Результаты, полученые при разных значениях порядка ДрФП показанные на Рис.4.

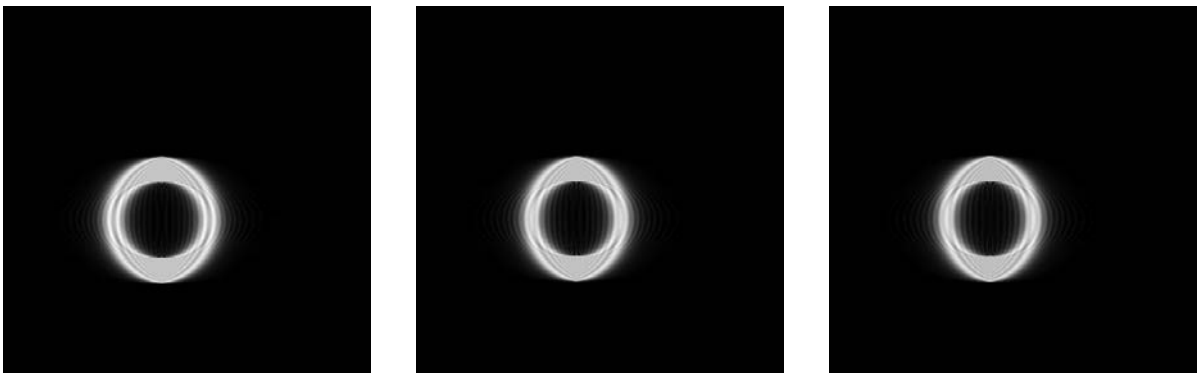
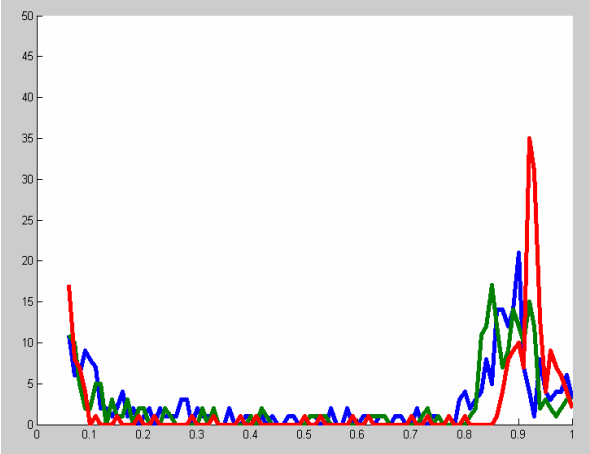


Рис. 4 Обратное преобразование строк -0.2, -0.3 и -0.34 ДрФП

Когда функция известна, критерий оптимизации можно искать как самой маленькой средной квадратической ошибки разницы оригинального и восстановленного образов. Если оригинал неизвестен и изменяется только эго амплитудная характеристика, а фаза постоянная, можно приложить подход минимизации имагинерной составной восстановленного образа. Это случай перехода паралелного лазерного пучка через амплитудной маской.



Фиг. 5 Гистограмма значений изображения при различных порядках ДрФП

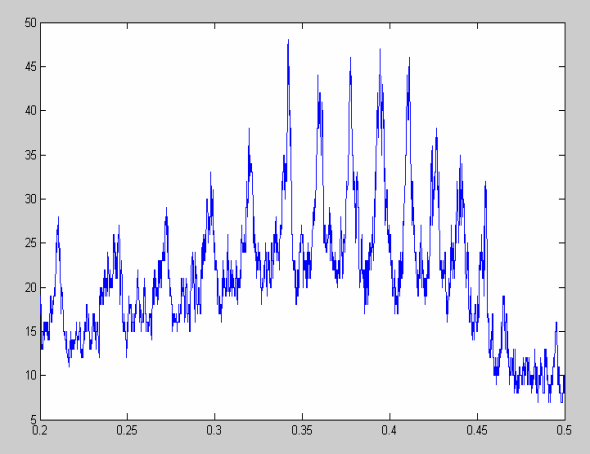


Рис. 6. Максимум в гистограмме в зависимости от порядка ДрФП

В выбранный нами объект, амплитуда имеет только две значения: 1 и 0. Таким образом, чистота белой части изображения может использоваться как критерий успешного восстановления. На Рис.5 представлена гистограмма значений для некоторых порядков ДрФП. При реальных измерений, вычисление ДрФП и нормализация данных теряют истинное значение амплитуды. Если имеются решения близкие до целевого, то значения будут группированы в две области: около 0 и около амплитуды. Когда нормировка выполняется после ДрФП, значение амплитуды немного ниже 1. Значения близких нулю не показаны, так как ищется максимум близко к 1.

Поиск решения задачи делается через последовательным изменением порядка ДрФП в интервале –1 до 0. Решение есть порядок, при которым получается наиболее высокий максимум в гистограммах, показанных на Рис.5. На Рис.6 показано изменение значения максимума в зависимости от порядка ДрФП. Из-за периодичности функции поиска глобального максимума, можно применить только сканирование в данным интервалом. Вычисление порядка ДрФП с точностью 0.01 позволяет применение быстро сходящих алгоритмов, например метод с разделением интервала пополам.

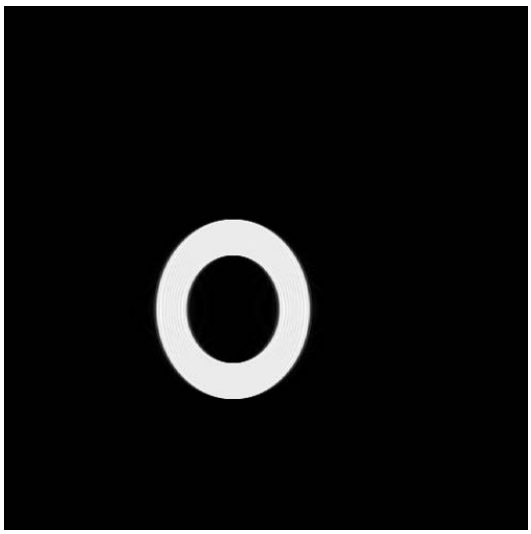


Рис. 7. Восстановленное изображение

Видно, что максимум является при значении порядка ДрФП приблизительно равно 0.34 (точное значение 0.3419).

Восстановленное изображение при этом значении показано на Рис.7. Масштабирование изображения не учитывается в процессе вычисления. Когда параметр принимает различные значения, изменяется и размер изображения. Проявление этого эффекта показано как деформирование изображения в горизонтальным направлением, так как вычисления проведены в этом направлении.

Программное обеспечение

Разработана компьютерная программа, состоящая из двух частей. Использован язык Microsoft Visual C для MS WINDOWS. Первая часть программы вычисляет интеграл Кирхгофа, при котором интегрирование проведено методом прямоугольников. Из-за колебательной природе кривых, ошибка интегрирования немного выше по сравнению с приложением формулы трапеции или других приближений более высокого порядка. Вычислительный процесс осуществлен на базе двуядерного процессора Атлон 64 4400+ и операционная система Windows XP. Обработка массива из 256x256 пикселей длится 16 минут, обработка массива 512x512 – 8 часов, а 1024x1024 – 5 суток.

Второй модуль программы ответствен для БДрФП и оптимизации эго порядка. Вычисление БДрФП для изображения, состоящее из 1024x1024 пикселей, длится 1 минуту. В процессе оптимизации применяется БДрФП только для одной строки, содержащий 1024 пикселей. При этом, сканирование для цели оптимизирования порядка с -1 до 0 с шагом 0.01 длится 3 секунды.

Выводы

Предложен алгоритм для вычисления дифракции света в зоне Френеля посредством обнаружения самого подходящего значения порядка ДрФП в одном сечении и эго применение для вычисления всего изображения. Показаны результаты обработки тестового изображения для каждого этапа алгоритма. Обработка сделана только по направление одной координатой с целью получения лучшей визуализацией.

Литература

1. Papoulis A., Systems and Transforms with Applications in Optics, McGRAW-HILL Book Company, 1968.
2. John M.Cowley, Diffraction physics, Amsterdam: North Holland, 1975.
3. V. Namias, The fractional order Fourier transform and its applications to quantum mechanics, J. Inst. Math Appl., 25, 241–265 (1980).
4. H.M. Ozaktas, M.A. Kutay, and G. Bozdagi. , Digital computation of the fractional Fourier transform. IEEE Trans. Sig. Proc., 44:2141{2150, 1996.
5. Ozaktas H., Zalevsky Z., Kutay M., The Fractional Fourier Transform with Application in Optics and Signal Processing, John Wiley & Sons, Ltd, 2001.
6. Bultheel A., K.U.Leuven H. Martnez-Sulbarany , Recent developments in the theory of the fractional Fourier and linear canonical transforms, Bulletin of the Belgian mathematical Society- Simon Stevin ,2006

Информация о авторе

Георги Стоилов – ЦЛОЗОИ БАН, н.с. I ст., София 1113, ул. „Акад. Г.Бончев” 101, П.К. 95, e-mail: gstoilov@optics.bas.bg

Software Engineering

ACCESS RIGHTS INHERITANCE IN INFORMATION SYSTEMS CONTROLLED BY METADATA

Mariya Chichagova, Ludmila Lyadova

Abstract: *All information systems have to be protected. As the number of information objects and the number of users increase the task of information system's protection becomes more difficult. One of the most difficult problems is access rights assignment. This paper describes the graph model of access rights inheritance. This model takes into account relations and dependences between different objects and between different users. The model can be realized in the information systems controlled by the metadata, describing information objects and connections between them, such as the systems based on CASE-technology METAS.*

Keywords: *access control mechanisms, graph model, metadata, CASE-technology.*

ACM Classification Keywords: *D.2 Software engineering: D.2.0 General - Protection mechanisms.*

Introduction

As information systems grow larger and more complex, and as the number of their users increase, there are growing needs for methods that can simplify and even partly automate the process of access rights assignment.

The main problem of traditional access control mechanisms is that they don't take into account the relations between information objects.

The technology METAS [Lyadova, 2003], [Ryzhkov, 2002] allows to create dynamically adjusted information systems. Means of adaptation are based on use of the metadata describing information objects and connections between them from the various points of view. Functioning of information system is carried out through interpretation metadata by MDK METAS The metadata can form a basis for realization of mechanisms of the rights equivalence, in particular, the mechanism of the access rights inheritance that allows to lower labour input of assignment of the rights the manager of system. At the same time there are problems at definition of the rights of access on the objects accessible on several relations from parental objects to which various rights are appointed.

The model proposed in this paper take such relations in consideration and the rules for it are formulated.

To define the inheritance mechanism we shall formally describe model of distribution of the access rights. As basic elements of the model are used access graph and rules of its transformation.

Graph Model

An information system consists of objects and subjects. Access control describes whether specific subject can access specific object.

Let O is a set of objects and S is a set of subjects. $G = (V, E)$ is a finite directed labelled graph, where $V = O \cup S$ is a set of *nodes* and E is a set of *arcs*.

Notation $v_i \rightarrow v_j$ means that there is an arc $(v_i, v_j) \in E$ in graph G. A node $s_i \in S \subseteq V$ is called *subject-node* and a node $o_i \in O \subseteq V$ is called *object-node*.

If $v_i \rightarrow v_j$ and both of these nodes are subject-nodes (or both of them are object-nodes) node v_i is called *parent* of node v_j and node v_j is called *child* of node v_i.

Subject-nodes which have not got any children are called *users* all other subject-nodes are called *groups*.

Object-nodes which have not got any children are called *leaves* all other subject-nodes are called *roots*.

Arcs between objects present the relations between them. The arc's direction depends on type of relationship between objects:

- 1 : 0..1 $\Rightarrow$ arc from "0..1" to "1";
- 1 : M $\Rightarrow$ arc from "1" to "M";
- 0 : 1..M $\Rightarrow$ arc from "0" to "1..M";
- M : M $\Rightarrow$ bidirectional arc.

For example, for part of the database scheme which is shown on fig. 1 the subgraph on fig. 2 is fitted.

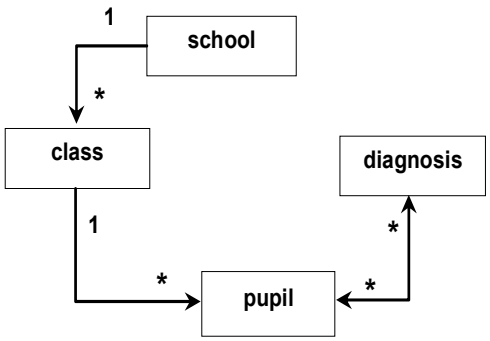


Figure 1. Database Scheme Fragment

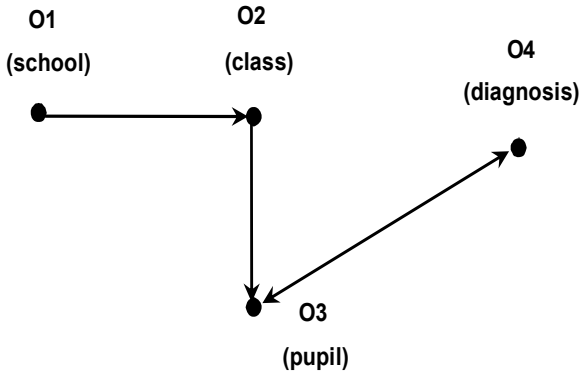


Figure 2. Object-nodes Subgraph

Each of the subjects must be connected by the arc to each of the objects. An arc between subject-node and object-node is called *access arc*.

An access arc's label determines the assigned access right of this subject to the object. *Assigned access right* is determined by the information system's administrator and it can deny or allow access to information objects.

Access rights that take into account relations between subjects and objects are called *actual access rights*.

Subject-nodes

In this section relations between subjects are considered and the rules that can take them into account are formulated. Access rights that allow for relations only between subjects are called *actual subject's access rights*.

Let *parent*(s_i) = {$s_j \in S \mid \exists (s_j, s_i) \in E$} – is a set of parents for subject-node $s_i \in S$.

Let access arcs have following labels:

- *right*(s_i, o_j) ∈ {0, 1, 2, 3} is an *assigned access right* of subject $s_i \in S$ to object $o_j \in O$, where 0 means that right is not assigned, 1 – access is denied, 2 – subject has a partial access and 3 – access is allowed. *Partial access* means that access is allowed only if certain conditions are fulfilled. These conditions are determined by administrator.
- *a_right*(s_i, o_j) ∈ {1, 2, 3} is an *actual subject's access right* of subject $s_i \in S$ to object $o_j \in O$, where 1 means that access is denied, 2 – subject has a partial access and 3 – access is allowed.

The subject's access rights can depend on its parents' rights. The process of determination the actual subject's rights is called *subject's rights inheritance*.

Let $s_i \in S$. The inheritance should be done according to the following two rules.

Rule 1. If subject has an assigned right *right*(s_i, o_j) ≠ 0 to object $o_j \in O$, the actual subject's right is determined as *right*(s_i, o_j), i.e.

$$a\_right(s_i, o_j) = right(s_i, o_j) \quad (1)$$

The main idea of this rule is that explicit assignment is more significant than inheritance.

Rule 2. If an access right to object $o_j \in O$ isn't assigned, i.e. *right*(s_i, o_j) = 0, the actual subject's right is determined as maximum of its parents' actual rights :

$$a\_right(s_i, o_j) = \max_{s_k \in parent(s_i)} (a\_right(s_k, o_j)) \quad (2)$$

If an access right to object $o_j \in O$ isn't assigned and the subject hasn't got any parents its actual right is determined as 1. It means that access is denied.

Note that by definition the maximum value for *a_right* is 3 (that means that access is allowed).

For finding actual subject's rights the above two rules are recursively applied.

Object-nodes

In this section relations between objects are considered and the rules that can take them into account are formulated. Access rights that allow for relations only between subjects are called *actual object's access rights*.

Note that the same object can be connected with different objects and rights can depend on the object from which the access is done.

As access rights depend on access context let define *context*(s_i, o_j) as a current *access context*, i.e. list of object-nodes (path from one of the roots to current object-node).

Parent from context is an object-node from the access context which is the parent for node $o_j \in O$. Let *c_parent*(o_j) is a parent for object-node $o_j \in O$ from context.

Let access arcs also have following labels:

- *o_right*(s_i, o_j) ∈ {1, 2, 3} is an *actual object's access right* of subject $s_i \in S$ to object $o_j \in O$, where 1 means that access is denied, 2 – subject has a partial access and 3 – access is allowed.

Let arcs between object-nodes have the following labels:

- *inherit*(o_k, o_j) ∈ {*true*, *false*} shows the possibility of inheritance. Let *inherit*(∅, o_j) = *false* that means that inheritance from empty object is forbidden.

The process of determination the actual object's rights is called *object's rights inheritance*.

Let $s_i \in S$. The inheritance should be done according to the following three rules.

Rule 3. If subject has an assigned right $right(s_i, o_j) \neq 0$ to object $o_j \in O$, the actual object's right is determined as $right(s_i, o_j)$, i.e.

$$o\_right(s_i, o_j) = right(s_i, o_j) \quad (3)$$

This rule is the same as the rule 1 for subjects' rights inheritance.

Rule 4. If an access right to object $o_j \in O$ isn't assigned, i.e. $right(s_i, o_j) = 0$, and $inherit(o_k, o_j) = false$ where $o_k = c\_parent(o_j)$ the actual object's right is determined as prohibition of access, i.e.

$$o\_right(s_i, o_j) = 1 \quad (4)$$

This rule means that if the inheritance is forbidden in current context and there are no assigned rights the access is denied.

Rule 5. If an access right to object $o_j \in O$ isn't assigned, i.e. $right(s_i, o_j)=0$, and $inherit(o_k, o_j)=true$ where $o_k=c\_parent(o_j)$ the actual object's right is determined as follows:

$$o\_right(s_i, o_j) = o\_rigth(s_i, c\_parent(o_j)) \quad (5)$$

In order to determine actual access rights in rules 3, 4, 5 we should use *a_right* instead *right*.

Conclusion

Using of access rights inheritance allows to simplify the access rights assignment by automatic taking into account relations between object and subject. This method also permits to avoid some kind of mistakes which can be made by information system's administrator.

In addition to the means described in the given article means of the control of correctness of obvious assignment of the access rights for objects and their attributes are offered also [Mikov, 2003].

The architecture of the CASE-system METAS, the models of the metadata used in this system, and principles of its functioning are described in several articles [Lyadova, 2003], [Ryzhkov, 2002].

The system METAS is developed on the .NET platform. The technology ADO.NET, providing independence from DBCS, is used for access to the objects in database.

Bibliography

[Lyadova, 2003] L.N. Lyadova, S.A. Ryzhkov. CASE-technology METAS. In: Scientific Articles Collection "Mathematics of Program Systems". Perm University, Russia, 2003, pp. 4-18.

[Mikov, 2003] A.I. Mikov, M.V. Chichagova. The Control over Rights Assignment. In: Scientific Articles Collection "Mathematics of Program Systems". Perm University, Russia, 2003, pp. 207-212.

[Ryzhkov, 2002] S.A. Ryzhkov. The Concept of the Metadata in the Development of Information Systems. In: Scientific Articles Collection "Mathematics of Program Systems". Perm University, Russia, 2002, pp. 25-35.

Authors' Information

Mariya Chichagova – Perm State University, Assistant of the Department of Computer Science; PSU, 15, Bukirev St., Perm, 614990, Russia; e-mail: chichagova@dom.raid.ru

Ludmila Lyadova - Institute of Computing, Deputy Director; 19/2-38, Podlesnaya St., Perm, 614097, Russia; e-mail: lnlyadova@mail.ru

THE APPLICATION OF GRAPH MODEL FOR AUTOMATION OF THE USER INTERFACE CONSTRUCTION

Elena Kudelko

***Abstract**: The ability of automatic graphic user interface construction is described. It is based on the building of user interface as reflection of the data domain logical definition. The submitted approach to development of the information system user interface enables dynamic adaptation of the system during their operation. This approach is used for creation of information systems based on CASE-system METAS.*

***Keywords**: User interface, metadata, CASE-technology, dynamically adapted information systems, graph model.*

***ACM Classification Keywords**: D.2 Software Engineering: D.2.2 Design Tools and Techniques – Computer-aided software engineering (CASE); G.2 Discrete Mathematics: G.2.2 Graph Theory – Graph algorithms.*

Introduction

The aim of the working out of application user interface is the reflection of the inner structure of information system objects on the level of user understanding about data domain that means the determination of screen objects which let user co-operate with the information system (IS). User interface must include the set of screen forms with the help of which information input and editing can be possible, as well as the navigation system which let data domain objects be catalogued for speeding-up access to them. In this work the approach for automation of the logical description data domain reflection on user interface level, based on the use of graph model, is described. The working out of the models of user interface is made in the context of the creation of CASE-technology METAS based on the metadata, which are multilevel and describe IS from different points. User interface management, described in this work, is based on the metadata of presentation level, which are built on the base of logical level. Both levels can be represented as graphs.

The basic concepts of the logical level are entities, attributes, relation between entities and also instances of these concepts. *Entity* is a type of data domain objects which is characterized by the set of its attributes and relations with entities. For example, entity can be «Person» characterized by the qualities «Surname», «Name», «Birthday», which are attributes of this entity. A person must have an address that means that entity «Person» must be connected with entity «Address». Metadata of the presentation level describe user interface elements: screen forms, controls of different type, navigation system of application. The idea of automatic creation of forms is based on the building of presentation level graph as a reflection of the logical level graph.

Screen Forms

Let us describe the graph of the logical model G_l, on the base of which we will build the graph of the presentation model G_{pr}. The nodes of the logical model graph are corresponded to entities of data domain; there are relations between entities, which are directed arcs in the graph of logical model:

$$G_l = (V_l, E_l), \text{where } V_l = \{e_1, e_2, ..., e_n\},\ E_l = \{r_1, r_2, ..., r_m\},\ n,m \in N$$

$$r_i = (e_j, e_k), \text{where } i = 1..m;\ j,k = 1..n$$

Incoming into the node *e arcs* mean relations of «1:M» type, in which entity *e* is on «M» side that means it is child entity. Outgoing arcs present relations of «1:M» type, in which entity *e* is on «1» side that means it is parent entity. Relations of «M:M» type are represented by two-forked arcs of graph G_l.

Each node of the presentation model graph G_{pr} is a form of some entity; arcs between nodes are possible

transitions between forms (corresponding to arcs in the logical model graph); arcs are directional, direction is given from the form involved to the forms which can be called from the current form:

$$G_{pr} = (V_{pr}, E_{pr}), \text{where } V_{pr} = \{f_1, f_2, ..., f_n\}, E_{pr} = \{r_1, r_2, ..., r_m\}, n, m \in N$$
$$r_i = (f_j, f_k), \text{where } i = 1..m; j, k = 1..n$$

Here pairs (e_j , e_k) and (f_j , f_k) are directed. That means graphs G_l and G_{pr} are oriented.

Graphs G_{pr} and G_l are, in general, multigraphs, in which cycles and loops are possible, and so, the incident nodes of the arc is not enough to identify this arc. That's why arcs must be marked with unique names, for $\forall e_1, e_2 \in V_l$ there must not be two arcs $(e_1, e_2) \in E_l$, having identical types and names. It is the same for graph G_{pr}.

The building process of the presentation model graph is in the reflection of the logical model graph G_l on the set of nodes and arcs of the presentation graph G_{pr}. In every moment the presentation model graph can be determined short. Let us see the building process of a new node f of graph G_{pr}.

The Elementary Graph of the Presentation Model

Node $f \in G_{pr}$ goes to some form for entity $e \in G_l$. This entity can be named the main entity form (we write $ME(f) = e$). So, we have: $(\forall f \in G_{pr})(\exists e \in G_l : ME(f) = e)$. The inverse proposition is also true. If we review the completely specified presentation model graph G_{pr}, i.e. the graph where there are't any undetermined nodes any more, then $(\forall e \in G_l)(\exists f \in G_{pr} : ME(f) = e)$. Arcs, incoming into node e of graph G_l, that means arcs for transition to parent entities in the relation of «1:M», «0..1:M», «0..1:1» type, are included into the presentation model graph (except two-forked graphs). Outgoing and two-forked arcs (arcs connecting reviewing entity with child entities) will be described later. Any arc in graph G_{pr} goes to some arc in graph G_l, i.e. $(\forall r_i = (f_j, f_k) \in E_{pr})(\exists r_i = (e_k, e_j) \in E_l : ME(f_j) = e_j, ME(f_k) = e_k)$.

Such graph model corresponds to the simplest case when for the reflection of every entity its own form is used and transitions between forms are realized by the relations of «1:M», «0..1:M», or «0..1:1» type, which exists between main entities of forms, in direction from child entity (from «M» side) to parent one (to «1» side). Let us view the example in Figure 1.

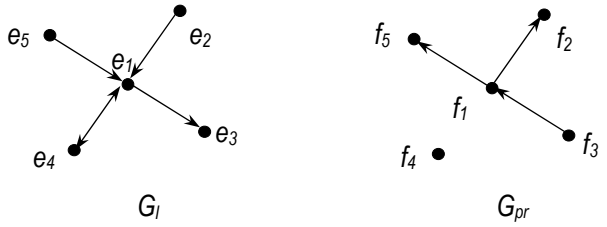


Figure 1. Reflection of the logical model graph on the presentation model graph

In graph G_l there are 5 nodes corresponding to 5 entities of data domain. Its own form is built for each of them: $e_i = ME(f_i)$, where $i = \overline{1,5}$. Entity e_1 has got two parent relations with entities e_2 and e_5, entity e_3 has got one parent – entity e_1. Therefore, there will be three arcs in graph G_{pr}, corresponding to the relations «1:M» of graph G_l. The building of graph G_{pr} goes according to gradual addition of nodes and its' arcs into sets V_{pr} and E_{pr} accordingly. That's why while building new node $f \in V_{pr}$ for node $e \in V_l$ it can happen that there is no node $f' \in V_{pr}$ for node $e' \in V_l$ yet, to which we must prolong arc from node f. Missing node f' must be built to avoid «hanging» arcs. We mark built node f' as indefinite (*«udef»*). Such node has got one or several incoming arcs (in particular, arc $(f, f') \in E_{pr}$).

Attributes Reflection

Besides relations with other entities every entity is characterized by the set of its attributes. These attributes must be reflected by the elements of user interface application and must be included in the presentation model.

The creation of node *f* in graph G_{pr} is the reflection of logical model attributes to the controls of the presentation level. Any node $e \in V_l$ has got the set of attributes $Attr(e) = \{a_0, a_1, ..., a_n\}, n \in N$, where attribute a_0 is the key attribute of entity ($a_0 = key(e)$). In common case entity can have several key attributes (a compound key). But this situation can be taken to the viewed case by the addition of artificial key.The key attribute is used only on the logical level and is inaccessible for user on the presentation level.

Any non-key attribute a_i, where $i = \overline{1,n}$, can be either own attribute of the entity (the set of such attributes – $Attr_{own}(e)$), or an attribute, realizing the relation with parent entity, i.e. outer attribute ($Attr_{parent}(e)$). The key attribute is also in the set of own attributes. All own attributes have got a type $Type(a_i)$, which sets possible attribute values and operations, applicable to these values, and also the method of attribute values input and output. Every attribute a_i connecting with parent node $e' \in V_l$ is corresponding to any incoming in node *e* (but not in two-forked) arc $(e', e) \in E_l$ ($rel(a_i) = (e', e)$). The reflection of such arcs in graph G_{pr} is set by the algorithms, described in the work. We can also speak about the type of such attribute. The type coincides with the type of the key attribute of parent entity e'.

Node *f* of graph G_{pr} includes in itself the set of the controls $AttrCtrl(f) = \{ac_1, ..., ac_n\}$, corresponding with attributes of entity $e = ME(f)$. The key attribute does not go into the presentation model. Each control will have the type, defining by the type of corresponding attribute or parent relation.

Compound Forms

Sometimes it is comfortable to include into the form information not only about one object of data domain but also the information connected with this object, i.e. the information about objects of several entities. Then each node of the presentation model graph will correspond to the subset of nodes of the logical model graph. The fulfillment of such reflection depends on the semantics of data domain and it can be done by user-administrator.

On the base of the built presentation model graph the extended graph can be built, nodes of which will be compound forms. Let us view the process of building such a graph. Let us have node *f* of graph G_{pr}, which corresponds to node *e* of graph G_l, i.e. $ME(f)=e$. Let's mark by $G_l(f)$ *some* set of nodes and arcs, complying with node *f* of the presentation model graph. In graph $G_l(f)$ every node is corresponded to node of the logical model graph, and arc is corresponded with the arc of the logical model graph. Now in the set of nodes there can be several nodes, corresponding to one and the same entity. The same is for arcs: in the set of arcs of graph $G_l(f)$ there can be several arcs, reflecting one and the same arc of the logical model graph. The set of nodes and arcs of graph $G_l(f)$ is marked correspondingly $V_l(f)$ and $E_l(f)$. Initially, the set of nodes of graph consists from one node, corresponding to entity *e*, and the arcs set is empty: $V_l(f) = \{ME(f)\}, E_l(f) = \phi$.

Let us view the arbitrary node *e* of the logical model graph G_l. Node *e* can be connected with the other nodes of graph G_l, i.e. there is a set of incident arcs to this node. Let us break this set on two subsets: the set of incoming arcs $E_{parent}(e) \subset E_l$ and the set of outgoing and two-forked arcs $E_{child}(e) \subset E_l$. The first set connects node *e* with the set of parent entities for this node

$$E_{parent}(e) = \{rp_1, ..., rp_n\}, n \in \{0\} \cup N, \text{where } rp_i = (e_i, e), e_i \in V_l, i = \overline{1,n},$$

and the second set – with the set of child entities

$$E_{child}(e) = \{rch_1, ..., rch_n\}, n \in \{0\} \cup N, \text{where } rch_i = (e, e_i), e_i \in V_l, i = \overline{1,n}.$$

In the example in Figure 1, relations (e_5, e_1) and (e_2, e_1) make up the set of parent relations for entity e_1, and relations (e_1, e_3) and (e_1, e_4) – the set of child relations.

Let us suppose that it is necessary to reflect on the form the information about the child entity of the main entity of form *f*. Let us examine the main entity *e* (*ME(f)=e)* and identical node to it e_{ch}, representing child entity for entity *e*. So, $\exists r_{ch} \in E_{child}(e) : r_{ch} = (e, e_{ch})$.

For including of node e_{ch} into graph $G_l(f)$ it is necessary to:

1. Include node e_{ch} into the set $V_l(f)$, and arc r_{ch} – into the set $E_l(f)$;
2. Include arc (f, f_{ch}), where $ME(f)=e$, $ME(f_{ch})=e_{ch}$, into the set E_{pr}. This arc will correspond to arc $(e, e_{ch}) \in E_l$.

To delete node e_{ch} from graph $G_l(f)$ it is necessary to fulfil the reverse transformation of the sets $V_l(f)$, $E_l(f)$ и E_{pr}:

1. Exclude arc (f, f_{ch}), where $ME(f) = e$, $ME(f_{ch})=e_{ch}$, from the set E_{pr};
2. Delete arc r_{ch} from the set $E_l(f)$, node e_{ch} must be excluded from the set $V_l(f)$.

Let us now have node e_p, incidental to the main entity *e* of form *f* (*ME(f)=e)* which is parent for it, i.e. $\exists r_p \in E_{parent}(e) : r_p = (e_p, e)$. Including and deleting node e_p from the set $V_l(f)$ occurs in the following way. For adding parent node to the main entity of the form it will be enough:

1. Include node e_p into the set $V_l(f)$, and arc r_p – into the set $E_l(f)$;
2. Include arcs, corresponding to the relations between node e_p and its parent nodes, i.e. for $\forall r \in E_{parent}(e_p)$ we include arc (f, f_r), where $ME(f_r) = e_r, r = (e_r, e_p)$;
3. Delete arc (f, f_p), where $ME(f)=e$, $ME(f_p)=e_p$, from the set E_{pr} of graph G_{rp}. This arc corresponds to arc $(e_p, e) \in E_l$.

Deleting of parent node from the main entity of the form includes the following steps:

1. Include arc (f, f_p), where $ME(f)=e$, $ME(f_p)=e_p$, into the set E_{pr} of graph G_{rp}. This arc corresponds to arc $(e_p, e) \in E_l$;
2. Exclude arcs corresponding with relations of node e_p with parent nodes, i.e. for $\forall r \in E_{parent}(e_p)$ we delete corresponding to it arc (f, f_r), where $ME(f_r) = e_r, r = (e_r, e_p)$;
3. Delete arc r_p from the set $E_l(f)$, node e_p must be excluded from the set $V_l(f)$.

Let us give the example, showing described operations of including of nodes into the form structure for entity *e*.

In Figure 2a the logical model graph is shown. Let us discuss the fragment of the presentation model graph, built for the form entity *e*. Figure 2b shows the simplest presentation model graph, built according the logical model graph. In Figure 2c there is the form structure *f* after including into it child relation (e, ech_1) and, so, adding relation (f, fch_1) into graph G_{pr}. In Figure 2d parent relation (ep_2, e) is included into the form structure, besides in graph G_{pr} connections (f, fp_{21}), (f, fp_{22}) are added and connection (f, fp_2) is deleted.

Similarly we can view adding and deleting operations for other entities, including in subgraph $G_l(f)$. In order to view these operations in details, we introduce the definition of parent entity of level *n*.

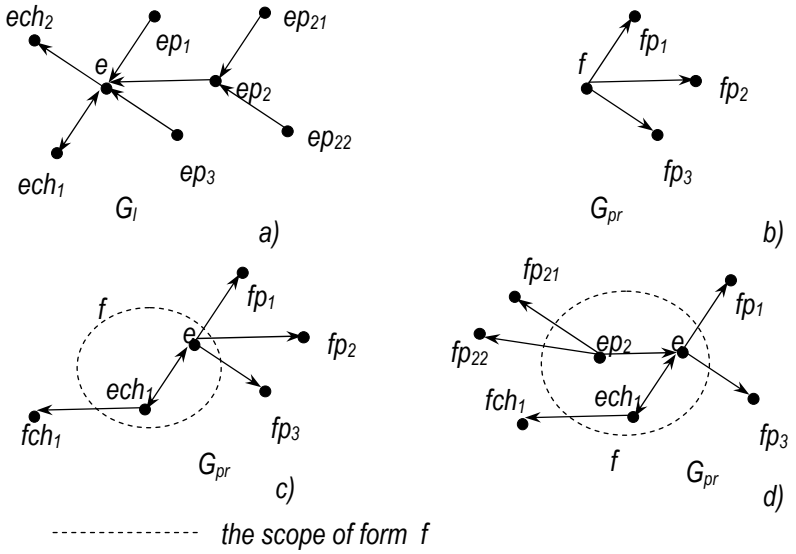


Figure 2. Including parent and child entities into the form structure

Let us have entity $e \in V_l$. The parent of the first level for this entity can be any entity, relation of which with entity *e* is in set $E_{parent}(e)$. If entity e_p^{n-1} is a parent of level *n-1* for entity *e*, then entity $e_p^n : (e_p^n, e_p^{n-1}) \in E_{parent}(e_p^{n-1})$ will be a parent of level *n* for entity *e*. The parent set of level *n* for entity *e* we will mark $E_{parent}^n(e)$. $E_{parent}^1(e) = E_{parent}(e)$.

For example, in Figure 2a $E_{parent}^3(ech_2) = \{(ep_{21}, ep_2), (ep_{22}, ep_2)\}$.

The discussed above operation of the addition of node $e_p \in E_{parent}(e)$ to the set $V_l(f)$ was described for e_p, which is the parent of the first level for node *e*. For node e_p, including in the set $V_l(f)$, and also for any node from the set $V_l(f)$, which is a parent of arbitrary level for node *e*, we can define the similar operations of deleting\adding parent and child entities.

We can examine node $f \in G_{pr}$ and entity *e=ME(f)*. We will take any node e', which is parent of arbitrary level of node *e*. We must notice that described below algorithms will be right for the case, when $e' = e$.

Let us have node e_{ch}, which is incidental to node e' of form *f* and which is child for her:

$$r_{ch} \in E_{child}(e') : r_{ch} = (e', e_{ch}).$$

For including node e_{ch} into graph $G_l(f)$ it is necessary to execute the following algorithm:

1. Add node e_{ch} into the set $V_l(f)$, and arc r_{ch} – into the set $E_l(f)$;
2. Add arc (f, f_{ch}), where $ME(f)=e$, $ME(f_{ch})=e_{ch}$, into the set E_{pr}. This arc corresponds to arc $(e', e_{ch}) \in E_l$;

In order to delete node e_{ch} from graph $G_l(f)$ it is necessary to fulfill the reverse transformation of the sets $V_l(f)$, $E_l(f)$ and E_{pr}:

1. Exclude arc (f, f_{ch}), where $ME(f)=e$, $ME(f_{ch})=e_{ch}$, from the set E_{pr};
2. Delete arc r_{ch} from the set $E_l(f)$, node e_{ch} must be deleted from the set $V_l(f)$.

Let us now have node e_p, incidental to node e' of form *f* and which is parent for it, i.e. $r_p \in E_{parent}(e') : r_p = (e_p, e')$.

Addition of parent node into the form structure is in the following:

1. Add node e_p into the set $V_l(f)$, and arc r_p – into the set $E_l(f)$;
2. Add arcs, corresponding to relations between node e_p and parent nodes, i.e. for $\forall r \in E_{parent}(e_p)$ we add

(f, f_r), where $\mathrm{ME}(f_r) = e_r, r = (e_r, e_p)$;

3. Delete arc (f, f_p), where *ME(f)*=*e*, *ME(f$_p$)*=*e$_p$*, from the set E_{pr} of graph G_{rp}. This arc corresponds to arc $(e_p, e') \in E_l$.

Deleting of parent node from form structure consists of the following steps:

1. Add arc (f, f_p), where *ME(f)*=*e*, *ME(f$_p$)*=*e$_p$*, into the set E_{pr} of graph G_{pr}. This arc corresponds to arc $(e_p, e') \in E_l$;
2. Delete arcs, corresponding to the relations between node e_p and parent nodes, i.e. for $\forall r \in E_{parent}(e_p)$ we delete corresponding arc to it (f, f_r), where $ME(f_r) = e_r, r = (e_r, e_p)$;
3. Delete arc r_p from the set $E_l(f)$;
4. Fulfill cascading deleting of all parent nodes which are in $G_l(f)$, i.e. apply recursively the algorithm to every node $e_p^i \in V_l(f) : e_p^i \in E_{parent}(e_p)$;
5. Delete cascadely all child connections of node e_p, which are in $G_l(f)$, i.e. we must apply the deleting algorithm of child node to every node $e_{ch}^i \in V_l(f) : e_{ch}^i \in E_{child}(e_p)$;
6. Node e_p must be excluded from the set $V_l(f)$.

Described above the rules of adding nodes can be used in series until the expansion of graph $G_l(f)$ is possible, i.e. till set nodes $V_l(f)$ have parent and child nodes and connections which can be included into graph $G_l(f)$. Node *e* (and incidental to it arc) can be included into $G_l(f)$, if the graph does not have the path, including the same consecution of nodes and arcs, starting in the root, as well as the way from root node up to including node *e*.

So the repeated bringing in of one and the same path to graph $G_l(f)$ is impossible. And graph $G_l(f)$ is a tree in which the root is a node, corresponding to the main entity of form.

Attributes Reflection in Compound Forms

Including of parent node $e \in V_l$ into graph $G_l(f)$ of any node $f \in V_{pr}$ is a reflection of the attributes of the logical model on the controls of the presentation level. Deleting node *e* from graph $G_l(f)$ is a deleting of corresponding controls. Node *f* of graph G_{pr} includes a set of controls $AttrCtrl(f) = \{ac_1, ..., ac_m\}, m \in N$, corresponding to attributes of entities of the set $V_l(f)$. So,

$$AttrCtrl(f) = \bigcup_{e \in V_l(f)} AttrCtrl(f, e).$$

When adding new parent node e_p from relation (e_p, e'), where $e' \in V_l(f)$, to set $V_l(f)$:

1. The element corresponding to parent relation (e_p, e'), i.e. $ac \leftrightarrow a \in Attr(e') : rel(a) = (e_p, e')$ is deleted from the set of controls.
2. We add controls *ac* into the set of controls of node *f*, corresponding to all non-key attributes *a* of entity e_p, i.e. $\forall a \in Attr(e') : a \neq key(e_p)$ into the set *AttrCtrl(f)* we put $ac \leftrightarrow a$.

When deleting parent node e_p, taking part in relation $(e_p, e') \in E_l(f)$ from the set $V_l(f)$:

1. We delete all controls, corresponding to entity e_p from the set of controls.
2. Into set *AttrCtrl(f)* we add $ac \leftrightarrow a : rel(a) = (e_p, e')$.

Entity Tree

A tree is a building of hierarchy on the set of entities and relations between them. Together with the set of forms the set of tree nodes must provide access to any entity. The described structure is a tree only on a user's screen (its' name comes from here). From the point of view of the structure it is an oriented graph $G_T = (V_T, E_T)$, maybe with cycles. The set of nodes $V_T = (nd_1, nd_2, ..., nd_n)$ includes two types of nodes (two subsets). The first subset $V_T^g \subset V_T$ consists of grouping nodes, the second subset $V_T^o \subset V_T$ has got object nodes. Any object node corresponds to the entity of the logical model and so, the form of the presentation level. Correspondence of object nodes and nodes of the logical model graph can be given as function $Ent : V_T^o \to V_l$. Then $(\forall nd \in V_T^o)(\exists e \in V_l : Ent(nd) = e \wedge \exists f \in V_{pr} : ME(f) = e)$. In that way, form object node *nd* we can draw arc *(nd, f)* to corresponding node-form *f*. The set of such arcs forms the additional set (let us call it E_{ext}), connecting two graphs G_{pr} and G_T. Nodes of the logical model do not correspond to nodes of the entity V_T^g. Such nodes are only created for convenient reflection of the information on the user's screen. There are arcs between nodes of the set V_T. Arcs can exist as between nodes of one subset (V_T^o или V_T^g), as also between nodes of different subsets. There is one group node in the tree from which the tree scanning starts. Such node can be marked as a *root*. The root node does not have incoming arcs. Any node of the set V_T is reachable from the root, i.e. between any node of the set V_T and root node there is a path.

The path between two object nodes can correspond to subgraph of the logical model which includes some sequence of arcs and nodes, which are between entities, which are reflected by two viewed nodes.

For example, in Figure 3, graph of the logical model G_l can be corresponded to the graph of the object tree G_T. In the tree graph object nodes e_1, e_2, e_3 are named corresponding to the names of nodes-entities of graph G_l, which they correspond to. Arc $(e_1, e_3) \in G_T$ is corresponded to the path $\langle e_1, r_{14}, e_4, r_{43}, e_3 \rangle$, arc (e_2, e_2) – to the path $\langle e_2, r_{22}, e_2 \rangle$, arc (e_2, e_3) – to the path $\langle e_2, r_{23}, e_3 \rangle$.

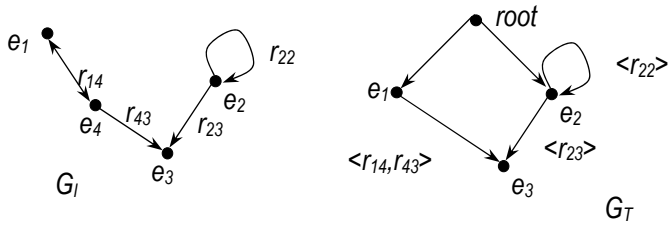


Figure 3. Example of entity tree

Let us view two object nodes $nd, nd' \in V_T^o : Ent(nd) = e, Ent(nd') = e'$, between which there is a path $\langle nd_1, (nd_1, nd_2), nd_2, (nd_2, nd_3), ..., (nd_{n-1}, nd_n), nd_n \rangle, nd_1 = nd, nd_n = nd', n \in N$. This path is corresponded to the path of the logical model graph $\langle e_1, (e_1, e_2), e_2, (e_2, e_3), ..., (e_{m-1}, e_m), e_m \rangle, e_1 = e, e_m = e', m \in N$.

Every node in graph G_T can be reachable from the root by different paths, and each of these paths can be concerned with the path in the logical model graph. While building the tree on the user's screen, i.e. during data loading, only important in this context relations are considered. While building a tree part of the ways can be reflected on the user's screen and the other part serve for assignment of additional dependence. Let us say that additional ways of the logical model connect with the way of graph G_T, consisting from one arc. For unification of

the ways assignment we can also take that visual paths in graph G_T do not have corresponding paths in the logical model graph and all entities relations specify by non-visual additional connections.

Let us make it clear on the example. There is a fragment of graph G_T, shown in Figure 4. Here while building the branch which includes nodes *root, nd1, nd2, nd3*, for loading nodes of type nd_3 relations (nd_3, nd_2) and (nd_3, nd_1) are used, but while loading the branch which goes over nodes *root, nd4, nd5, nd2, nd3*, relations (nd_2, nd_5) and (nd_2, nd_4) are used, while loading nodes of type nd_2 and the relation (nd_3, nd_2) while loading nodes of type nd_3.

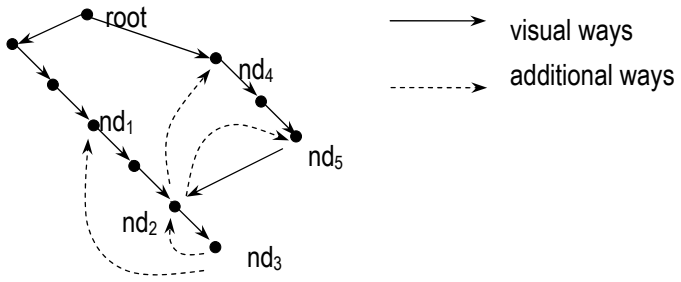


Figure 4. Paths assignment in the objects' tree

In such a way, a tree is a kind of visual presentation of entities' connections and it reflects the view of the end user on the interconnection of the objects of data domain (or vice versa on the absence of the connection between some entities).

At last we get extended graph of the presentation model G'_{pr} (V'_{pr} – the set of its' nodes, E'_{pr} – the set of arcs), consisting of nodes and arcs of graphs G_{pr} and G_T; besides, this graph includes the set of arcs between object tree nodes and nodes-forms:

$$G'_{pr} = (V'_{pr}, E'_{pr}), \text{where } V'_{pr} = V_{pr} \cup V_T, E'_{pr} = E_{pr} \cup E_T \cup E_{ext}.$$

Conclusion

This work describes the base model of user interface of application, including objects of user interface and specifying interconnection between them. Based on described operations with graph of the presentation model there have been worked out additional algorithms that are not described in the work.

In order to realize suggested above graph model of user interface we can offer the approach of CASE-technology METAS, including in itself tools for automation of working out large scale IS, based on using multilevel structure of dynamically changing metadata. Case-tools contain the set of program components, processing metadata of different levels and building on this base application, which has the meanings of interface designing and graphic interface of the end user of IS. Program components, realizing Windows-interface of applications, are built based on the present graph and they use algorithms suggested above.

Author's Information

Elena Kudelko – Perm State University, Department of Computer Science, Assistant; 15, Bukirev St., Perm, 614990, Russia; e-mail: kudelko_elena@mail.ru

IMPLEMENTING AJAX BASED SPREADSHEET ENGINE WITH RELATIONAL BACK-END[1]

Ivo Marinchev

***Abstract**: In this paper we represent our implementation of a web based spreadsheet engine that uses state-of-the-art technologies for development of interactive server-centric applications – AJAX and web services. Initially an overview of web application technologies and spreadsheet-centric application development is presented. Then a general architecture of a web based spreadsheet engine is discussed. Finally we introduce one concrete implementation of this architecture that is suitable for small and middle-size deployment scenarios.*

***Keywords**: spreadsheet engine, AJAX, web applications, web services, rich-clients, web 2.0.*

Introduction

Recently, there is an obvious tendency of moving more and more applications to the web based technologies. As a result the paradigm of building "classical" desktop applications turns into the paradigm of building web-based applications. New applications are projected from the very beginning as web applications and many old ones are re-implemented or extended to web-based analogues. In reality web applications are actually server-centric applications that can be started and run through regular web browser. A distinguishing feature of the web applications is that significant parts of their code is located and executes on remote servers (or cluster of servers) and only user interface related components are transferred through the network to the users' location (system) and executed there.

The first distributed applications with centralized core application logic (business logic) were rich-client applications. They require more sophisticated and responsive user interface than what HTML, CSS, and JavaScript can offer at that time and run on the users' computers communicating with a business logic that is physically located on the centralized servers. Rich client applications and technologies appear about 10 years ago but could not become widespread. They remain in use mainly in intranet environment inside the organizations or shared between affiliate organizations. Their failure to become widespread was due to many factors some of which are:

- they require the users to install additional software on their systems;
- security concerns;
- not having enough support from big application vendors;
- high price tag.

The most prominent technologies in this area are Microsoft's ActiveX, Java Web Start (JWS), Eclipse Rich Client Platform (Eclipse RCP), and Macromedia Flex.

Recently widespread adoption of web applications became feasible with the introduction of several key technologies in practically all of the modern web browsers - Internet Explorer, Firefox, Mozilla, Opera, Safari, and Konqueror. These technologies are CSS [CSS1, CSS2], JavaScript [JavaScript], DOM [DOM] and DHTML (dynamic screen re-flow). Despite most of them were available even 6 or 7 years ago, the biggest boost started just recently with the widespread adoption of so named XMLHttpRequest [AJAX] object. It allows web pages (utilizing JavaScript) to perform asynchronous request to their originating server and fetch updated data from it.

[1] The research has been partially supported by "Technologies of the Information Society for Knowledge Processing and Management" - IIT-BAS Research Project No. 010061.

On the next step these data are used to update part of the web page information in timely and responsive manner without requiring page reloads. As the XMLHttpRequest object has support for transferring data in the XML format (presentation/view neutral encoding) the corresponding technology was named AJAX [AJAX] (Asynchronous JavaScript And XML). Hence it becomes possible to develop web applications that look and feel in a way very similar to the regular desktop ones, providing the user with similar usage experience and capabilities. Keeping the business logic on the centralized servers allows the highest level of security, easier management, support, maintenance, and upgrades. New schemes for application distribution, delivery and usage have become feasible – pay-per-use, application service providers, click-and-run (no installation is required), application delivered as a service, etc.

The first widespread web applications were Google Mail (www.gmail.com) and Flickr (www.flickr.com). Other more recent and complex web applications are Writely (www.writely.com) word processing application (since March 2006 owned by Google), web calendar applications - 30 boxes (www.30boxes.com), CalendarHub (www.calen darhub.com), and many others. In the field of web based spreadsheets applications the key players are NumSum (www.numsum.com), and iRows (www.irows.com), and open source applications TrimSpreadsheet (http://trimpath.com /project/wiki/TrimSpreadsheet), WikiCalc (http://www.softwaregarden.com/wkcalpha). Unlike the rest of the web applications, at the moment of this writing (April 2006), web spreadsheets are not feature complete, but built with specific purposes in mind, and non-customizable. Open source ones are mostly unusable and are actually just a proof of concept than real applications.

Spreadsheet-Centric Application Development (SCAD)

SCAD is a software development methodology that uses the spreadsheet component as the primary user interface for the application. The following is a list of key features (presented in no particular order) of SCAD:

- Rapid Application Development - the spreadsheet engines are extremely rich in functionality. Today's SCAD components provide functionality on the par with the best off the shelf spreadsheet applications such as Excel. This provides for a vast array of features to tap into and use within the SCAD software.
- Excel-Compatibility – because of the dominance of Microsoft Office in today's marketplace, spreadsheet engines always provide some degree of compatibility with Excel. This feature provides a powerful and easy way to use Excel as design tool for initial development of the application or for easy transition from legacy Excel based applications to their web based analogues.
- Calculation Engine - the vast array of calculation functions that exist within the spreadsheet engine provide an extremely powerful tool to create a robust and reliable calculation engine that churns through complex algorithms effectively and expeditiously.
- Segmented Programmability - it is not only entirely possible but also feasible to segment the spreadsheet development from control coding and assign them to professionals with varying degrees of expertise. In the case of the former, the spreadsheet professional must be proficient in design and formulation of the core spreadsheet files while the latter must have command of the native language used to control and customize the SCAD environment.
- Cost-Efficiency - spreadsheet development generally requires less time and expertise and therefore it is more cost-efficient than control coding.
- Familiar Look and Feel - the spreadsheet interface is one of the most familiar if not the most familiar look and feel in the computing arena today. Developing a user interface based on the spreadsheet grid provides an additional level of comfort to the users and positively impacts their introduction into the SCAD software.

SCAD has been around since 1983 (with advent of lotus 1-2-3). This development methodology was never widely adopted because companies such as Lotus or Microsoft never promoted it as such. The marketers within these

companies designated these packages as “end-user” products and it was more profitable for them to segment the market in this fashion.

With the advent of Microsoft Office and Visual Basic for Applications (VBA), SCAD took a giant leap into being recognized as serious application development tool. But Microsoft continued to resist the complete independence of VBA from the Office application by refusing to create runtime versions of Excel or Word.

Having recognized this gap in the market, several alternatives have emerged to fulfill a demand for SCAD. The best, most reliable, fastest, and most Excel-compatible of these products is Formula One e.Spreadsheet Engine. It is geared toward delivering enterprise reporting applications, which presently are the most visible adoption of SCAD. Formula One is implemented in Java and can be used as a component in Java desktop applications and applets.

In the following sections we introduce our implementation of a web-based spreadsheet engine. Unlike the applications mentioned above the architecture of our application is not restricted to particular language or environment. We employ only ubiquitously adopted standards and loosely coupled component architecture that allows building of extremely portable and flexible application that can be deployed in many different usage scenarios.

System Architecture

Fig. 1 depicts general architecture of our system. It comprises of 3 tiers – user interface layer, application layer and persistent storage layer.

User interface layer consists of all of the code and software components that are executed on the clients’ workstations (i.e. run client-side). This is usually a thin layer which means it is mainly a code that is related to the interface presented to the user and translates user interactions to the corresponding messages that are sent to the application layer. There are many possible implementations of this layer but they are divided in two general groups – web based (HTML, Javascript) and rich clients (ActiveX, Java Web Start, Macromedia Flash). Web based implementations can run out-of-the box (they need just a modern web browser) whereas the rich clients usually require additional software to be installed beforehand on the users’ workstations. In both cases communications between the user interface layer and the application layer employ some ubiquitous (programming) language independent high-level protocol stack as HTTP and REST [REST] or SOAP web services. Language independence of the communication protocols between the layers allows different implementation to be changed easily and different implementations to be used by different groups of clients. For example a typical deployment scenario can allow internal or trusted users to use rich client interface whereas the rest be restricted to use a more restricted and secure web interface.

Application layer comprises all algorithms and internal data structures that are involved in spreadsheet management and processing. Upon client requests (that come from the user interface layer) it fetches the data from the persistent storage layer and builds internal data model of the manipulated spreadsheets. Then it uses these data models to re-calculate the spreadsheets values based on the user changes and sends the updated data (user changes) back to the persistent storage layer and forth (recalculated fields values) to the user interface layer.

Persistent storage layer contains database logic for storing, retrieving organizing and managing internal application layer’s data structures on persistent media. This layer is normally implemented with the help of some kind of data management systems. In practice they can be relational databases or xml databases. The communication protocols can be SQL and XPath or XQuery correspondingly. It is even possible this layer to be implemented with web services as well. The later will decouple the application layer from the need to know the exact representation of the database back-ends used.

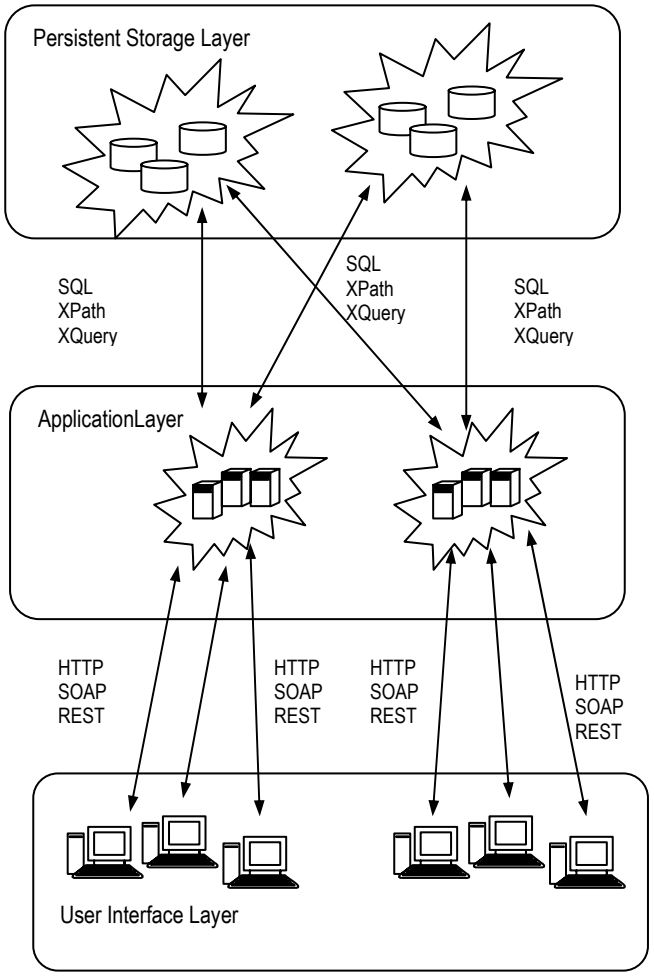


Fig.1. 3-tier architecture of our system is suitable for small and middle-size deployments.

Implementation

Our implementation of afore mentioned architecture complies with the following basic requirements:

- Many different people must be able to use it from any physical location provided Internet connection and a contemporary version of widespread web browser is available.
- The engine must support at least 3 different user roles (groups of users):
  - Spreadsheet Designers - they can create/import and edit spreadsheets' definitions.
  - Regular Users – they can use available spreadsheets i.e. just filling the data in the cells that are not locked.
  - System Administrators – manage all of the aspects of the system that are not application specific – installation, deployment, maintenance, support, etc.
- The data entered by the regular users must be kept separated from the spreadsheet definitions so that it can be reused in different spreadsheets and/or other related applications.
- Ajax version must be able to work on IE and Firefox. If it is possible Opera and Safari must be supported as well.
- The system must be able to import MS Excel spreadsheets and convert them into its internal form – reuse already created spreadsheets and use MS Excel as a primary design tool until comparable web based spreadsheet designer is developed.

The current version of our implementation is a standard 3-tier web application that utilizes the following technologies:

- *User interface layer* – implemented as a JavaScript library that uses Ajax requests to send, update, and fetch data to/from the application layer. It works on IE (5.5, 6.0), Firefox (1.0, 1.5) and to some degree on Opera (8.5). It is possible to build Java Web Start client in the future.
- *Application layer* – this layer contains all of the algorithms and data structures that implement the core spreadsheet engine functionality. Upon client requests (coming from the user interface layer) it fetches the data from the persistent storage layer and builds internal data model of the manipulated spreadsheets. Later it uses these data models to re-calculate the spreadsheets values based on the user changes and sends the updated data (user changes) back to the persistent storage layer and forth (recalculated fields values) to the user interface layer. Currently this layer consists of PHP pages (for stateless services) and Java servlets (for statefull services that manipulate big data models). The implementation uses REST services. Although REST is not a standard but an architectural style, its light-weighted, requires fewer resources, and is simpler and faster for quick-and-dirty implementations. Later we can convert inter-layer communications to the complete SOAP, WSDL, WS-I stack if it is needed. This layer contains also all of the algorithms and data structures that dynamically build user interface pages (screens) of the system. Another responsibility is getting and validating request parameters from the user interface layer and reformatting the response data if it is needed.
- *Persistent storage layer* – it uses relational database storage. It was tested with MySQL and SQL Server, but as it uses standard SQL queries it should work with any complaint relational database system. This layer contains also a tool that creates spreadsheet definitions from existing MS Excel spreadsheets.

Figures 2 and 3 show the same spreadsheet opened in MS Excel and its converted version opened with our application. Our conversion tool preserves not only computational logic and cell merging but also visual formatting as much as possible. It is required as our tool has not full-featured visual designer yet. We use MS Excel for initial creation of the spreadsheet visual representation.

Fig. 4 shows another spreadsheet opened in IE with the help of our web application. This screenshot depicts one unique feature of our system – it can show the cell's formula as tooltip when the mouse pointer hovers over the corresponding cell. Although it is currently not implemented, the same technique can be used for visualization of other types of information – for example, displaying the inter-cell dependencies as a tree rooted in the current cell with a nodes and leafs the cells it depends on in the corresponding order and level.

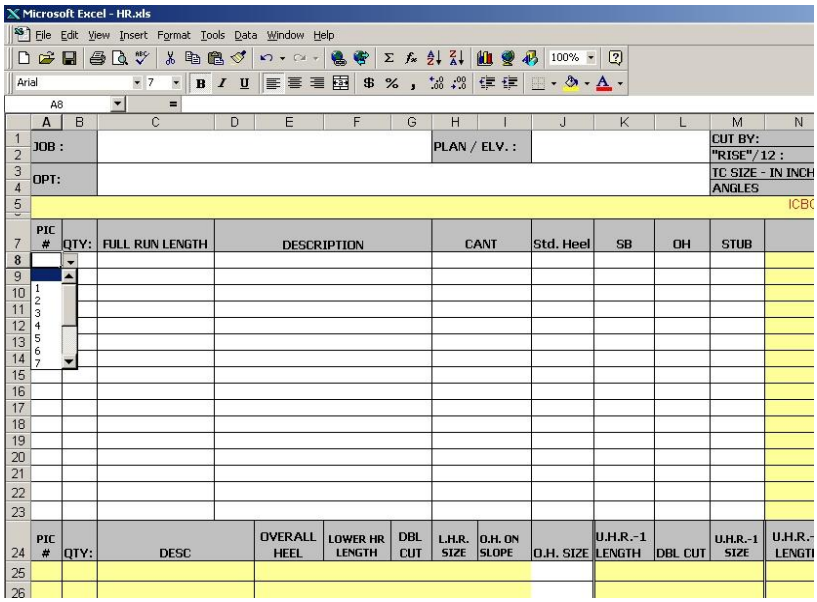


Fig.2. A complex spreadsheet opened within Microsoft Excel.

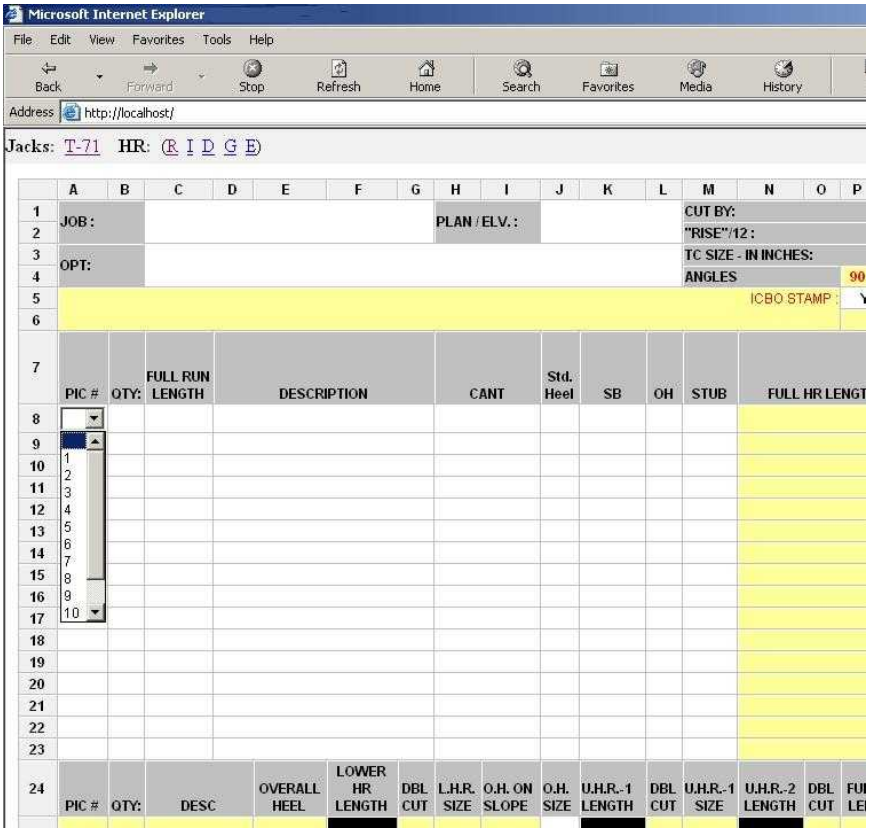


Fig.3. The spreadsheet from fig. 2 converted and managed by our system. Opened within Internet Explorer

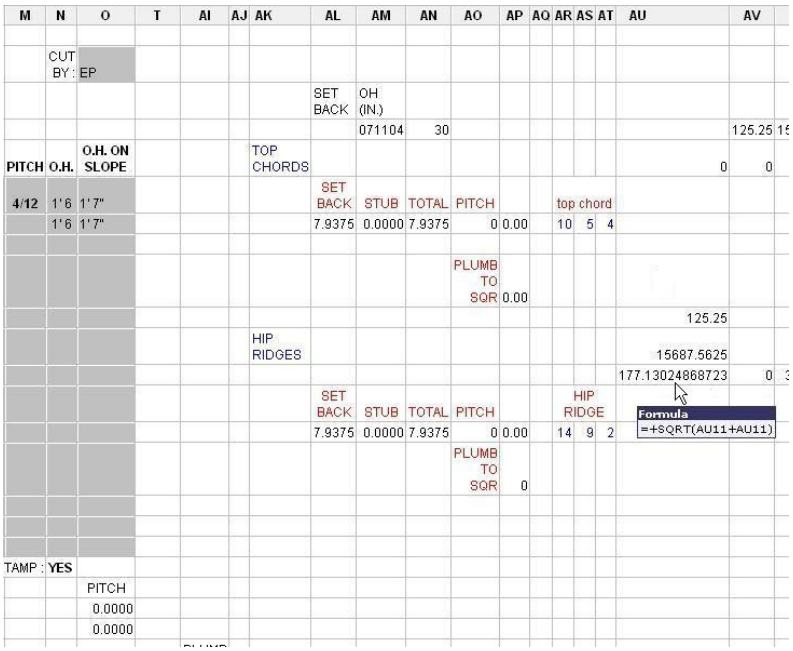


Fig.4. One unique feature of our system is showing cell formulas as tooltip when mouse hovers over the cell.

Conclusion and Future Work

In this paper we presented our implementation of web-based spreadsheet engine. It employs many contemporary technologies for building scalable and responsive web applications. Several improvements and additions need to be implemented in order to make a system more usable and user friendly. Among them are - support more browsers (Opera, Safari); support more MS Excel features and formulas; add some unique features as the ability to use CGI scripts or web services in formulas (getting weather conditions, exchange rates in real-time); many other optimizations and improvements.

Bibliography

[AJAX] AJAX, http://en.wikipedia.org/wiki/AJAX/

[CSS1] CSS1 Specification, http://www.w3.org/TR/REC-CSS1/

[CSS2] CSS2 Specification, http://www.w3.org/TR/REC-CSS2/

[DOM] W3C Document Object Model, http://www.w3.org/DOM/

[JavaScript] ECMA-262, ECMAScript (JavaScript Specification), http://www.ecma-international.org/publications/standards/Ecma-262.htm

[REST] R. Fielding. Architectural Styles and the Design of Network-based Software Architectures. PhD Thesis, University of California, Irvine, 2000

Author's Information

Ivo Marinchev – Institute of Information Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev Str., Bl. 29A, Sofia-1113, Bulgaria; e-mail: ivo@iinf.bas.bg

DIGITAL ART AND DESIGN

Khaled Batiha, Safwan Al-Salaimeh, Khaldoun A.A. Besoul

Abstract*: The desire to create unique things and give free rain to one's imagination served as a powerful impetus to the development of digital art and design software. The commoner was the use of computers the wider variety of professional software was developed. Nowadays the creators and computer designers are receiving more and more new and advanced programs that allow their ideas becoming virtual reality. This research paper looks at the history of the development of graphic editors from the simplest to the most modern and advanced. This brief survey includes the history of different graphic editors' creation, their features and abilities. This paper highlights the two basic branches of graphic editors – these that are in free use and commercial graphic editors design software. The researcher selected the most powerful and influential graphic editors design software brands like Paint.NET and GIMP among free software and commercial Adobe Photoshop. This paper also dwells upon the way digital art transferred from the exclusively professional business into the hobby for ordinary users. This research paper bears implications for those who are interested in features and potentiality of most popular graphic editors design software.*

Keywords: *Digital Art, Graphic information, DPaint, Image Manipulation Program, Paint Shop Pro, Photopaint, Photoshop.*

1. Introduction

Imagination is extremely refined work of the human mind. It is the easiest medium for to creation out of nothing. Human mind constantly works on creating something that has never existed before and does not now exist. This is the approach with which any professional creator of Digital Art, or in other words, creator of design will gain success. Digital culture is neither new nor determined by technology, but rather that technology is a product of digital culture. The term "digital" originally referred to data organized in discreet units in any system, linguistic, and numerical systems included.

Since the use of computers became an everyday occurrence the wide variety of software has been emerging to assist designers. From the simplest and primitive up towards professional graphics editors computer software has undergone the complicated evolution and development. I shall not mention vector graphics editors; however I'd prefer to concentrate on bitmap graphics editors, which are mainly used to produce images.

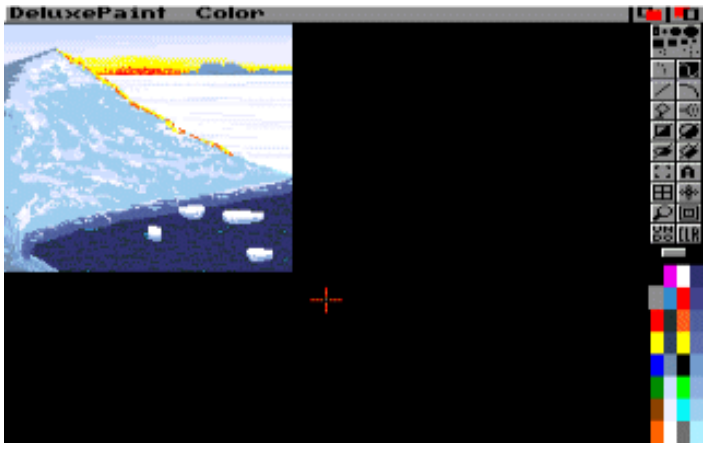
Graphic information is stored in computer memory in "bitmap" or "raster" formats such as JPEG, PNG, GIF and TIFF. Besides that, every company that creates graphics editor sets up its own format of storing raster graphics.

I would like to start the review of graphics editors with RIP editors, those that currently represent solely historical interest. Among the first editors the following deserve mentioning Deluxe Paint, Personal Paint and Photogenic.

2. Deluxe Paint

Deluxe Paint (DPaint) is a bitmap graphics editor created by Dan Silva for Electronic Arts (EA) [1]. The original version was created for the Amiga OS and was released in November 1985.

DPaint was the product of an in-house art development tool called Prism. As Silva added more features to Prism, it started to have market-place potential. When the Amiga was released in 1985, DPaint was quickly released for it. It was quickly embraced by the Amiga community and became the standard graphics development tool for the platform. Amiga manufacturer Commodore International later struck a deal with EA to have DPaint (and later its four "sequels", versions 2, 3, 4 and 5) bundled with every new Amiga sold. This deal lasted until Commodore's bankruptcy in 1994.



Screenshot and image designed in Deluxe Paint.
Taken from http://amiga.emucamp.com/dpaint4.htm.

The program DPaint enables us to create gradients, draw in anti-alias mode, change the palette, make "stencils", and transform any group of pixels into a "brush." It also allows special brush techniques "smooth" and "smear," features that are also found on Adobe Photoshop. The maximum number of colors we can work with is 256, which makes it satisfactory program for altering GIF images.

Other two programs Personal Paint and Photogenics had similar characteristics. Thus I will not dwell on them.

3. Free Graphics Design Software

Some significant position is occupied by graphics editors considered as free software. One can mention here Paint.NET i GIMP.

Paint.NET [2] is a project developed at Washington State University and mentored by Microsoft. It is a free graphics editing program for use on Windows XP and 2000 based operating systems, with the source freely available for download. It is programmed in C# and is released under the open source MIT License. Paint.NET is the unofficial successor to the older Microsoft Paint graphics program.

Graphics editors GIMP [3] deserve more particular attention. At the same time I am going to draw up some comparison of its abilities with those of Adobe Photoshop. The **GNU Image Manipulation Program** or **The GIMP** is a bitmap graphics editor and also has some support for vector graphics. The project was started in 1995 by Spencer Kimball and Peter Mattis and is now maintained by a group of volunteers; it is licensed under the GNU General Public License.

Overview

GIMP originally stood for General Image Manipulation Program; in 1997, the name was changed to GNU Image Manipulation Program. It is an official part of the GNU project.

The GIMP can be used to process digital graphics and photographs. Typical uses include creating graphics, resizing and cropping photos, changing colors, combining images using a layer paradigm, removing unwanted image features, and converting between different image formats.

The GIMP is also notable as perhaps the first major free software end-user application. Previous work, such as GCC, the Linux kernel, and so on, were mainly tools by programmers for programmers.

Features

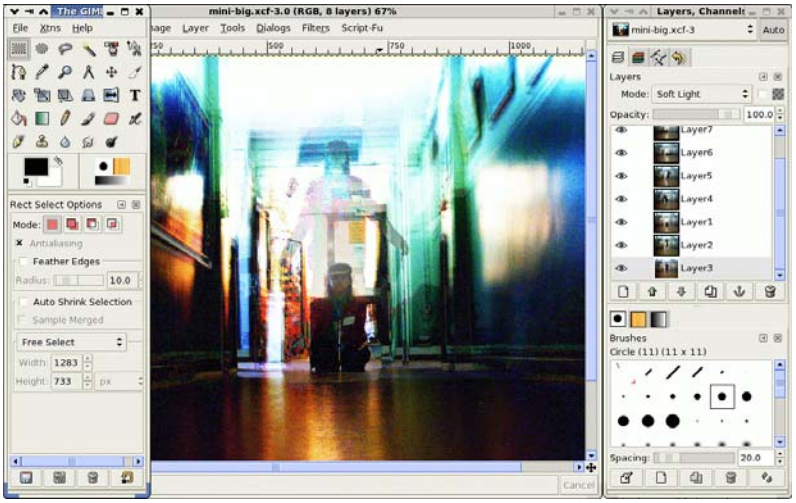
The GIMP was intended as a free (as in speech) alternative to Adobe Photoshop, but the latter still dominates in the printing and graphics industries:

- Photoshop includes licensed support for the Pantone color matching system.
- The number of plugins and other add-ons available for Photoshop is larger.
- GIMP has only experimental CMYK separation support.
- GIMP has almost no spot color support.
- GIMP has limited gamma support.
- GIMP has limited color management through LCMS

There is a plugin called PSPI for the Microsoft Windows version of the GIMP only, which allows the use of the 8bf Adobe Photoshop filters in the GIMP.

The peculiarity of graphic design is the ability of graphics to interact with different projects. It is necessary to mention that such feature became indispensable at Internet technologies development. As well as interactive use, the GIMP can be automated with macro programs. The built-in Scheme can be used for this, or alternatively Perl, Python and Tcl can also be used. This allows the writing of scripts and plugins for the GIMP which can then be used interactively; it is also possible to produce images in completely non-interactive ways (for example generating images for a webpage on the fly using CGI scripts) and for batch color correction and conversion of images.

The current (as of March 2005) stable version of the GIMP is 2.2.7. Major changes compared to version 1.2 include a more polished user interface and further separation of the user interface and back-end. For the future it is planned to base GIMP on a more generic graphical library called GEGL, thereby addressing some fundamental design limitations that prevent many enhancements such as native CMYK support.



Screenshot of The GNU Image Manipulation Program 2.0.0 running on XFce on Linux

4. Adobe Photoshop

The most outstanding graphic image editor is **Adobe Photoshop** [4]. **Adobe Photoshop** is a bitmap graphics editor (with some text and vector graphics capabilities) developed and published by Adobe Systems. It is the market leader for commercial bitmap image manipulation. As with most of Adobe's other applications, Photoshop is available for Mac OS and Microsoft Windows; versions up to Photoshop 7 can also be used with operating systems such as Linux using software such as CrossOver Office. Past versions of the program were ported to the SGI IRIX platform, but official support for this port was dropped after version 3.

Features

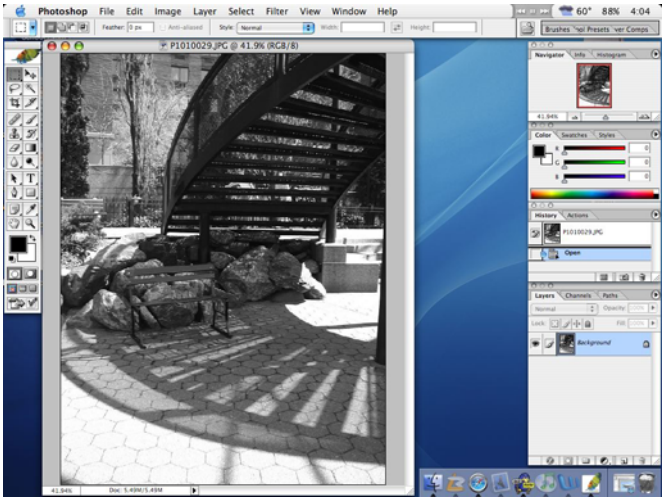
Although primarily designed to edit images for paper-based printing, Photoshop is used increasingly to produce images for the World Wide Web. Recent versions bundle a related application, Adobe ImageReady, to provide a more specialized set of tools for this purpose.

Photoshop also has strong links with software for media editing, animation and authoring. It works with Adobe Illustrator, Adobe Premiere, Adobe After Effects & Adobe Encore DVD to make professional standard DVDs, provide non-linear editing and special effects services such as backgrounds, textures and so on for television, film and the web. Photoshop's native file format (PSD or PDD) can be exported to and from Adobe Illustrator, Adobe Premiere, After Effects and Adobe Encore DVD. Photoshop CS broadly supports making menus and buttons for DVDs. For PSD or PDD files exported as a menu or button, it only needs to have layers, nested in layer sets with a cueing format and Adobe Encore DVD reads them as buttons or menus.

PSD or PDD is a widely accepted file format. Competing bitmap image editing programs (such as Macromedia Fireworks, Corel Photo-Paint, Discreet Combustion, WinImages, GIMP, etc.) can import and edit layered PSD or PDD files.

The most recent version, as of 2006, is version 9. This iteration of the program is marketed as "Photoshop CS2". In an effort to break away with previous versions of the application and to reinforce its belonging with the new line of products, Photoshop even dropped one classic graphic feature from its packaging: the Photoshop eye, which was present in different manifestations from versions 4 to 7. Photoshop CS versions now use feathers as a form of identification.

Photoshop CS features a revolutionary command: 'Shadow/Highlight' which allow user to 'suppress' highlights and/or 'push out' shadows while maintaining most of the 'image details' (i.e. the histogram would remain virtually unchanged).



Screenshot of the Photoshop CS under Mac OS X

5. From Novelty To Everyday Use

Only one and a half decade ago paint programs were treated like novelty and something very exquisite. But the development of such software turned out to be extremely rapid. Thus for comparatively short period they became an every-day need for representatives of a fine art community. Even more, the development of graphics editors caused the emergence of a new term “photoshopping”. The term photoshopping is a neologism, meaning "editing an image", regardless of the program used. The name comes from Adobe Photoshop, the image editor most commonly used for the practice, although other programs, such as *Paint Shop Pro*, *Photopaint*, or the *GIMP* may be used.

The practice of photoshopping is possible because modern image editors made the work of altering images extremely easy, particularly with the clone tool. Nowadays actually anyone who possesses elementary computer skills can use photoshopping to edit his photographs.

Professional photographers also use photoshopping in their work. Thus, practically, no photograph in the magazines Popular Photograph, Nature Photographer, Close Up and others is free of retouch. The example of professional photograph, retouched with the help of Photoshop is presented below:

Fine Art Photograph by Richard Seiling Breaking Clouds, Half Dome (20 x 24 inches)
Image is take from http://www.yosemitestore.com/

Photographer Richard Seiling captured this image in Yosemite Valley, as sunlight broke through the clouds of a winter storm. Taken on 4x5 transparency film, Rich scanned the image into the computer, and performed traditional darkroom techniques, such as dodging and burning, using Adobe Photoshop. This image comes matted and overmatted on 8-ply white Archival Mat Board.

It also should be mentioned that today no well-colored magazine can do without the imaged edited in graphics editors. The results of such work are apparent everywhere – from ads in the magazines to the billboards.

Conclusion

The present study presents a concise review of historical development of graphic editors with the particular consideration of the most representative examples. The comparative approach to the most powerful graphic editors that represent two different, in principle, branches of software - free (Paint.NET and GIMP) and commercial (Adobe Photoshop) revealed that potentialities of commercial software is still leading on software market due to its advanced and newest features that satisfy the most refined aspirations of professional users. That's why the professionals prefer Adobe Photoshop; however, the amateurs may be well satisfied by Paint Shop Pro, Photopaint, or the GIMP. The last but not the least point is that when making a choice for a particular digital art and design program one should remember the rapid progress of this kind of software and the abilities that used to be pertinent to expensive commercial products now are the characteristics of more simple free software products.

References

[1] http://amiga.emugaming.com/dpaint.html official site of Electronic Arts Company

[2] http://www.eecs.wsu.edu/paint.net/ official site of Paint.NET

[3] http://www.gimp.org/ official site of the GIMP

[4] http://www.adobe.com official site of the Adobe Company

In the research the material from the following sites was used:

http://www.guerillapixel.com/pages/digital_ill_pages/software_mainr.htm

http://screamdesign.daz3d.com/

http://www.photoshopcafe.com/PhotoshopCS2.htm

Authors' Information

Khaled Batiha – e-mail: batihakhalid@yahoo.com

Safwan Al-Salaimeh **–** e-mail: safwan_71@yahoo.com,

Khaldoun A.A. Besoul – e-mail: dr_khaldoun69@hotmail.com

Irbid National University, Faculty of information technology, Irbid, The Hashemite Kingdom of Jordan

IMAGE PARTITIONS METRIC PROPERTIES IN IMAGE UNDERSTANDING PROBLEMS

Vladimir Mashtalir, Vladislav Shlyakhov

Abstract: *A new distance function to compare arbitrary partitions is proposed. Clustering of image collections and image segmentation give objects to be matched. Offered metric intends for combination of visual features and metadata analysis to solve a semantic gap between low-level visual features and high-level human concept.*

Keywords*: partition, metric, clustering, image segmentation.*

Introduction

There has been a tremendous growth of the image content analysis significance in the recent years. This interest has been motivated mainly by the rapid expansion of imaging on the World-Wide Web, the availability of digital image libraries, increasing of multimedia applications in commerce, biometrics, science, entertainments etc. Visual contents of an image such as color, shape, texture and region relations play dominating role in propagation of feature selection, indexing, user query and interaction, database management techniques. Many systems combine visual features and metadata analysis to solve the semantic gap between low-level visual features and high-level human concept, i.e. there arises a great need in self-acting content-based image retrieval task-level systems.

To search images in an image database traditionally queries 'ad exemplum' are used. In this connection essential efforts have been devoted to synthesis and analysis of image content descriptors. However, a user's semantic understanding of an image is of a higher level than the features representation. Low-level features with mental concepts and semantic labels are the groundwork of intelligent databases creation. Short retrieval time independent of the database size is a fundamental requirement of any user friendly content-based image retrieval (CBIR) system. Characteristics of different CBIR schemes, similarities or distances between the feature vectors of the query by example or sketch and those of the images collection are sufficiently full explored [see, e.g. 1-3]. To optimize CBIR schemes it is necessary to minimize a total number of matches at a retrieval stage. Thus there arises a problem to find novel partition constructions for the fast content-based image retrieval in video databases and furthermore we have be able to compare different partitions.

A Metric for Partitions Matching

As retrieval is computationally expensive, one of the most challenging moments in CBIR is minimizing of the retrieval process time. Widespread clustering techniques allow to group similar images in terms of their features proximity. The number of matches can be greatly reduced, but there is no guarantee that the global optimum solution is obtained. We propose clustering of image collections with objective function encompassing goals to number of matches at a search stage.

The problem is in that under given query $y \in \mathrm{Y}$ one needs to find the most similar image (or images) $x_v \in \mathrm{X}$. In other words, it is necessary to provide $min_{v \in \mathrm{V}} \rho(y, x_v)$ (here $\rho(\circ,\circ)$ is arbitrary distance function, V is an indexing set) during minimum possible warranted time. If $\mathrm{Y} \subseteq \mathrm{X}$, the exact match retrieve is required. We shall name elements $[\mathrm{X}]_\alpha$, $\alpha \in \mathrm{A}$ of power set 2^{X} as clusters if they correspond to the partition of set X. Let us consider such partitions that any elements of one cluster do not differ from each other more than on ε, i.e. $\forall x' \neq x''$ we have $[x'] = [x'']$, if $\rho(x', x'') \leq \varepsilon$ and $[x'] \cap [x''] = \varnothing$ otherwise. The given or obtained value ε used at a clustering stage is connected with required accuracy of retrieve δ, if it is specified, as follows. There arise two cases:

$\delta > \varepsilon$ – any representative of the cluster nearest to the query y can be used as the image retrieval result, i.e. minimal number of matches is defined by the number of clusters; in other words it is necessary to provide

$$N_1 = card\ \{[\mathrm{X}]_\alpha\} \to min; \quad (1)$$

$\delta \le \varepsilon$ – the element of more detailed partition will be the result of the image retrieval. In simplest situations it is necessary to fulfill a single-stage clustering, i.e. to optimize retrieval under worst-case conditions we have to ensure

$$N_2 = card\{[\mathrm{X}]_\alpha\} + max(card\ [\mathrm{X}]_\alpha) \to min. \quad (2)$$

At the multilevel clustering the repeated clusters search inside of already retrieved clusters is fulfilled and only on the last step required image is searched by complete enumeration. Let us assume that the cluster $[\mathrm{X}^{(i-1)}]_p$ is selected on $(i-1)$ level of hierarchy from a condition $\rho(y, [\mathrm{X}^{(i-1)}]_q) \to min$, $q = \overline{1, card\{[\mathrm{X}^{(i-1)}]\}}$, i.e. $[\mathrm{X}^{(i-1)}]_p = \ = [\mathrm{X}^{(i)}]_1 \cup [\mathrm{X}^{(i)}]_2 \cup \ldots \cup [\mathrm{X}^{(i)}]_{\alpha_p}$ where for any k and l the equality $[\mathrm{X}^{(i)}]_k \cap [\mathrm{X}^{(i)}]_l = \varnothing$ holds. Then the minimization of matches amount is reduced to the clustering with the goal function

$$N_3 = \sum_{i=1}^{m-1} \left\{ card\ [\mathrm{X}^{(i)}]_{p,(i)} \middle| x \in [\mathrm{X}^{(i-1)}]_{p,(i-1)} \right\} + max\ \left(card\ [\mathrm{X}^{(m-1)}]_{p,(m-1)} \right) \to min, \quad (3)$$

where m is a number of hierarchy levels, $[\mathrm{X}^{(0)}]_{1,(0)} = \mathrm{X}$. The method of search (1) was proposed in [4], the solution of problem (2) was offered in [5], searching of (3) one can see in [6].

Content of an image may be often summarized by a set of homogeneous regions in appropriate feature space. Therefore, there is a great need for automatic tools to classify and retrieve image content on the base of segmentation.

Segmented images are formed from an input image by gathering its elements into sets likely to be associated with meaningful objects in the scene. That is, the main segmentation goal is to partition the entire image into disjoint connected or disconnected regions. Unfortunately, the effectiveness of their direct interpretation depends heavily on the application area and characteristics of an acquisition system. Possible high-level region-based interpretations are associated with a priori information, measurable region properties, heuristics, plausibility of computational inference. Whatever the case, often it is necessary to have dealings with a whole family of partitions and we must be able to compare these partitions which are produced by a variety of segmentation algorithms. At least splitting and merging techniques make us to match segmentation results which ultimately may be corresponded to indirectly images comparisons.

For region-based similarity analysis novel approaches are required since usually early processing scheme consists of following steps: images are segmented into disjoint (or weakly intersecting) regions, features are extracted from each region, and the set of all features is used for high-level processing. It should be emphasized that quite often simultaneous processing of partitions or coverings is wanted to produce reliable true conclusion. In this connection we propose and vindicate a new metric providing all partitions (and consequently images) matching

$$\rho(\mathrm{P}, \mathrm{Q}) = \sum_{k=1}^{m} \sum_{l=1}^{n} |\, \mathrm{X}_k \Delta \mathrm{Y}_l \,|\, |\, \mathrm{X}_k \cap \mathrm{Y}_l \,| \quad (4)$$

where Δ denotes a symmetric difference, $\mathrm{P} = \{\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_m\}$, $\mathrm{Q} = \{\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_n\}$ $\mathrm{X}_k, \mathrm{Y}_l \subseteq \mathrm{D}$, D is a field of view (generally, arbitrary finite signal or feature space with found partitions). Note, these partitions are segmentation results, representing pairwise disjoint family of nonempty subsets whose union is the image and each subset may contain required target, may belong to a carrier of object image or may be a part of that. From this follows to provide possibilities of reliable low-level feature selection and reasonable semantic concepts accommodation often it is necessary to analyze partition collections.

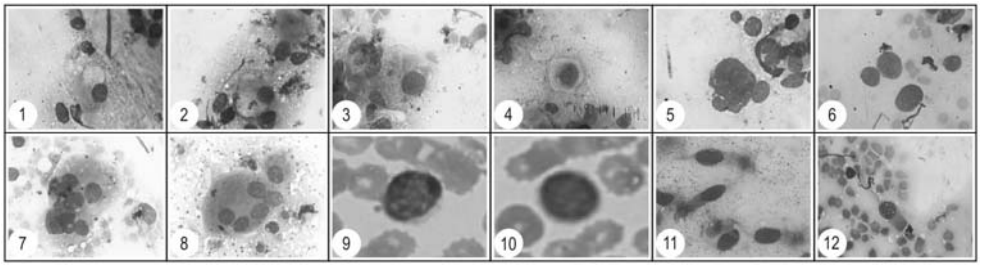


Fig. 1. Examples of input images

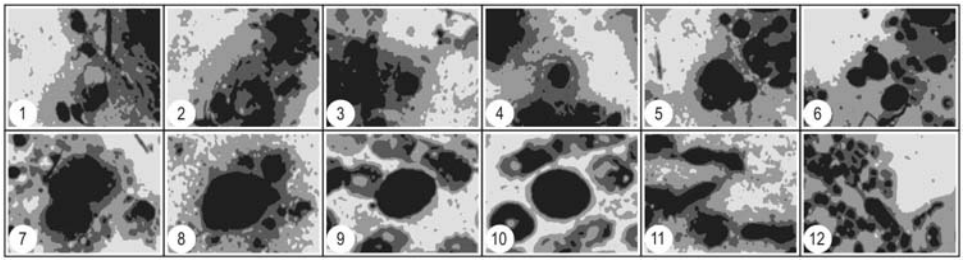


Fig. 2. Multithresholding segmentation of images shown in fig. 1

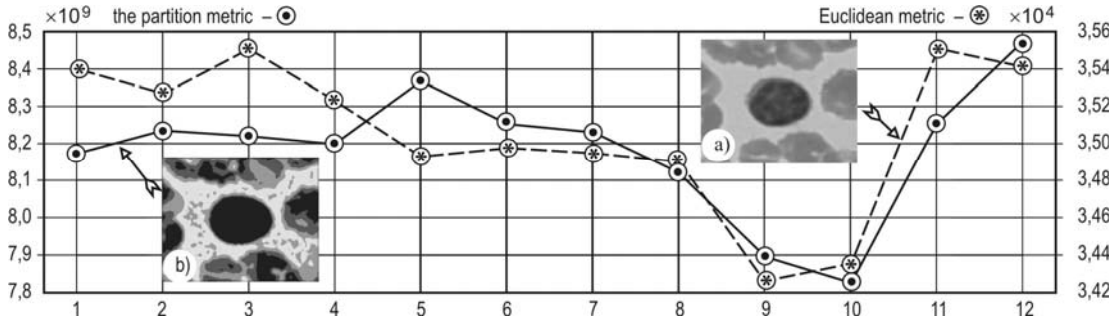


Fig. 3. Examples of image and partitions matches (query a) and b) correspondingly)

The analysis of experimental results has shown that the application of partitions as features provides a sufficient relevance at access to an image in database with queries 'ad exemplum'. Figures 1 and 2 illustrate images and partitions which were compared by means of traditional and proposed metrics. Examples of dependences, query image and its partition are shown in fig. 3. We see comparability of obtained results for Euclidean metric and distance function (4). The reliability of matching can be increased by an intellectual processing (via relations analysis of region-based models) which provides conditions for entirely correct and complete segmentation.

Conclusion

An intensive experimental exploration with the collection of histologic specimens images with final goal classification as an aid in cancer detection vindicates the efficiency of proposed metric.

Bibliography

1. Müller H., Müller W., Squire, D.McG., Marchand-Maillet S., Pun T. Performance Evaluation in Content-Based Image Retrieval: Overview and Proposals. Pattern Recognition Letters, Vol. 22, 2001, pp. 593-601.
2. Li J., Wang J.Z. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, 2003, pp. 1075-1088.
3. Käster T., Wendt V., Sagerer G. Comparing Clustering Methods for Database Categorization in Image Retrieval. Lecture Notes in Computer Science, Springer-Verlag, Berlin Heidelberg, Vol. 2781, 2003, pp. 228-235.

4. Mashtalir V.P., Yakovlev S.V. Point-Set Methods of Clusterization of Standard Information. Cybernetics and Systems Analysis, Kluwer Academic Publishers, Vol. 37, 2001, pp. 295-307.
5. Kinoshenko D., Mashtalir V., Yegorova E. Clustering Method for Fast Content-Based Image Retrieval. Computer Vision and Graphics, K. Wojciechowski et al., eds., Computational Imaging and Vision, Springer, Vol. 32, 2006, pp. 946–952.
6. Kinoshenko D., Mashtalir V., Yegorova E., Vinarsky V. Hierarchical Partitions for Content Image Retrieval from Large-Scale Database. Machine Learning and Data Mining in Pattern Recognition, P. Perner., A. Imlya, eds., Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol. 3587, 2005, pp. 445-455.

Authors' Information

Mashtalir Vladimir – Doctor of Technical Sciences, Professor of Computer Science Department and Dean of Computer Science Faculty, Kharkiv National University of Radio Electronics, Lenina ave, 14, Kharkiv, Ukraine, 61166, e-mail: mashtalir@kture.kharkov.ua

Shlyakhov Vladislav – Candidate of Technical Sciences (equivalent Ph.D.), Senior Researcher of Computer Science Department, Kharkiv National University of Radio Electronics, Lenina ave, 14, Kharkiv, Ukraine 61166.

DEVICE FOR COUNTING OF THE GLASS BOTTLES ON THE CONVEYOR BELT

Ventseslav Draganov, Georgi Toshkov, Dimcho Draganov, Daniela Toshkova

Abstract*: In the present paper the results from designing of device, which is a part of the automated information system for counting, reporting and documenting the quantity of produced bottles in a factory for glass processing are presented. The block diagram of the device is given. The introduced system can be applied in other discrete productions for counting of the quantity of bottled production.*

Keywords*: device for counting, automated information system*

Introduction

In all discrete productions it is needed the ready production to be counted as well as reporting and documenting of the received data. In the present paper a device for counting the quantity of the produced glass bottles, moving on conveyor belt and which is designed by the authors is presented. It is a part of the automated information system for reporting and documenting of the ready production in a factory for glass processing [Draganov, 2006]. The information system has to meet following requirements: collecting data for the ready production, moving in one direction on the conveyor belts; archiving the data for each shift; reporting the quantity of the production for a shift (eight hours).

Different company developments of production counting systems are known [Solid Count, 2006; Fast Counts, 2006; Patent 0050111724, 2005]. One of them is the system SolidCount™ [Solid Count, 2006], which is designed for an automatic collecting of data for the ready mixed (of different kinds) production from a single production line, reporting the quantity of the production and receiving statistical data for the production in real time. The system Fast Count™ [Fast Counts, 2006] serves for: collecting data from several lines; reporting of the quantity of the production in different formats; monitoring of the productivity; archiving of the data; statistics and diagnostics in real time. For counting of the ready production a method and apparatus for counting is suggested in [Patent 0050111724, 2005]. The data for the ready production are received by comparison between the image of the product on programmable zoned arrays of light sources and photo detectors and known images.

The software and hardware products, which are considered, are of general use. They are expensive, very complicated and less reliable. These disadvantages are avoided in the system for counting, reporting and documenting of ready production, moving in unidirectional way on four conveyor belts as well as the entire production of the factory for glass from the four conveyor belts. The system is developed by the authors and it is introduced in a factory for glass.

Structural Diagram of the Automated Information Systems

The structural diagram of the automated information system in the factory for glass is depicted in Fig.1.

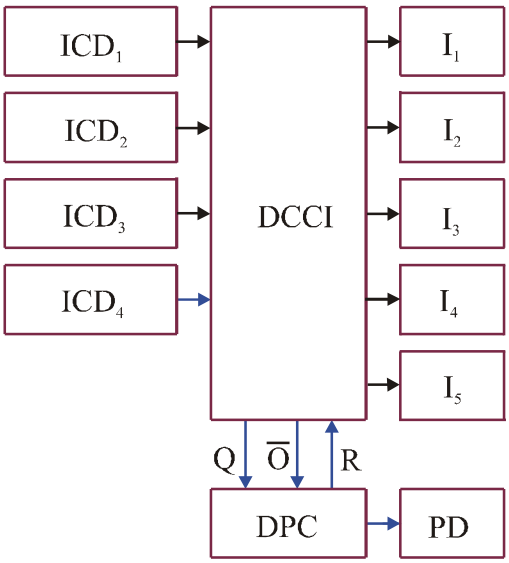


Fig.1 Structural diagram of automated system for counting of bottles on conveyor belt

Each of the four input conversion devices (ICD) feeds an electric impulse to the device for counting control and indication (DCCI) when a ready production unit passes the conveyer belt in front of the input conversion device.

In DCCI information about the quantity of the impulses, which have come from the four ICD, is gathered. On the basis of this information the necessary signals for control of the indications I1÷I5 are depicted. In the presence of danger of overflow of any of the counters, registering the input impulses, DCCI sends a signal for overflow ($\overline{O}$) to the device for printing control (DPC). The last also receives information for the state of all counters in DCCI (Q). DPC gives a command to the printing device (PD) for printing the results and after that to DCCI – a command to clear the counters (R). The printing with consequent clearing is also accomplished by external signal from an operator through clearing button, lying on the command panel, which is a part from the DCCI, at the end of the shift. In case of power failure DPC saves the current information and after restoring the electricity supply the necessary commands for printing and clearing are passed to PD.

Scheme Solution of the Device for Counting of Bottles

To receive reliable information for the quantity of the produced glass bottles it is necessary each input converting device from the automated information system for reporting and documenting of the quantity of ready production to be designed. The device has to meet the following requirements: to convert the information for the number of the glass bottles, which move on the conveyor belt separately or in groups in electrical impulses with TTL level in contactless way; the number of the electrical impulses to correspond strictly to the number of the passing glass bottles and errors, caused by bottles, which are contiguous one to another or by the uneven optical density of the glass from which the bottles are made or by vibrations of the conveyor belt have to be expelled; the device to be

simple and cheap at most and with high reliability of the scheme solution; the construction to be with high mechanical stability and manifacturability.

The main goal of the work is to design a device, which meets the attached requirements and free of the indicated disadvantages.

Devices for counting of objects, based on electro-contactable, capacitive, inductive and other principles are known. One of the most perspective one is the photo-converting principle, which has following advantages: broad field of application; contactless way of operation; high reliability and long exploitation time; high promptitude; low feeding voltages and small consumption of electrical power; broad temperature range of operation; possibility for miniaturization and integration and etc.

The photo-converting devices frequently operate in a mode of transmission [Bergmann, 1980], in which the counted objects cross and modulate a ray, emitted from light source to a light receiver, situated on the other side of the object. There is a possibility for operating in another mode – mode of reflection [Bergmann, 1980], in which the light source and the light receiver are situated on one and the same side of the moving object, reflecting directly or diffusely part of the light, emitted by the light source to the light receiver. An operation in a mode of autonomous emitting [Bergmann, 1980] at which the object itself is a light source is possible.

The photo-converters may operate with unmodulated and modulated light [Bergmann, 1980]. The schemes of the photo-converters with unmodulated light are simplified but they are adversely influenced by the disturbing light – daylight or artificial, emitted by other sources of light. The photo converters with modulated light are protected from the influence of the disturbing light in a high degree, but their scheme solution is complicated and expensive.

In the designed device the photo converting principle of operation, based on mode of transmission of the unmodulated light is used. Thus a simplified scheme solution is obtained.

The disadvantages of principle of the devices operating with unmodulated light are not substantial in the concrete case as the application of the device to be designed is characterized by a small distance between the light source and the light receiver and lack of parasitic lighting. For the purpose an appropriate construction is developed.

The possible errors, caused by vibrations of the conveyor belt and by the uneven density of the bottles may be avoided by transmission of light ray at the height of the mouth of the bottles. But even in this case the light ray is discontinued repeatedly when a single bottle is passing and the number of the obtained output impulses is arbitrary.

Scheme solutions by which this disadvantage may be avoided – with using of integrator, by their processing with monostable multivibrator are known. The difficulty in using them in the concrete case is caused by their irregular movement of the conveyor belt because of the vibrations, which strongly hampers the specifying of the time constant of the delay circuitry.

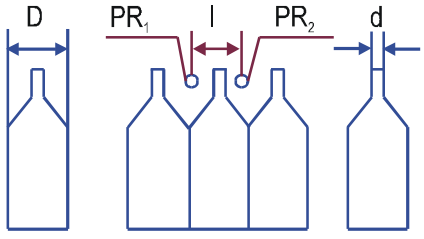


Fig.2 Scheme solution of the device for counting of bottles.

The problem is solved by using of $\overline{R}$ - $\overline{S}$ trigger, to which both inputs impulses are entering from the both photoconverters (Fig.2). Each photoconverter contains emitter and receiver. When the mouth of the bottle passes between the source and the receiver of the first photoconverter PR_1, the light ray is discontinued repeatedly. The obtained output impulses enter the first (for example ”S”) input of the trigger. The first impulse fixes a certain

state - in the case logical “1” at its output and the succeeding ones do not change the output state regardless their number. When the second light ray crosses the mouth of a bottle, the obtained output impulses from the second photo receiver PR_2 enter the second (“R“) input of the trigger. The first one of them alters the output state of the trigger into logical “0” and the succeeding ones are not of importance. Thus the obtaining of only one output impulse when a bottle passes is guaranteed.

The chosen scheme solution is characterized by extremely high reliability, high stability, simplicity and lack of necessity of adjustment at producing and in the process of exploitation.

The main problem in designing of the construction is the right choice of the distance l between both photoconverters. In order the impulses not to enter the both inputs of the $\overline{R}$ - $\overline{S}$ trigger simultaneously this distance has to be as big as possible. But its excessive augmentation would lead to errors from missing of bottles if they do not move closely one to another. From Fig.2 it can be seen that if the ray diameter is small enough following condition has to be fulfilled:

$$d < l < D \qquad (1)$$

where d - maximal diameter of the mouth of the bottle; D – minimal diameter of the body of the bottle.

On the basis of the described principle the entire block scheme of the device for counting of glass bottles on the conveyor belt (Fig.3) is developed.

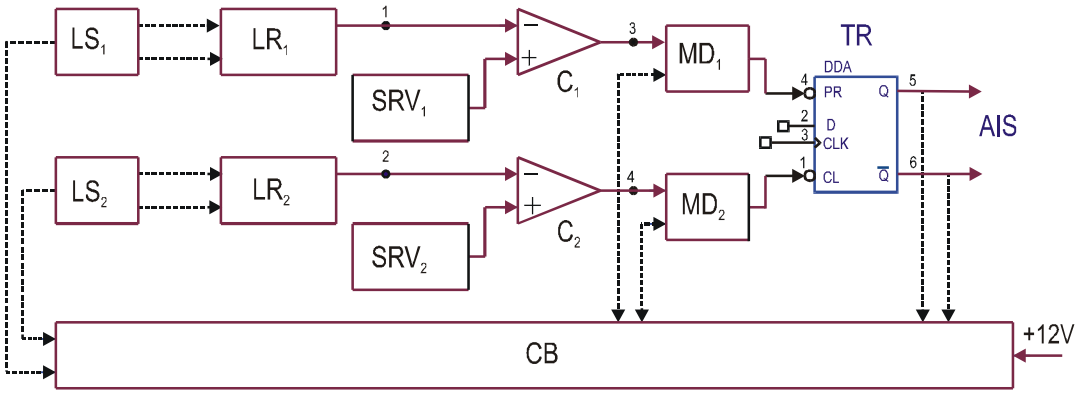


Fig.3. Structural scheme of the device for counting of bottles

Two identical channels, each one including light source (LS_1 and LS_2), light receiver (LR_1 and LR_2), source of reference voltage (SRV_1 and SRV_2), comparator (C_1 and C_2) and matching device (MD_1 and MD_2) are used.

The principle of operation is illustrated through the time diagram from Fig. 4. When the light receiver LR is lighted up, the voltage of the inverting input of the comparator C is higher than the reference one (Ur). The corresponding output voltage of the comparator is low. At the output of the amplifier MD a high TTL - level is obtained as the amplifier is an inverting one. When the light receiver LR_1 is shaded by a passing bottle at the output of the comparator C_1 a high level is obtained and at the output of MD_1 – low level. The $\overline{R}$ - $\overline{S}$ trigger TR is established in condition “logical 1”. When the light receiver LR_2 is shaded analogous processes occur and the trigger TR is cleared. The trigger TR eliminates the influence of the winkings.

The device has a symmetrical output. This enables sharp decreasing of the disturbances, which may penetrate through the line, connecting the output of the device to the input of the Automated information system (AIS) as well as for possible errors, caused by the disconnecting of the connecting wires at their connecting to “ground” etc. For the purpose in the receiving block of AIS a circuitry “sum of modulus two” is connected.

A control block (CB) for diagnostics and control of the normal operation [Marinov, 1980] is provided as a part of the device and through which the good working order of the LS_1 and LS_2; the output signals of the comparators,

received from MD_1 and MD_2; the signals, received from the outputs of the trigger; the presence of supply voltage are supervised.

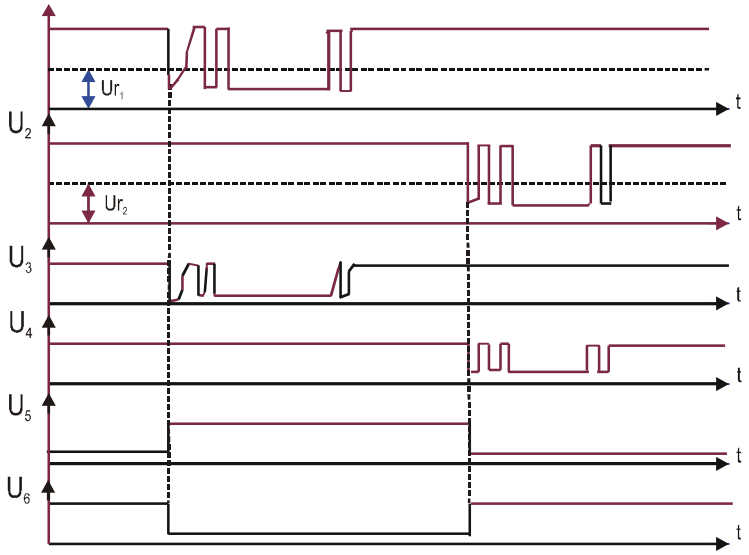


Fig.4. Operation time diagram

Conclusion

The designed device is a composite part of the automated information system for control, reporting and documenting the quantity of produced glass bottles, which is introduced in the factory for glass processing in town of Elena. The device enables the counting of empty bottles, discolored or of different coloring, of different form and size. It also may be successfully applied for counting of full bottles regardless the content and it level. These qualities of device provide its comparatively wide application in different branches of industry.

Bibliography

[Draganov, 2006] Draganov, V.D. , G.P. Toshkov, D.T. hristov. System for reading and documenting of the ready production quantity, Acta Universitatis Pontica, Volume 6, Number VII, 2006

[Solid Count, 2006] www.cornerstoneautosys.com/solidcount.htm - Solid Count

[Fast Counts, 2006] http://www.accusort.com/applications/counting.html - Fast Counts

[Patent 0050111724, 2005] United States Patent Application: 0050111724, A1, May 26, 2005. Method and apparatus for programmable zoned array coundter

[Bergmann, 1980] Bergmann H. Fotoelektrische Schalter, Radio fernsehen elektronik, 12/1980, pp. 769-770, in German

[Marinov, 1980] Marinov, Ju., E. Rangelova, V. Dimitrov. Technical diagnostics of radio-electronic systems and devices, Tehnika, Sofia, 1980, in Bulgarian

Authors' Information

Ventseslav Draganov – Technical University of Varna, 1, Studentska Str, Varna, 9010, e-mail: v2draganov@abv.bg

Georgi Toshkov – Technical University of Varna, 1, Studentska Str, Varna, 9010, e-mail: g_toshkov2006@abv.bg

Dimcho Draganov – "Technotrade**",** Varna, 9000.

Daniela Toshkova – Technical University of Varna, 1, Studentska Str, Varna, 9010, e-mail: daniela_toshkova@abv.bg

Information Systems

BUILDING DATA WAREHOUSES USING NUMBERED INFORMATION SPACES

Krassimir Markov

Abstract: *An approach for organizing the information in the data warehouses is presented in the paper. The possibilities of the numbered information spaces for building data warehouses are discussed. An application is outlined in the paper.*

Keywords: *Data Warehouses, Operational Data Stores, Numbered Information Spaces*

ACM Classification Keywords: *E.1 Data structures, E.2 Data storage representations*

Introduction

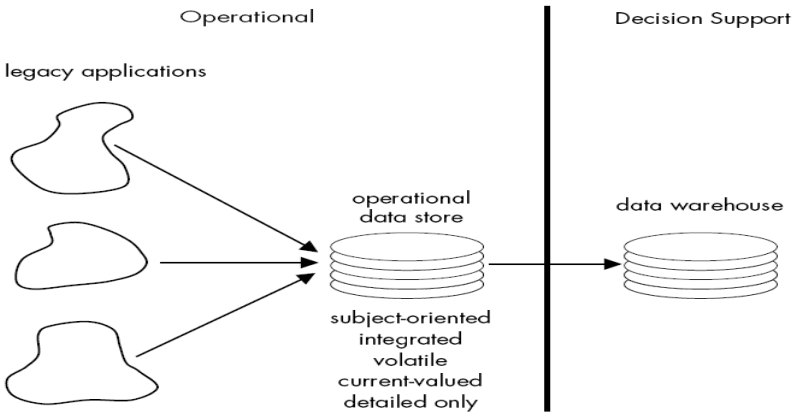
The origin of the Data Warehouses (DW) can be traced to studies at MIT in the 1970s which were targeted at developing an optimal technical architecture [Haisten, 2003]. The initial conception of DW had been proposed by the specialists of IBM using the concept “information warehouses” and its goal was to ensure the access to data stored in no relational systems. In 1988, Barry Devlin and Paul Murphy of IBM Ireland tackled the problem of enterprise integration head-on. They used the term "business data warehouse" and defined it as: “a repository of all required business information” or “the single logical storehouse of all the information used to report on the business” [Devlin and Murphy, 1988]. At present, the conception of “data warehouse” becomes popular mainly due to activity of Bill Inmon. In 1991, he published his first book on data warehousing. W.H. Inmon’s definition is: “Data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision making process“ [Inmon, 1991]. Let remember, the data warehouses allow long term information about an enterprise to be recorded, summarized and presented. Usually the data warehouse is a passive observer object that takes no part in business processes, and is not part of the business model. The axes of a multidimensional data warehouse are not arbitrary, but represent real aspects of the business. Axes should represent the purpose, process, resource and organization aspects. The summary hierarchies on each of these axes should parallel the fractal structures in the business model. Roll up and drill down to zoom from summary to detail information is therefore based on the structure of the business, so is meaningful to management and other users. [Marshall, 1997].

As a rule, the typical enterprise has many different systems for operative processing with very incompatible data. In such case, the main task is to convert the existing archives of data into a source for new knowledge which will give to the users a uniform integrated and consolidated notion of the corporate data. The old systems for operative information processing have been developed without foreseeing the support of the requirements of modern business and the need of automated support of decision making. Because of this, the converting the usual systems for online transaction processing (OLTP) in the systems for decision support (resp. – DW) were very complicated task. To solve this problem, an intermediate level has been proposed – the “operational data stores”. The **Operational Data Store** (ODS) is a database designed to integrate data from multiple sources to facilitate operations, analysis and reporting. Because the data originates from multiple sources, the integration

often involves cleaning, redundancy resolution and business rule enforcement. An ODS is usually designed to contain low level or atomic (indivisible) data such as transactions and prices as opposed to aggregated or summarized data such as net contributions. Aggregated data is usually stored in the DW [Wikipedia, ODS].

The definition of ODS given by Bill Inmon is: "an ODS is a subject-oriented, integrated, volatile, current-valued, detailed-only collection of data in support of an organization's need for up-to-thesecond, operational, integrated, collective information". [Inmon, 1995]

At first glance the ODS appears to be very similar to the data warehouse in structure and content. In some respects there are strong similarities between the two types of architectural constructs. But the ODS has some very different characteristics from the data warehouse. Both the ODS and the data warehouse are subject-oriented and integrated. In that regard, the two environments are identical. Both environments require that data be integrated and transformed as it passes into the ODS and/or the data warehouse. But here the similarities between the ODS and the data warehouse end. The ODS contains volatile data while the data warehouse contains non-volatile data. Data is updated in the ODS while data is not updated in the data warehouse. Another important difference between the two environments is that the ODS contains only very current data while the data warehouse contains both current data and historical data. The data in the data warehouse is not nearly as fresh as the data in the ODS. The data warehouse contains data that is no more current than the last 24 hours. The ODS contains data that may be only seconds old. Another major difference between the two architectural constructs is that the ODS contains detailed data only. The data warehouse contains both detailed and summary data. There are then some major differences between the types of data found in the two environments. One of the most important features of the ODS is the system of record. The system of record is the formal identification of the data in the legacy environment that feeds the ODS. (Pic.1) [Inmon, 1995]



Pic.1. The Operational Data Store [Inmon, 1995]

So, an operational data store (ODS) is a type of database often used as an interim area for a data warehouse. Unlike a data warehouse, which contains static data, the contents of the ODS are updated through the course of business operations. An ODS is designed to quickly perform relatively simple queries on small amounts of data (such as finding the status of a customer order), rather than the complex queries on large amounts of data are typical of the data warehouse. An ODS is similar to your short term memory in that it stores only very recent information; in comparison, the data warehouse is more like long term memory in that it stores relatively permanent information.

In the early 1990s, the original ODS systems were developed as a reporting tool for administrative purposes. They were usually updated daily and provided reports about business transactions for that day, such as sales totals or orders filled. This type of system is now referred to as a **Class III ODS**. With changes in technology and business needs, the **Class II ODS** evolved to track more complex information such as product and location codes, and to update the database more frequently (perhaps hourly) to reflect changes. **Class I ODS** systems

arose from the development of customer relationship management (CRM). In Class I systems, synchronous or near-synchronous updates are used to provide customers with consistently valid and organized information. Another version, the **Class IV ODS**, was recently developed with an added capacity for more interaction between the data warehouse or data mart and the ODS. [Oracle ODS]

The milestone for the work presented in this paper is the simple idea that we may use a special kind of organization of the information and this way to develop easy to use and compact ODS of Class I with facilities of DW with very high speed for response which enables ***the real-time analytical processing*** (RTAP). (The RTAP multithreaded processing engine needs to support extremely large volumes of data in real time. The analytics performed are composed of combinations of algorithmic, statistical and logical functions. [B-Jensen 2002])

The investigation presented in this paper is based on the fact that a specialized form of data warehouse is the corporate financial ledger. The segments of an account code serve the same purpose as the values on the axes of a data warehouse [Marshall, 1997]. In the same time, there exist a lot of account codes in a financial ledger and it is needed to operate with great complex of tables, descriptions, reports, etc. This leads to very complicated realizations which in the most cases are paid by more and more external memory for hundreds files as well as by growing quantity of processing operations.

In other hand, well-known considerable information complexes are offered by “SAP” (Germany), “Oracle”, “PeopleSoft” (USA), etc., but the prices of such software are very high. This is serious problem for the middle and small enterprises, especially in Bulgaria, which will bankrupt if decide to implement so rich automated systems. Because of this the narrow versions of such software are offered at the market. Unfortunately those versions are not as convenient as they are advertised and provoke many additional problems during the implementation process and exploitation.

Our approach is to build information complexes for information service of business accounting and decision making based on numbered information spaces [Markov, 2004a], which may support RTAP on the level of ODS Class I and this way to reduce the expenses for maintenance separate DW. This goal may be achieved using the FOI Archive Manager (ArM) ®.

FOI Archive Manager (ArM) ®

The FOI Archive Manager (ArM) ® is a tool for building numbered information spaces. ArM is based on the “Multi-Domain Information Model” (MDIM). It has been established more than twenty years ago. For a long period it has been used as a basis for organization of the information bases. The first publication which contains some details from MDIM is [Markov, 1984] but as a whole the model was presented in [Markov, 2004a]. There exist several realizations of FOI Archive Manager (ArM) ® for different hardware and/or software platforms. The newest ArM Version 9 for IBM PC developed using DELPHI for MS Windows XP is called **ArM32.**

Let remember the main possibilities of **ArM32** [Markov, 2004b] using some definitions of MDIM.

Basic information element of MDIM is an arbitrary long string of machine codes (bytes). When it is necessary the string may be parceled out by lines. The length of the lines may be variable. In ArM32 the length of the string may vary from 0 (zero) up to 2^{30} (1G) bytes. There is no limit for the number of strings in an archive but theirs total length plus internal indexes could not exceed 4G bytes in a single file.

Let $\boldsymbol{E}_1$ is a set of basic information elements: $\boldsymbol{E}_1 = \{e_i \mid e_i \in \boldsymbol{E}_1,\ i=1,\ldots, m_1\}$.

Let μ_1 is a function which defines a biunique correspondence between elements of the set $\boldsymbol{E}_1$ and elements of the set $\boldsymbol{C}_1$ of positive integer numbers: $\boldsymbol{C}_1 = \{c_i \mid c_i \in \boldsymbol{N},\ i:=1,\ldots, m_1\}$, i.e. $\mu_1 : \boldsymbol{E}_1 \leftrightarrow \boldsymbol{C}_1$. The elements of $\boldsymbol{C}_1$ are said to be number codes of the elements of $\boldsymbol{E}_1$. The triple $\mathcal{S}_1 = (\boldsymbol{E}_1, \mu_1, \boldsymbol{C}_1)$ is said to be a ***numbered information space of range 1***.

The triple $\mathcal{S}_2$ = ($\mathcal{E}_2$, μ_2, C_2) is said to be a ***numbered information space of range 2*** iff $\mathcal{E}_2$ is a set which elements are numbered information spaces of range 1 and μ_2 is a function which defines a biunique correspondence between elements of $\mathcal{E}_2$ and elements of the set C_2 of positive integer numbers: $C_2 = \{c_j \mid c_j \in N,\ j:=1,\ldots,m_2\}$, i.e. $\mu_2 : \mathcal{E}_2 \leftrightarrow C_2$.

The triple $\mathcal{S}_n$ = ($\mathcal{E}_n$, μ_n, C_n) is said to be a ***numbered information space of range n*** iff $\mathcal{E}_n$ is a set which elements are information spaces of range ***n-1*** and μ_n is a function which defines a biunique correspondence between elements of $\mathcal{E}_n$ and elements of the set C_n of positive integer numbers: $C_n = \{c_k \mid c_j \in N,\ k:=1,\ldots,m_n\}$, i.e. $\mu_n : \mathcal{E}_n \leftrightarrow C_n$.

The sequence $A = (c_n, c_{n-1}, \ldots, c_1)$ where $c_i \in C_i$, i=1,…,n is called ***multidimensional space address*** of range **n** of a basic information element. Every space address of range **m**, **m<n**, may be extended to space address of range **n** by adding leading **n-m** zero codes. Every sequence of space addresses $A_1, A_2, \ldots, A_k$, where k is arbitrary positive number, is said to be a ***space index***.

Every index may be considered as basic information element, i.e. as a string, and may be stored in a point of any information space. In such case it will have a multidimensional space address which may be pointed in the other indexes and, this way, we may build a hierarchy of indexes. So, every index which points only to indexes is called ***metaindex***.

Let $G = \{S_i \mid i=1,...,m\}$ is a set of numbered information spaces.

Let $\tau=\{\nu_{ij} : S_i \rightarrow S_j \mid i=const, j=1,\ldots m\}$ is a set of mappings of one "main" numbered information space $S_i \subset G$, i=*const*, into the others $S_j \subset G$, j=1,…m, and, in particular, into itself. The couple: $Đ = (G, \tau)$ is said to be an "***aggregate***".

The **ArM32** elements are organized in numbered information spaces with variable ranges. There is no limit for the ranges the spaces. Every element may be accessed by correspond multidimensional space address (coordinates) given via coordinate array of type cardinal. At the first place of this array the space range needs to be given. So, we have two main constructs of the physical organizations of ArM32 – numbered information spaces and elements.

The main **ArM32** operations with basic information elements are: **ArmRead** (reading a part or a whole element); **ArmWrite** (writing a part or a whole element); **ArmAppend** (appending a string to an element); **ArmInsert** (inserting a string into an element); **ArmCut** (removing a part of an element); **ArmReplace** (replacing a part of an element); **ArmDelete** (deleting an element); **ArmLength** (returns the length of the element in bytes).

The **ArM32** numbered information spaces are ordered and main operations within spaces take in account this order. So, from given space point (element or subspace) we may search the previous or next empty or non empty point (element or subspace). In is convenient to have operation for deleting the space as well as for count its nonempty elements or subspaces.

The **ArM32** logical operations defined in the multi-domain information model are based on the classical logical operations - intersection, union and supplement, but these operations are not so trivial. Because of complexity of the structure of the spaces these operations have at least two principally different realizations based on codes of information spaces' elements and on contents of those elements.

The **ArM32** information operations can be grouped into four sets corresponding to the main information structures: elements, spaces, aggregates, and indexes. Information operations are context depended and need special realizations for concrete purposes. Such well known operations are, for instance, transferring from one structure to another, information search, sorting, making reports, etc.

At the end there exist several operations which serve information exchange between **ArM32** archives (files) such as copying and moving spaces from one to another archive.

ArM32 engine supports multithreaded concurrent access to the information base in real time.

Very important feature of ArM32 is possibility not to occupy disk space for empty structures (elements or spaces). Really, only non empty structures need to be saved on external memory.

Complex FOI®

Complex FOI® is an integrated software environment for economical information processing and business analysis. The main features of **Complex FOI** [Markov et al, 1994] are built on three levels, which correspond to the Pyramidal Information Model (PIM) presented in [Markov et al, 1993]. The levels of this model are “Strategy”, “Analysis”, and “Service”. Every level contains three parts, which correspond to “Human Resources”, “Materials”, and “Finances” of the enterprise. It easy to see that there exist correspondence between PIM and ODS and DW.

The main set of concrete systems for information processing is included on "Service" level. They are aimed to service the operative work and control. For instance, there exist systems for service the enterprise financial tasks such as computing of salaries [Markov et al, 1996a], systems for managing different material stores using appropriate information access - by names or by numbers of goods [Markov et al, 1995a], systems for maintenance of fixed assets [Markov et al, 1996b], etc. An example of another class of service systems is one for automated payment of consumption of water and other communal services in a town as well as the specialized service systems, such as one for computing the price of building of some architectural object. It is clear, ***the legacy applications*** of the enterprise are assumed to be on this level too.

All these systems are integrated with the upper level (“Analysis”) via very convenient interface – the natural language standard accounting records which are the usual transaction form for accounting process. Furthermore, the information in **Complex FOI** is distributed in correspond numbered information spaces in accordance to usual every day financial accounting information structures. This make integration possible and automated information exchange is simple and comprehensible.

There is only one system on level "Analysis". It is an ODS with possibilities for accounting as well as for account analysis [Markov et al, 1995b]. This is the main tool for enterprise financial control and managing which support automated day-to-day operations (purchasing, banking etc), transactions access and modifying a few records at a time, application oriented database design, and metric: transactions/sec. The main structure of this level is the financial ledger - usually it is a numbered information space of range up to 10. Its subspaces represent accounting divisions, groups and accounts, as well as sub-accounts on several sub-levels. Every space may contain operational and historical data in the same time.

The main feature of the level "Strategy" is the decision support. All information from low levels can be used for supporting the processes of business decisions in the group of leaders of the enterprise. The functionality of this level covers the usual understanding of data warehouse but it is realized as distributed RTAP engine which support complex queries that access records with operational and/or historical data for trend analysis.

Because of special multidimensional organization, in Complex FOI the analytical pre-computation can be provided in real time during the operative work and its results (elements, spaces, aggregates, and indexes) can be stored in corresponded structures of the multidimensional hierarchical information base. So, in query response time, it is easy to process ***multidimensional modeling*** (for instance - compute total *sales* volume per *product* and *store*); ***operating with dimensions and hierarchies*** (for instance - roll-up: move up the hierarchy e.g. given total salaries per department, we can roll-up to get salaries per enterprise; drill-down: move down the hierarchy more fine-grained aggregation; pivoting: aggregate on selected dimensions usually 2 dims (cross-tabulation)); ***comparisons*** (for instance - this period vs. last period - show me the sales per store for this year and compare it to that of the previous year to identify discrepancies); ***ranking and statistical profiles*** (for instance – top N / bottom N - show me sales, profit and average call volume per day for my 10 most profitable salespeople); ***custom consolidation*** (for instance - market segments, ad hoc groups - show me an abbreviated income statement by quarter for the last four quarters for my northeast region operations); etc.

Conclusion

The approach to build information complexes for information service of business accounting and decision making based on numbered information spaces which may support RTAP on the level of ODS Class I and this way to reduce the expenses for maintenance separate DW has been presented in the paper. This goal may be achieved using the FOI Archive Manager (ArM) ® and "Multi-Domain Information Model" (MDIM). An application of presented approach named "Complex FOI" was outlined.

Acknowledgments

Author is indebted to Ilia Mitov and Krassimira Ivanova for support and collaboration. Due to theirs hard work the approach presented in this paper has been widely implemented in practice.

This work is a part of the project "**ITHEA XXI**", partially financed by the Consortium FOI Bulgaria.

Bibliography

[B-Jensen 2002] M.T. B-Jensen. High Tower Software's Tower View is the Odds-On Favorite of International Game Technology for Real-Time Data Management. Product Review published in DM Review Magazine July 2002 Issue. http://www.dmreview.com/article_sub.cfm?articleId=5403

[Devlin and Murphy, 1988] B.A. Devlin and P.T. Murphy. An Architecture for a Business and Information System. IBM Systems Journal. Volume 27, No. 1, 1988. http://www.research.ibm.com/journal/sj/271/ibmsj2701G.pdf

[Haisten, 2003] M. Haisten. The Real-Time Data Warehouse: The Next Stage in Data Warehouse Evolution http://www.damanconsulting.com/company/articles/dwrealtime.htm

[Inmon, 1991] W.H. Inmon. Building the Data Warehouse, QED/Wiley, 1991.

[Inmon, 1995] W.H. Inmon. The Operational Data Store. InfoDB February 1995 http://www.evaltech.com/wpapers/ODS2.pdf

[Markov 1984] K. Markov. A Multi-domain Access Method. // Proceedings of the International Conference on Computer Based Scientific Research. Plovdiv, 1984. pp. 558-563.

[Markov et al, 1993] K. Markov, K. Ivanova, I. Mitov, J. Ikonomov. Pyramidal Model of the Firm Information Activities. IJ ITA, 1993, Vol. 1, No. 2. (in Russian)

[Markov et al, 1994] K. Markov, K. Ivanova, I. Mitov. Basic concepts and main information structures of the Complex FOI. FOI-COMMERCE, Sofia, 1994. (in Bulgarian)

[Markov et al, 1995a] K. Markov, K. Ivanova, I. Mitov. Automated service of the storehouses. FOI-COMMERCE, Sofia, 1995. (in Bulgarian)

[Markov et al, 1995b] K. Markov, K. Ivanova, I. Mitov. Automated service of the financial accounting using System "ANALYSE". FOI-COMMERCE, Sofia, 1995. (in Bulgarian)

[Markov et al, 1996a] K. Markov, K. Ivanova, I. Mitov. Automated service of the accounting the staff and salaries FOI-COMMERCE, Sofia, 1996. (in Bulgarian)

[Markov et al, 1996b] K. Markov, K. Ivanova, I. Mitov. Automated service of the accounting of the fixed assets. FOI-COMMERCE, Sofia, 1996. (in Bulgarian)

[Markov, 2004a] K. Markov. *Multi-Domain Information Model*. Proceedings of the ITC&P-2004 - International Conference "Information Technologies and Communications & Programming", Varna. FOI-COMMERCE, 2004, pp. 79-88. Int. Journal "Information Theories and Applications", 2004, Vol. 11, No. 4, pp. 303-308

[Markov, 2004b] K. Markov. *Coordinate Based Physical Organization of Computer Representation of Information Spaces*. Proceedings of the Second International Conference "Information Research, Applications and Education" i.TECH 2004, Varna, Bulgaria. Sofia, FOI-COMMERCE – 2004, стр.163-172 (in Bulgarian).

[Marshall, 1997] Cr. Marshall. *Business Object Management Architecture*. OOPSLA'96 Workshop Business Object Design and Implementation II: Business Objects as Distributed Application Components - the enterprise solution? http://jeffsutherland.com/oopsla97/marshall.html

[Oracle ODS] - whatis.com - http://searchoracle.techtarget.com/sDefinition/0,,sid41_gci786730,00.htm

[Wikipedia, ODS] http://en.wikipedia.org/wiki/Operational_data_store

Authors' Information

Krassimir Markov – Institute of Mathematics and Informatics, BAS, Information Research Department; e-mail: foi@nlcv.net

INTERNATIONALIZATION AND LOCALIZATION AFTER SYSTEM DEVELOPMENT: A PRACTICAL CASE

Jesus Cardenosa, Carolina Gallardo, Alvaro Martin

***Abstract**: Internationalization of software as a previous step for localization is usually taken into account during early phases of the life-cycle of software development. However, the need to adapt software applications into different languages and cultural settings can appear once the application is finished and even in the market. In these cases, software localization implies a high cost of time and resources. This paper shows a real case of a existent software application, designed and developed without taking into account future necessities of localization, whose architecture and source code were modified to include the possibility of straightforward adaptation into new languages. The use of standard languages and advanced programming languages has permitted the authors to adapt the software in a simple and straightforward mode.*

***Keywords**: Localization, Internationalization, XML.*

***ACM Classification Keywords**: D. Software, D.2.7 Distribution, Maintenance and Enhancement*

Introduction

Any technical device devoid of human interaction operates and yields an expected level of productivity regardless of the cultural environment where it is located. The same can be said for software, as long as it does not call for any human interaction. However, many software applications require human interaction for a correct functioning. In this case, the level of productivity of the software will depend not only on software's intrinsic technical characteristics but on external human factors.

When a software application is used in a context with a different cultural environment (like different mother language, different icons, symbols, etc.) from its original one, a process of adaptation into the new work culture is required. This process is known as **localization**. The adaptation into a new culture not only comprises evident factors like the language of the interface and messages to the user, measure units or data formats (also known as overt factors according to [Mahemoff et al, 1998]); but also other slippery and fuzzy issues that finally distinguish a culture, like mental disposition, perception of the world, rules of social interaction, religion, etc., which are referred to as the covert factors of a culture. More specifically, the process of localization consists on the "**adaptation** of a product, application or document content to meet the language, cultural and other requirements of a specific target market (a *locale*)", as expressed by the W3C [W3C, 2005].

On the other hand, **internationalization** refers to the design and development of a product, application or document content that **enables** easy localization for target audiences that vary in work culture, region, or language. In this sense, it can be said that internationalization precedes and facilitates the task of localization.

Besides, the processes involved in localization of software applications changes significantly depending on whether it is done over a pre-existent application or over a developed application.

The next section sketches the most frequent practices of software internationalization and localization in software design. However, in pre-existent applications, and depending on the system development methodology, the localization process can become very expensive in terms of time and resources. We will show how we internationalized and subsequently localized a pre-existent application in a cheap and quick manner, by means of advanced standard implementations languages like Visual .Net and XML.

Software Architectures for Internationalization and Localization

As we commented in the previous section, the internationalization and localization (I&L) processes deal with more that mere language issues. However and for the purposes of this paper, we will consider only the language adaptation, which is the most prominent and visible aspect of I&L.

Apparently, an **internationalized** product does not entail structural changes in order to adopt a new language. Internationalization consists on abstracting the functionality of a product of any given language, in a way that the support of the information of the new language can be added afterwards, without facing the source code (dependent of a given language) when the product is localized into a new language. Currently, main development platforms offer support and tools to facilitate the internationalization of over factors of applications [Hogan, 2004], [Huang et al, 2001], in a way that currently problematic questions are centered on the optimization of the internationalization processes within the life cycle of the application.

There are three main approaches for internationalizing an application. The first one is the system where messages, menus and other culture-sensitive factors are embedded in the source code of the application. This approach obliges to develop a different version of the system for each of the target cultural environments. Each version requires independent process of testing, maintenance and upgrading, multiplying the costs of localization.

The second approach consists on extracting messages to the user of a given application into an external library. The application is generated from a common source code that links to the culture-sensitive libraries. Although this architecture resorts on a unique source code, only the languages contained in the external library could be incorporated, and it is required to test and maintained each of the supported languages individually.

The third and last approach consists on an architecture composed of the core of the application comprising all the functionalities but independent of cultural factors, which dynamically access to files of external resources that contain information about the corresponding culture (localization packages). The difference with the previous approach lies in the fact that the culture-independent code dynamically calls to the information of culture, so that only one executable must be tested and maintained. Once the set of supported cultures is tested, the addition of new cultures does not imply modifications. From this general idea, each author develops his/her own way of acting. For example, [Stearns, 2002] describes the process of developing systems sensitive to cultures using JAVA and XML for resources files, whereas the environment GNU/Linux [Tykhomyrov, 2002] and the Free Software Foundation [FSF, 2002] prefer the use of special libraries that facilitates the extraction of the localizable contents of the application and the construction of localization packages.

Regarding the aspects related to the life cycle of the internationalized software, [Mahemoff et al, 1999] presents a methodology for requirements specification to develop culture-sensitive systems. On the other hand, [Huang, 2001] offers a description of the processes to be followed to create culture-sensitive software, emphasizing the fact that the internationalization tasks should be included in the corresponding phases of the life cycle of software.

The work on the area of localization is complemented with research on the problem of localizing software already internationalized. Even when the technical procedure for software internationalization is optimized, the bottleneck lies in the **localization** processes of a product. The process of internationalized software localization resorts on the concept of repository and reuse of translation resources. That is, apart from the external file that contains the messages to the user and its translations, there is a repository where translations are stored for their subsequent reuse. In some cases, there are also repositories for terminology.

The following standards have been established to facilitate the task of managing the culture-sensitive resources files and their communication with repositories:

- XLIFF (XML Localization Interchange File Format) defines a standard format for resources files that stores the translated strings, in a way that tools for assisted machine translation can be developed

independent of the application to be localized, as well as transporting the translation information from one phase of the process to the following phase [OASIS, 2003].

- TMX (Translation Memory Exchange), allows for the storage and interchange of translation memories obtained after the use of automatic tools for translation [LISA, 2005].
- TBX (Term Base Exchange), defines a standardized model for terminological databases [LISA, 2003].

There are also some common practices among companies that have become a "*de facto*" standard [Hogan, 2004] aiming at minimizing the impact of localization on commercial software products, namely:

- Extraction of the fragment of texts used in the user interfaces into resources files.
- Control of the extracted texts, contexts and their translations.
- Outsourcing of the translation tasks to specialized companies.
- Simplification of the contents of the chains and their contents as a previous step to the sending to translation centres.

However, as can be seen, internationalization architectures and localization standards do not offer a solution for already existent applications that require international dissemination. That is, according to these architectures, localization is a bottleneck and it is only possible with an internationalized architecture. But what happens if we want to adapt a software application into many languages? The next section presents how an architecture can be changed in an afterwards-mode and how we internationalized and subsequently localized a pre-existent application in a cheap and quick manner, by means of advanced standard implementations languages like Visual.Net and XML.

Internationalizing an Existent Application: the Context

The starting point of this work is a software application for multilingual generation that allows for human interaction. It is an interactive application composed of a user interface where the user can manipulate semantic representations of the text to be translated.

The only requirement in the development of this tool was the use of UNICODE files, because of the almost certainty that the tool was going to be used for analysis and generation in a variety of languages. This obviously involved the future necessity of localization of the tool. It seems clear that the internationalization should be foreseen and reflected at the level of requirements specification and that it consists on something more than the mere use of UNICODE files. We will show how, in cases where internationalization has not been taken into account in the development processes, a pre-existent system can be adapted a posteriori for internationalization purposes. That is, our work is framed in the following context:

- There is a need of a future internationalization and localization processes, which is partly reflected on the requisites through the need to work with UNICODE files.
- The system is implemented in a development framework compatible with the use of UNICODE files (VB.NET), which guarantees the strict observation of the previous requisite, but nothing else.
- Apart from UNICODE files, there is not any other feature in the system oriented towards internationalization and subsequent localization.

The result of this situation is an environment able to import and deals with UNICODE files, that works with several languages (it is a translation aid tool) but with the totality of the user interfaces functionalities in just one language (in this case, Spanish).

The internationalization process that we are going to describe has been carried out **after** the complete development of the tool, proving that at least on of the most important and basic tasks of localization, such as language adaptation, can be done even without having internationalized the system in previous development phases.

Description and Preliminary Analysis of the Pre-existing System

The application is conceived as an environment for linguistic tasks, in which some external resources and components as language analyzers, language generators and dictionaries are integrated, with a powerful user interface and graphics management. From the architectural point of view, there are three main subsystems in the environment, which are:

- **Kernel**: it is the component in charge of managing most of the information and data flow of the application, as well as integration with external language analyzers, generators and dictionaries.
- **Graphic controller**: this component is in charge of managing the graphical display of abstract and semantic structures, as well as the correspondence between semantic structures and graphics.
- **Interface**: this component manages the communication of the application with the user. It mainly consists on the user interface with a few functionalities, which are delegated to the kernel or the graphic controller.

Figure 1 shows the application architecture and information flow graphically.

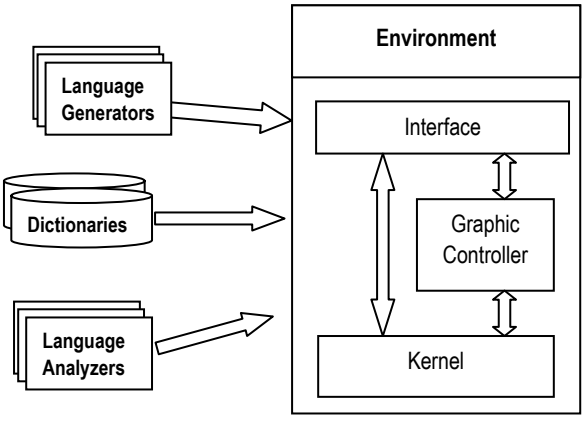


Figure 1. Architecture of the application

The entire interface is in Spanish. It is important to note that there are two types of **textual elements** in the application interface: "message errors" occurring in unexpected situations (also called emerging messages) and the text of the environment itself.

Each subsystem can generate a given number of emerging messages and windows with their corresponding text elements. In this way, the textual elements are scattered all over the source code.

A preliminary analysis of the source code shows that the textual elements follow two regular patterns. The first pattern corresponds to "emerging messages". These are created with the statement "*msgbox ()"* (an abbreviation for *message box*); the text assigned to the emerging message is written within the parenthesis. As an illustration, a real emerging message informing of a file that is not found will have the following code:

```
msgbox("Fichero no encontrado")
```

The second pattern corresponds to the text used in windows and buttons of the application, which have the general pattern:

```
component.text = "Text associated to this component"
```

Where the expression "component.text" is the convention in VB.NET to note that the string in double quotes is the text that will appear on that specific component. For example, to assign the text "Aceptar" (OK) to a given button, we write:

```
button.text ="Aceptar"
```

Since we are going to restrict the I&L process to just linguistic issues, these textual elements will be the subject of the I&L processes.

Strategy for I&L, Conceptual and Architectural Design

Our specific problem is the need to adapt the environment into the English language. The most obvious and even quick solution is to search for all the text elements in Spanish and create a new version of the application with the interface in English. However, there are some requirements on the I&L adaptation, such as:

a) The localization process should be done by translators / final users.
b) Maintainability of the system and translations should be guaranteed.
c) It is desirable to produce a core application abstracted from the linguistic issues.
d) The pre-existing components must not be functionally modified.

Therefore, the architecture should be modified with the addition of a new component in charge of the internationalization functionalities; so that the textual contents of new languages are stored as a new resource (in the form of an external file, for example) which can be read and processed by the application itself.

The new component is in charge of reading the external files with the translations of the textual elements and imports them into the environment so that messages and interfaces can be shown in different languages. The result is new software architecture as illustrated in figure 2.

Thus, the global strategy promotes the creation of a new specific component that once integrated in the original architecture is responsible for all the internationalization tasks. The basic functionalities of this new component should be:

1. Identification and labelling of all the text strings written in Spanish language of any kind (emerging messages, buttons, windows, and any other textual elements)
2. Extraction of these strings and generation of a XML file according to a predefined structure
3. Capture of the new XML file, once all the identified strings have been translated into the new language in the XML file.
4. Insertion of the translated strings according to the labelling.

The detailed description of this process is shown in the next section.

The Practical Case

The new component, called “Internationalization Manager”, serves a number of functions that guarantee that the required language changes are carried out over the existent environment, while intervening in the current software as less as possible. Figure 2 shows the new architecture of the environment and how the “Internationalization Manager”, together with its functional element the Text Management Module (TMM), is integrated in this new architecture. In the remaining, we will describe how the new component works and its main functionalities.

1. Text string identification and labelling.

This first functionality consists on identifying the textual elements in the original language (in our case, Spanish) following the two aforementioned search patterns, namely `msgbox` and the `component.text`. This function has been carried out by means of a script that identifies these text strings. The result of the script is a file where not only the text string is stored, but also additional information associated to the string, like its location, the component it belongs to, and other information that could be useful. All the information that the script gathers about a text string is labelled with a numeric identifier. The only modification that is done from this moment over the original software is the substitution of these strings by a function that calls for the identifier in the XML file of the required language and inserts the text string contained in the XML file. Let’s see an example of how it works.

Suppose the source code of the application in Spanish contains the following an emerging message:

`msgbox ("Error: Archivo no encontrado")` (English: "Error: *File not found*")

The Spanish text string is substituted by the following:

`msgbox (`**`InternationalizationManager.GetText(57)`**`)`

Where `InternationalizationManager` is the function that calls to the corresponding component of the TMM that executes the instruction `GetText(57)`. This instruction captures and temporally inserts the text string labelled with the identifier `57` in the language selected by the user in its place. Currently there are not text strings of a specific language in the environment anymore but functions like the aforementioned that allow for the incorporation of a new language in the environment without further changes over the original software.

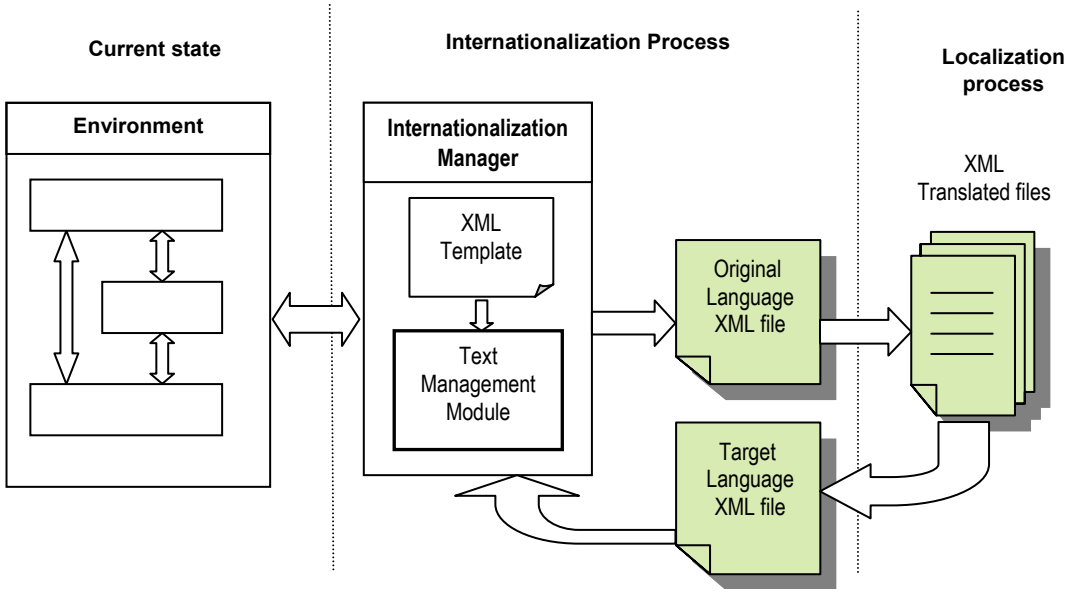


Figure 2. Global process

2. XML structure

The information about the text should be structured according to an XML template that permits to save a unique structure but modifiable in the data (in our case the text translations) that guarantees their interchangeability and maintainability. This XML file can be imported by the environment since the programming language (VB.net) is provided with an XML parser. This XML file is delivered to the translators and looks like as shown in figure 3.

The first line of the XML file indicates the version of the XML standard being used and the type of codification of the file (UNICODE in this case). The second line has an empty attribute `langID=""` that will indicate the target language of the translation of the strings. The rest of the XML file is divided in three elements `<userInterface>`, `<kernel>` and `<graphicController>`, each pertaining to the main components of the software. Each component is composed by a number of `<item>`. An `<item>` stores the following elements:

- The attribute `"id"` (in the example of figure 3, one "id" is 56) that uniquely identifies the linguistic text element and its presence in the software component.
- The `"orig"` attribute corresponds to the text string in the original language. One example is the Spanish string *"¿Desea continuar?"*.
- The `<translation>` element which is empty and will have to be filled with the translations into the target language.

```
<?xml version="1.0" encoding="UTF-8"?>
<localisation langID="">
 <userInterface>
   <item id="56" orig="¿Desea continuar?">
      <translation> </translation>
   </item>
      <item id="57" orig="Error: Archivo no encontrado">
      <translation> </translation>
   </item>
  ...
 </userInterface>
 <kernel>
   </item>
      <item id="64" orig="Atributo no válido">
      <translation> </translation>
   </item>
  ...
 </kernel>
 <graphicsController> … </graphicsController>
</localisation>
```

Figure 3. Original Language XML file

This file is distributed to translators so that they can perform the translations tasks in their corresponding working places, allowing for an absolute independence of the translation process and its integration in the software environment. The XML files in the target languages are delivered to the TMM and located in the corresponding directory so that they can serve as the different language options of the environment to be selected by the user. An example of an XML file containing the translations for English is shown in figure 4. This file is the result of the localization process.

```
<?xml version="1.0" encoding="UTF-8"?>
<localisation langID="English">
 <userInterface>
   <item id="56" orig="¿Desea continuar?">
      <translation>Do you want to continue?</translation>
   </item>
      <item id="57" orig="Error: Archivo no encontrado">
      <translation>Error: File not found </translation>
   </item>
  ...
 </userInterface>
 <kernel> ... </kernel>
 <graphicsController> … </graphicsController>
</localisation>
```

Figure 4. English Language XML file

Finally, the component “Internationalization Manager” is in charge of detecting XML files in the available languages and thus it offers them as options to the user of the environment. Once the user has select a language, the application dynamically imports the XML file that contains the text strings translated into the selected language and shows the environment in that language.

Conclusion

We have presented three approaches for software internationalization and subsequent localization. We have seen how the use of current programming languages which incorporate XML parsers allows the development of the third strategy, which the one that produces more flexible, adaptable and maintainable applications, in a convenient and easy and straightforward manner with a relatively low cost.

This approach also permits that the work of the developers can be initially kept apart from the linguistic questions and permits to maintain a single version of software. Major changes on the original software can be dealt with in the same way even if there appear new items.

Bibliography

[FSF, 2002] Free Software Foundation (2002), "Online GNU gettext manual" http://www.gnu.org/software/gettext/manual/gettext.html Accedido: Marzo 2006

[Hogan, 2004] Hogan M. J., Ho-Stuart C. & Pham B. (2004), "Key challenges in software internationalisation"

[Huang et al, 2001] Huang E., Hsu J. & Trainor H. (2001), "Unicode enabling for software internationalization" www.symbio-group.com/doc/Symbio%20Whitepaper%20on%20Unicode%20Enabling.pdf Accedido: Marzo 2006

[Huang, 2000] Huang E., Haft R., & Hsu J. (2000), "Developing a Roadmap for Software Internationalization" www.symbio-group.com/doc/Developing%20a%20Roadmap%20for%20Software%20Internationalization.pdf

[LISA, 2002] LISA (2002), "TBX Specification" http://www.lisa.org/stantdards/tbx Accedido: Marzo 2006

[Lisa, 2004] The Localization Industry Standards Association (2004), "Lisa Industry Primer 2nd Edition".

[LISA, 2005] LISA (2005), "TMX Specification" http://www.lisa.org/standards/tmx/tmx.html Accedido: Marzo 2006

[Mahemoff et al, 1998] Mahemoff M. J. & Johnston L. J. (1998), "Software Internationalisation: Implications for Requirements Engineering"

[Manemoff et al, 1999] Mahemoff M. J. & Johnston L. J. (1999), "The Planet pattern language for software internationalisation"

[OASIS, 2003] OASIS (2003), "XLIFF 1.1 Specification" http://www.oasis-open.org/comitees/xliff/documents/xliff-specifications.htm Accedido: Marzo 2006

[Stearns, 2002] Stearns B. et al (2002), "e-business Globalization Solution Design Guide: Getting Started" http://www.redbooks.ibm.com/abstracts/sg246851.html?Open Accedido: Marzo de 2006

[Tykhomyrov, 2002] Tykhomyrov O. Y. (2002) "Introduction to internationalization programming" The Linux Journal, Diciembre de 2002 http://www.linuxjournal.com/article/6176 Accedido: Marzo 2006

[W3C, 2005] W3C, "Localization vs Internationalization".http://www.w3c.org/International/questions/qa-i18n. 2005

[Yeo, 2001] Yeo A. W (2001), "Global-software development lifecycle: an exploratory study" Conference on Human Factors in Computing Systems, 2001

Authors' Information

Jesus Cardenosa – Department of Artificial Intelligence; Universidad Politécnica de Madrid; Madrid 28060, Spain; e-mail: carde@opera.dia.fi.upm.es

Carolina Gallardo – Department of Artificial Intelligence; Universidad Politécnica de Madrid; Madrid 28060, Spain; e-mail: carolina@opera.dia.fi.upm.es

Alvaro Martin – Department of Artificial Intelligence; Universidad Politécnica de Madrid; Madrid 28060, Spain; e-mail: martin@opera.dia.fi.upm.es

EXPERIENCES OF SPANISH PUBLIC ADMINISTRATION IN REQUIREMENTS MANAGEMENT AND ACQUISITION MANAGEMENT PROCESSES

Jose A. Calvo-Manzano, Gonzalo Cuevas, Ivan Garcia, Tomas San Feliu, Ariel Serrano, Magdalena Arcilla, Fernando Arboledas, Fernando Ruiz de Ojeda

***Abstract**: This paper shows the main contributions of the 1st Symposium on Improvement Process Models and Software Quality of Public Administrations. The obtained results expose the need to promote the implementation of Software Maturity Models and show possible advantages of its application in software processes of Public Administrations. Specifically, it was analyzed the current status in two process areas: Requirements Management and Subcontracting Management.*

***Keywords**: Requirements engineering, subcontracting management, improvement models, process.*

***ACM Classification Keywords**: D.2.9 Management, K.6.3 Software Management*

Introduction

In spite of the great advance in Information Technologies (IT) on the last years, we found that the majority of organizations, public or private, have the same problems in their software production process. These problems cause that:

- A software product is delivered, most of the time, with a 15% of defects.
- A fourth part of software projects are not finished or they are abandoned.
- A 30% or 45% of software resources are expensed in rewriting the software.
- Only the half of the times the plans and schedules established at the beginning are satisfied.

As an alternative to solve these problems, some of the research institutes of software engineering began the task to obtain and organize, in processes, the practices that are used to produce an maintain software and have demonstrated to be effective in some organizations.

For Software Engineering Institute (SEI) a process is a set of practices to perform and obtain a result, including tools, techniques, materials and people. This set of tools, techniques, materials and people is named “Software Process” [1]. The SEI has grouped the effective practices in references models.

A reference model is a set of processes that helps organizations to know their process status and is used as a guide to improve them. One of the most important improvement process models is the “Capability Maturity Model for Software (Sw-CMM)”, developed by the SEI [2]. Actually, this model is recognized for its integrated version (Capability Maturity Model Integration, CMMI) [3], [4]. The purpose of CMMI is to provide a “road map” for process improvement that works as a framework to improve in an effective way. This “road map” will be a useful guideline to improve all processes (develop, acquisition and maintain of products and services). In the same way, the CMMI offers a structured framework to evaluate the organization’s current process and establish priorities for the improvement task.

Motivation

The increasing use of the Internet and the continuous developing of new IT have changed the focus of Public Administrations to interest in improve the citizen attention process through continuous improving of their IT products. For this purpose, it is necessary that one of the system’s fundamental components, like software,

satisfy the user requirements, it is characterized by its reliability and can assure the accomplishment of costs and schedule associated with his development.

In this context, the Madrid Community, through the Autonomous Organism of Informatics and Communications (ICM) with collaboration of the Polytechnic University of Madrid, organized the 1st Symposium on Improvement Process Models and Software Quality of Public Administrations, its objectives were:

1. Determine the progress in Improvement Process Models, CMMI specifically.
2. Identify the advantages of CMMI application to the software process of Public Administrations.
3. Make a quickly evaluation of current situation of Public Administrations in processes like Requirements Management (RM) and Subcontracting Management (SAM) using CMMI as reference model.
4. Obtain general conclusions to develop process improvement initiatives in any Public Administration.

The Symposium was organized in:

- Conferences. The Conferences purpose was to show the current trends of the Improvement Process Models, particularly the models developed by the SEI. Besides, experiences of CMMI implementation in public organisms and private enterprises were presented [1], [5] [6] [7].
- Focus Groups. The principal objective of the Focus Groups was to obtain a quick assessment of processes in the Public Administrations using the CMMI as reference model.
  The RM and SAM processes were discussed in different focus groups, having special attention in do not accept, in the same group, two or more participants of the same Public Administration.
  Each focus group was integrated with a maximum of ten participants. The discussion was managed by one moderator with the objective of getting current data (issues) related to the Public Administrations.
  Another objective of focus group was the benefits identification through the analysis of Public Administrations on improving their current processes by a CMMI implementation. The obtained result was a list of actions and recommendations of short term.
- Workshops. Finally, the Workshops were organized by the Symposium sponsors to present, in this way, the commercial offer with solutions for the implementation of improvement process models.
  The Symposium had an audience of 133 participants from three Ministerial, eleven Autonomous Communities and two Local Administrations. Besides, eleven private enterprises participated in the Symposium and they sponsored the event too.

Requirements Management Group

The focus groups have one moderator. The moderator gave a brief introduction about concepts and practices of CMMI RM Process. After this, the moderator addressed the discussion with the objective that the Public Administrations expose their current situation.

To continue, the specific practices (SP) of CMMI RM Process are described. A brief description and the obtained findings are included.

- SP1.1. Develop an understanding with the requirements providers on the meaning of the requirements.
  Description. Establish the need of identifying the requirements providers to get the same understanding for each requirement and obtain the expected reliability.
  Findings. It was found that there are multiple providers for the Public Administrations. It is true that most of the time these providers could be identified by their name, but the communication process is informal, it means that do not exist any documentation to identify all requirements providers. Usually, the requirements are communicated in a horizontal way to the top level.

- SP1.2. Obtain commitment to the requirements from the project participants.
  Description. Determine the impact of agreed requirements with the project participants negotiating and registering the established agreements.
  Findings. Almost all of the participants in this group determined that they do not have defined processes for establish agreements with all requirements providers. The agreements are made orally through work meetings.
- SP1.3. Manage changes to the requirements as they evolve during the project.
  Description. Establish that the requirements and their associated changes must be managed from the beginning and during life cycle, in a manual way or using and automatic tool.
  Findings. A great number of Public Administration does not have a process for change management or, in a better case, they have a poor process but they do not document their changes. This is a weak point identified by the evaluation. In another hand, the participants found that it is too difficult manage the changes because sometimes the requirements providers are not aware of the impact generated by a change.
- SP1.4. Maintain bidirectional traceability among the requirements, project plans and the work products, from their source to a lower level.
  Description. This practice, called traceability, establishes the need of make a detailed and continuous tracking of each system requirement through the life cycle.
  Findings. On first place the participant had confusion with the "traceability" concept. This aspect exposed the lack of knowledge because it was the first time that they heard the word. In another hand, we found that many Public Administrations do not perform traceability practices and they do not use a traceability matrix for their requirements. Only some of them perform a poor change management process. We confirmed this practice as a weak point in the RM process in Public Administrations.
- SP1.5. Identify inconsistencies between the project plans and work products and the requirements.
  Description. This specific practice finds the inconsistencies between the requirements and the project plans and work products and initiates the corrective action to fix them.
  Findings. It was determined that the Public Administrations do not perform the revision of their projects plans, activities or work products for consistency with the requirements. For this reason, this practice was identified as a weak point too.

By the end, the "process institutionalization" to ensure that the process will be documented, effective, repeatable and lasting, is a long term objective. The Public Administration confirmed the use of one or two previous practices, but an institutionalized practice is a concept out of their hands in this moment.

Proposals of Short-term Actions on Requirements Management Process in Public Administrations

All the participants of focus groups identified the need of having an effective, repeatable and lasting RM process to obtain reliable and controllable requirements. They identified the following short-term actions:

- Involve all organization in the improvement project, mainly the top level.
- Promote training initiatives among the personal of RM process.
- Sensitize Senior Management of Public Administrations with the importance of having an effective, repeatable and lasting RM process, and the benefits that this process brings to the software development process.

- Make that users understand the cost that any change implies and the importance of a proper requirements definition process at the beginning of a development.
- Monitor the requirements through traceability techniques and promote the tools acquisition for easy implementation.
- Develop a list of the most common terms used in Requirements Engineering to avoid confusions.
- Develop a guideline of RM practices to obtain a successful process in future projects.

Conclusions about RM process in Public Administrations

The Focus Groups promoted the participation of Public Administration and they expressed their current issues. Also, with the ideas and concepts expressed in the Conferences, each participant made the assessment of their own RM process using the CMMI as reference model. The Public Administrations identified the gap that they have with respect to the model. Although it is true that each one of the Public Administrations has a RM process, the opportunity to compare it with a reference model helped them to identify their deficiencies and what they have to do to improve their RM process. In addition, all Public Administrations agreed the importance of having procedures that allow them to repeat the successes for every new project.

Subcontracting Management Group

The Acquisition Process is defined as the process of acquiring partially or totally the Information System (IS) Technologies from an external services supplier [8]. It means to delegate everything or part of the IT work through a contract with an external company that joins in the client organizational strategy and seeks to design a solution to existing informatics problems inside the latter. In the last years, the SAM process of IT functions has been gained the attention of many researchers and industries.

To continue, the two specific goals (SG) of CMMI SAM Process are described. A brief description and the obtained findings are included.

- SG1. Establish supplier agreements.

  Description. Determine the product to acquire and identify and select the suppliers that better adapt to the organization needs. The agreements must be established by a formal contract.

  Findings. The Public Administrations usually call for proposals to subcontract their projects. However, this context is made with a certain level of mistrust giving the budgets exceed, the lack of communication, and mainly, the loss of project control. One of the main issues was the poor knowledge before the signing of the acquisition contract, losing then the capacity of negotiation. Here arose the idea of “subcontracting with responsible supervision”. The real problem appears in the supplier selection activity, at least in “big projects”. Is the Regional Government who select the provider and it does not consider the selection criteria and procedures of Public Administrations.

- SG2. Satisfy supplier agreements.

  Description. Agreements with the suppliers are satisfied by both the project and the supplier.

  Findings. The Public Administrations have a subcontracting process (efficient or not, but they have it) but they do not have a monitoring and control process for lead the subcontracting project. This loss of control can be due to the lack of knowledge about subcontracting standards and models within the organizations. One of the mentioned alternatives was the use of subcontracting strategies; establishing process to accomplish the goals and design the strategy and control for the required service.

Proposals of Short-term Actions on Acquisition Management Process in Public Administrations

The participants exposed the following short-term actions:

- It is necessary to subcontract in a rational way, trying that the Public Administration maintains the strategy and functional analysis of project.
- The Public Administrations should have a deep knowledge about the product that they want to acquire: "If I do not have the knowledge, I do not know what we want to subcontract".
- Never forget that the subcontracting process does not avoid the work, generates new: monitoring and control the acquisition.
- Both parts (client and provider) must have clearly objectives and they must follow and efficient methodology to all levels (top level, functional, and more).

Conclusions about SAM Process in Public Administrations

The Public Administrations have certain mechanisms that lead the SAM process, usually they plan the project without metrics and they do not have monitoring process.

The lack of project control can be due to the lack of knowledge about subcontracting standards and procedures in the organizations.

It is important to mention that the Public Administrations have not known how to exploit the existing resources. A solution proposed by participants was not to imply in big projects or in its defect they can apply the "divide and conquer" theory. On this way, they would have many small projects that, by experience, are easy to monitor and control in an effective way.

We conclude that the Public Administrations use some model of effective practices, like Spanish methodology called METRICA3, to cover partially the SAM process. But their processes are not aligned with CMMI necessarily.

Conclusions

Before the Symposium on Improvement Process Models and Software Quality of Public Administrations, we invited to the General Administration of the State and 17 Autonomous Communities. With the purpose to make an extensive invitation, the Spanish Federation of Municipals and Provinces (FEMP) participated in the invitation visits.

In these visits we detect a poor knowledge on Improvement Models but we identify an increasing interest for participate in the Symposium discussions. Finally, three State Administrations, eleven Autonomous Administrations and two Local Administrations participated.

This represents a great advance, giving the poor knowledge and initial skepticism. We made a discussion forum to study and debate how the organizations could improve their current process, and to meet the recent studies on improvements models and their applicability in Public Administrations around the world.

With the assessment of Public Administration, in RM and SAM process specifically, we exposed the lack of control in these processes. This detection of "organization's weakness" promoted the organization's initiatives to begin an improvement process.

With these results, we accomplished the Symposium objectives. Firstly we obtain an initial assessment of RM and SAM processes of all Public Administrations and secondly we promote, among all Symposium participants, the idea that it is possible to improve and obtain the leadership in Public Administrations.

This Symposium was the first step, now it is the turn of Public Administrations, they should begin the formal assessment of their process to identify the strengths and weakness, and prioritize their improvements actions.

Bibliography

[1] Cuevas, G., "Training on Software Process Improvement Models across University", 1st Symposium on Improvement Process Models and Software Quality of Public Administrations. San Lorenzo del Escorial. Madrid, Spain. November, 2004.

[2] Paulk, M.C., Curtis, B., Chrissis, M.B., and Weber, C.V., "Capability Maturity Model for Software, Version 1.1", CMU/SEI-93-TR-024, Software Engineering Institute, Carnegie Mellon University, 1993.

[3] Product Development Team, "Capability Maturity Model Integration (CMMI), Version 1.1, Continuous Representation", CMU/SEI-2002-TR-011, Software Engineering Institute, Carnegie Mellon University, 2002.

[4] Product Development Team, "Capability Maturity Model Integration (CMMI), Version 1.1, Staged Representation", CMU/SEI-2002-TR-012, Software Engineering Institute, Carnegie Mellon University, 2002.

[5] Jordán, A., "SEI's Maturity Models Motivation and Evolution", 1st Symposium on Improvement Process Models and Software Quality of Public Administrations. San Lorenzo del Escorial. Madrid, Spain. November, 2004.

[6] Sivaramakrishnan, G. R., "Software Development in India, Why, For what, and How apply Software Process Models", 1st Symposium on Improvement Process Models and Software Quality of Public Administrations. San Lorenzo del Escorial. Madrid, Spain. November, 2004.

[7] Yard, J., "Maturity Models Implementation in Public Administration", 1st Symposium on Improvement Process Models and Software Quality of Public Administrations. San Lorenzo del Escorial. Madrid, Spain. November, 2004.

[8] Lee, J.-N., Huynh, M., Chi-Wai, R., and Pi, S.-M., "The Evolution of Outsourcing Research: What is the Next Issue," *presented at 33rd Hawaii International Conference on System Sciences*, 2003.

Authors' Information

Jose A. Calvo-Manzano – Universidad Politecnica de Madrid – Facultad de Informatica, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain; e-mail: jacalvo@fi.upm.es

Gonzalo Cuevas - Universidad Politecnica de Madrid – Facultad de Informatica, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain; e-mail: gcuevas@fi.upm.es

Ivan Garcia - Universidad Politecnica de Madrid – Facultad de Informatica, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain; e-mail: igarcia@zipi.fi.upm.es

Tomas San Feliu - Universidad Politecnica de Madrid – Facultad de Informatica, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain; e-mail: tsanfe@fi.upm.es

Ariel Serrano - Universidad Politecnica de Madrid – Facultad de Informatica, Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain; e-mail: aserrano@zipi.fi.upm.es

Magdalena Arcilla - Universidad Nacional de Educacion a Distancia – Escuela Tecnica Superior de Ingenieria Informatica, C/ Juan del Rosal 16, 28040 Madrid, Spain; e-mail: marcilla@issi.uned.es

Fernando Arboledas – Informatica y Comunicaciones de la Comunidad de Madrid (ICM), C/ Embajadores 181, 28045 Madrid, Spain; e-mail: fernando.arboledas@madrid.org

Fernando Ruiz de Ojeda - Informatica y Comunicaciones de la Comunidad de Madrid (ICM), C/ Embajadores 181, 28045 Madrid, Spain; e-mail: frdo2@madrid.org

GEN. A SURVEY APPLICATION GENERATOR

Hector Garcia, Carlos del Cuvillo, Diego Perez, Borja Lazaro, Alfredo Bermudez

***Abstract**: The National Institute for Statistics is the organism responsible for acquiring economical data for governmental statistics purposes. Lisbon agreements establish a framework in which this acquisition process shall be available through Internet, so each survey should be considered as a little software project to be developed and maintained. Considering the great amount of different surveys and all changes produced per year on each make impossible this task. An application generator has been developed to automate this task, taking as a start point the Word or PDF template of a survey, and going through a graphical form designer as all human effort, all HTML, Java classes and Oracle database resources are generated and sent from backoffice to frontoffice servers, reducing the team to carry out the whole set of electronic surveys to two people from non I.T. staff.*

Introduction

Complaining Lisbon agreements concerning e-Government, the Spanish National Institute for Statistics (INE) tackles the problem of translating all economical surveys from paper format into web applications. There exist hundreds of different forms, and for a particular survey, more than one version depending on the kind of target organization, so the required effort to create all infrastructures exceeds not only the capacity of I.T. Department, but the budget to carry out the gigantic task. A previous successful experience on metadata processing from INE and the pilot projects on Java application generation from Technical University of Madrid seem a proper combination to afford the trouble.

The idea consists of taking as a start point the current survey forms in Microsoft Word or PDF format, translating these into a tag based format appropriate for both browser representation and automated processing. This creates some kind of template used as a background for the application. Then a user may define the web form over the background painting components using a designer, and establishes properties for the components from those pre-defined in the designer. Finally only translating these definitions into source code is still to be done.

The technology of generated code shall meet the following requirements:

- HTML 4.01, later substituted by XHTML 1.1 by the research team at UPM, for the web user interfaces.
- XForms 1.0, for the definition of validation rules, with the premise to deploy complete surveys in XForms for future use.
- Java servlets, based on action struts architecture and their corresponding beans.
- Hibernate 3 as database connection tier.
- PDF format as receipt of the answered surveys.

The Basic Architecture

The main goal of the project, beyond any other, was to decrease sensitively the staff, effort and time to market for each survey application, and so, of the whole set of applications. The lack of I.T. professionals in the department in charge of the project also conditions the profile of the target user of the generator.

Four modules were found to be the core of the survey generator:

Format translation tool

As long as the forms corresponding to the different surveys are being created in other departments, the format and composition developed for hard copies is not valid for automated processing, some tool to extract the

contents from Word and PDF files and export them into tagged files, closer to web requirements and much more appropriated for processing.

For this purpose several options were evaluated. At first the best choice seemed to develop a specific translator, in Java language. The wide support for PDF processing available supposed a great advantage, but by that time the number of API or information about accessing Word files, especially about the structure of these files, was very poor. Only some arising APIs, such as Apache's POI, were available, so finally the decision went on a third party product, and then develop only the integration to the system.

Survey definition tool: the editor

Once the source document has been translated in a processable format, and before proceeding to its publishing, the system allows users to define the forms for the survey. Of course, there is not enough information in the templates, but visual aspect. In this sense, the captured survey is shown to the user as background in a screen, in which he may add or remove components that will later compose a web form.

For each field in the survey the editor allows to define some specific features, such as data type, length, etc. In case of combo boxes or lists, it is possible to define the valid values list. Also constraints have been implemented, such as date formats, decimal and thousands separator, ranges, allowed data sets, etc.

The components available while designing forms are the following:

Figure1.- Toolbar

Label: consists of a read only text. The user may define component name, text, font type, size and color, bold and italics style and background color.

Text: it is a read/write field containing characters, and it is possible to configure component name, length, maximum capacity, data type (string, date, time, year, day, month, float, double, integer, positive, negative, long) for validation rules to be applied. If the validation finds that the content of the field does not match the data type, the user shall be advised, so the application allows the user to set an error message to be displayed in a dialog. It is possible to apply some modifiers to the text fields, such as mandatory, read only, hidden and calculate value automatically. In case of selecting calculated field, a calculator is shown to define the formula. A formula may contain both values and fields in the same survey. There is also a description of the field to be shown as a hint. For some specific surveys, with repetitive contents, such as tables with a row per city or state, it is necessary to associate the field with a column in the table where data is to be stored.

Text from database: it is a read/write field associated to a column in a table from a database. The purpose is to set a default value for the field when the form is loaded. The field is processed as a text field once the form is submitted. It is very useful when defining, for instance, headers in a form, with the name, address and essential data for the final user. Possible configurations are identical to text fields.

Text area: consists of a text field with several rows. The user may define the size in terms of rows and columns, and data type, read only and mandatory options are available. For this component the validation rules may be disabled.

Radio, check boxes and lists: a new feature is added regarding the previous components: an interface to manage the choices and value for each choice in the available options, and establish a default selection if any.

Buttons: it is possible to draw buttons and associate a number of actions to them. Currently a button can save the form, add or delete items to repetitive contents (such as tables), validate the form and generate the errors list, or open the help page. Also specific separate components are available to create reset and submit buttons.

Of course a survey can be defined in several working sessions, so the functionalities for saving a survey in the current definition state, open a saved survey, and delete are available, as well as preview of the current survey. Finally, in the bottom of the toolbar, generate survey and exit commands are located.

If the user has already defined other surveys, the system enables the import functionality, so that the components from other surveys can be brought. This feature is especially useful to create surveys that are simple modifications from the previous year format, or to create the same survey in a different language, so only some modifications on descriptions and hints are needed.

Most surveys are so long that page breaks are needed. The editor allows defining these line breaks, which also are considered when generating the PDF receipts.

Application generator

The application generator retrieves the form already defined with the previous tool, starting a analysis process, in which the definition data are separated into modules, and afterwards the different application components are generated as a set of XHTML pages that are a composition starting with the original HTML from the PDF or Word, adding the information defined in the editor, and some GEN specific attributes for some tags that allow further processing. These attributes are used mainly to show the specific user information (i.e. tags with organization information, or the form with values that have been saved in a previous session).

Then the XForm files needed to validate the data in the forms are built to be used in web and bulk load processes, plus the resources for XHTML, such as background images, help files, etc.

This entire infrastructure is stored in the database, and when a request to fill in a form is produced a set of servlets, action struts and filters that parse the XHTML page to be showed, together with the required resources, get the result.

To generate PDF receipts from the submitted forms XSL:fo is used. As long as the forms are, after the process, HTML pages, a previous transform has to be done. Fo processes basically XML documents, and HTML does not provide the closing tags, for instance
, so the document is analyzed and translated into XHTML before processing. The PDF shall have same number of pages as defined in the survey using the page breaks.

Publishing tool

The first approach to publish all the resources generated by GEN was to send to the production server all classes, JSPs, servlets, images, etc. It is a remote server in a different network, and should be configured for allowing the access to file structures, execute compilations, modify Tomcat settings, … at the same time that the network devices where also reconfigured. Of course this is not an affordable task, obviously a software tool of this kind cannot cause all this changes.

Finally, this led us to refactor all design, and instead generating all the mentioned resources, two meta-servlets where produced, so that one of them is capable to draw any form and the other one is capable to receive (after submitting) any kind of form. This reduces sensibly the publishing process, now only access to the database is required, and all resources are stored there. The database is accessed through the corporate intranet.

Case Study

Probably the best way to describe how GEN works and the simplicity of the process to generate the new application to manage a survey is to follow an example.

First step: translating a PDF survey into HTML

The first of all, we need the file containing the survey designed by the corresponding department. Usually surveys have been translated into PDF, anyway this is an immediate process, usually covered by Adobe Acrobat Writer or Distiller, as well as by any tool complaining the standard.

This translation is based on an external tool, so we do not mind how the translation is done, the only important issue is the quality of the HTML we get after the translation. GEN only needs as input a HTML complaining the W3C standard. We always avoid HTML resulting from Microsoft Word exports, FrontPage, etc.

In this case study, we are using a translation tool which converts PDF into an HTML composed of layers. Each component of the document is set into a different layer, and the absolute position for each layer is written in the code. This is very useful for future processing of the template, while painting the components in the proper location over their relative texts and tables.

Second step: design the form for a survey

Now GEN user shall define a form to capture the data corresponding to the survey. The HTML from the previous step is considered as a simple background to help the user drawing all components. This is the main task while using the editor, but also some other functions are to be completed: buttons to send, save or step through the forms, page breaks for viewing and generating PDF receipts, or dependencies and constraints specification.

First of all, the typical screen where user is going to define the working directory, survey name and type. GEN allows defining multilingual surveys, so that the language specification is also important for further publishing. Each survey may be also converted in a template that may be used to create compositions in which a survey is composed of modules that are really survey templates. This is very useful to define headers or footers, or to include forms common to many applications.

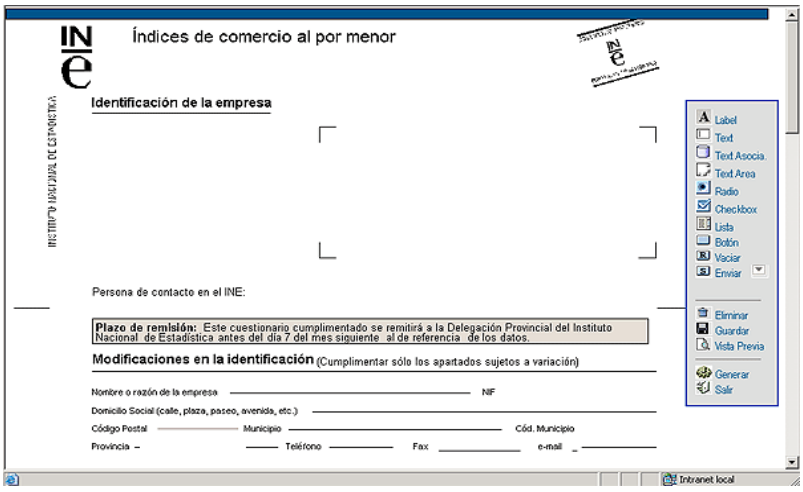


Figure 2.- The survey is loaded in the editor to create the form

Once the survey is loaded in the editor, the user may define the components and its location over the background. The layer information from the HTML obtained in the first step is very important now; when a user draws a component in the form, the editor calculates the position trying to associate the component to a field in the survey, matching them depending on the proximity.

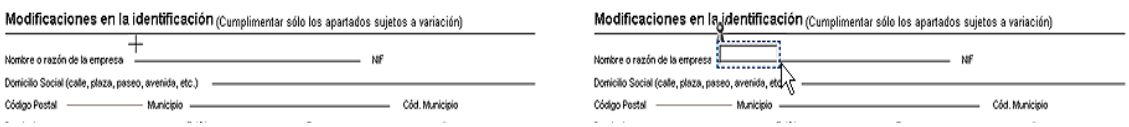


Figure 3.- Drawing a component from the toolbar

Components are managed as usual in an IDE, including the possibility of copy, paste, align vertically or horizontally, size, resize or set equal size for several of them, defining default values for each component type, etc.

In some cases it is very interesting to define a domain of values for some components, especially in case of combo boxes or lists, in the figure below a example on how to define a fixed domain for states is shown. The possibility of recovering dynamically the domain from a table in a database has been also implemented.

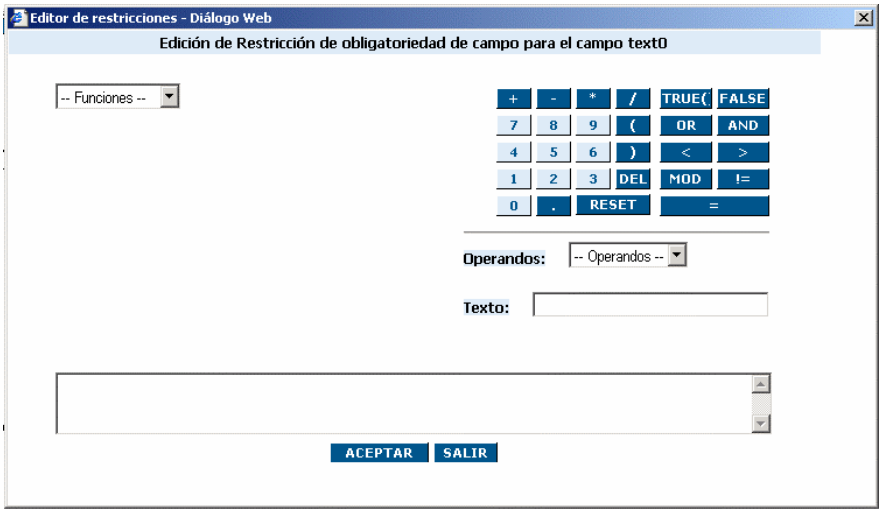


Figure 4.- Constraint definition

Another interesting functionality while defining surveys is to set some kind of constraint to the values of the different fields. The user may define both individual and group constraints. Individual constraints are to be applied to the value of a field, while group constraints are applied to the set of values of some fields. In any case, the constraints may obly to fill the field (*mandatory*) or define restrictions on the value of the field if any, through formula specification, as shown in the figure below. In the combo box the user may select some common predefined functions.

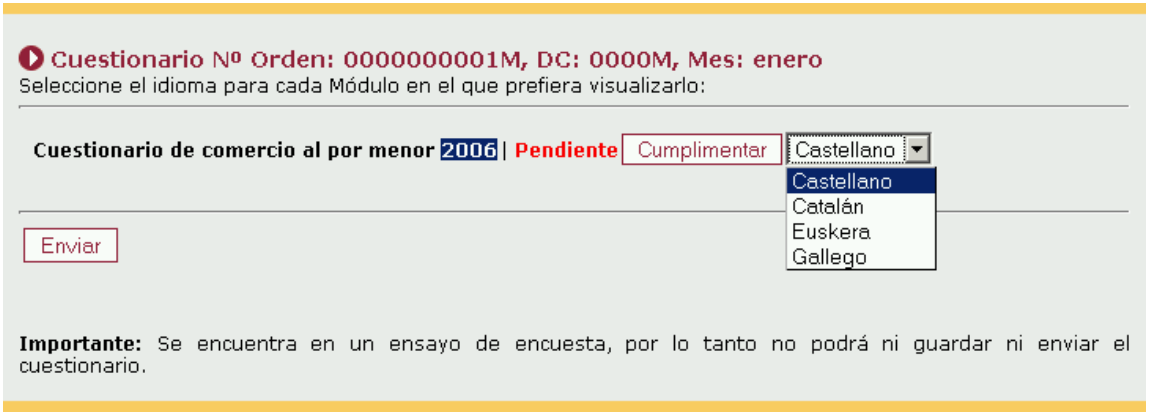


Figure 5.- Language selection. GEN generates multilingual applications

After some time defining the form associated to a survey, the user may have added and removed several components (in fact, usually, dozens of them). Each new component is named automatically by the editor and introduced in the tab order. The user may alter both, the given name to make the application comprehensive (very recommendable) and also the tab order, so that the final user may navigate through the fields with the tab control, which is one of the most accepted (and requested) features from final users.

Not only the form to publish the survey is important, but also to provide the users an identical aspect and information to reduce the difference to the hard copy surveys. In the same terms the background is the same that in that case, the designer provides the chance to add help buttons to the form, and then these buttons may be linked to a help file in HTML format that usually contains the same help provided to non Internet users.

The results: an external user fulfilling a survey

When the process is finally finished, the user gets a result which may be considered in two perspectives:

- An application that supports the publication and collection of data from the web, on a surprisingly low cost and effort, and which is integrated in the portal of the organization.
- A service for the information providers, those are obliged by law to compliment the surveys, and may now carry out this task easily.

The applications generated by GEN are integrated in the web site through shared tables in an Oracle database, and the development of a section in which the user may select the survey in which he is interested. The access to the section is based on the typical user and password that has been replaced by an order number and control digit. This information is sent to each organization together with customized instructions and paper surveys via mail as official notifications, reusing the existing infrastructure and the old procedure, which is mandatory to maintain.

When a survey is generated, the user may define whether the survey shall exist also in test mode. This allows final users to practice with real surveys without writing results or submitting information. The only difference between tests and real surveys is a parameter switching between them.

The surveys can be displayed in all Spanish official languages as mandatory by law, but the data are considered to correspond to the same survey, this is logical until a problem arises: the staff in charge of collecting data and extracting statistical information do not use to speak all of them. It is considered a great lack not for the application but for the legal procedure.

In the figure 6 a fragment of the web showing a survey is presented. Note the final alignment of the fields, the different component types and the buttons in the bottom, offering the functions of *Save*, *Send form*, *Cancel* and *Show errors*.

Before sending the final answer to a survey, the user may save any number of times, and before submitting may query the errors in the different fields in order to make corrections. Sending a correct answer is mandatory by law.

The final reply of a user to a survey is stored in XML format into the database, and will be later exported to the analysis application.

This is the common process for most users, however, a number of enterprises, mainly holdings, do need to fill lots of surveys that they may automate by using their information systems. A complementary application has been developed allowing this kind of users replying massively to the surveys through bulk loads. This application, called G2G, receives XML files containing several surveys, validates the data using the XForms from GEN, after formatting incorrect data or truncating data too long for the precision in database.

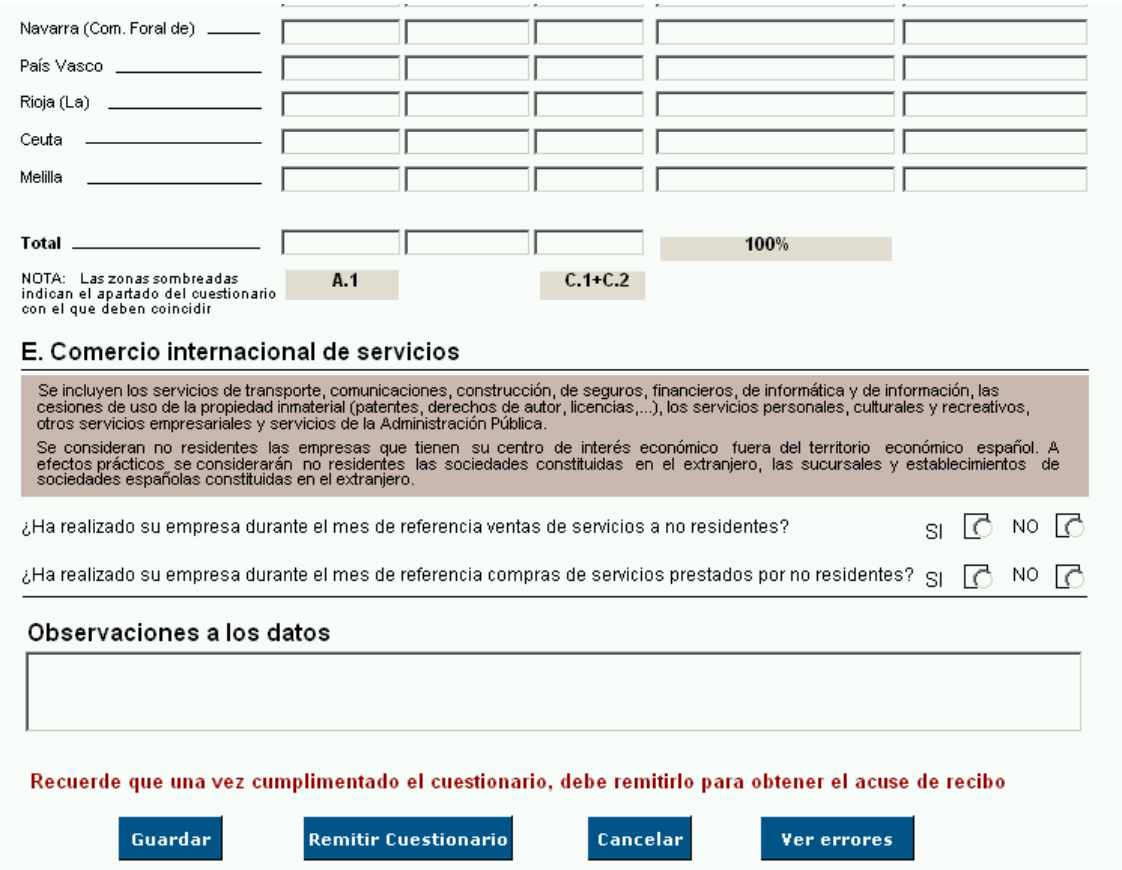


Figure 6. A piece of the final presentation of a survey on commerce

Conclusion

Sometimes the public organisations shall face political commitments with little help from the Government, same budgets and staff, allowing the development of beautiful projects in which technology makes life easier. Probably this project had lost the chance if there where enough staff to develop each application as usual software projects.

A tool has been developed that allows a non technical user, with no knowledge on software architectures, design, HTML, nor Java, struts or databases, carry out the task of creating applications. Of course this does not replaces technical staff, but get them to develop only technical high level tasks.

The fact is that a complex technology, that has been tested by Technical University of Madrid, and that shuns from the typical theoretical aspects from code generation, trying to open a new spectrum of possibilities, has been applied successfully to a specific application domain demonstrating that the approach is good and feasible.

For public administrations, the project allows a rise in the corporate image, and the observance of the Lisbon agreements in one of the most difficult aspects it was facing to.

The Spanish National Statistics Institute has been capable to meet the goals described in its services white-paper, minimizing the impact in the staff and budget, and reducing sensitively the time to market for each survey, which is available in Internet in less than a week, so that the complete process increases a minimum percentage from the previous one, with no digital survey. Also the time in which all data has been collected is lower, because there is no need to transcript all data to a computer.

Collaborators that fill in the forms, now have the chance to select the preferred choice: digital or paper surveys, knowing that digital ones provide the proper mechanisms to automate the calculations, report possible errors reducing the time to finish the job, etc.

The success in figures: in the first two months more than 10% of the surveys where submitted via Internet, more than 600 per day are being received, and 7 multilingual surveys have been generated; 168.433 organizations are working with the generated applications.

Bibliography

[Alves2002] Alves, L., von Staa, A., 2002. A construction process for artifact generators using a CASE Tool. Proceedings of Workshop on Generative Programming 2002. Austin, Texas, USA, pp. 7-10.

[Arisholm1998] Arisholm, E. et al. 1998. Incorporating Rapid User Interface Prototyping in Object-Oriented Analysis and Design with Genova. The Eighth Nordic Workshop on Programming Environment Research.

[Batory 1994] Batory, et al. 1994. The GenVoca Model of Software-System Generators. IEEE Software, 0704-7459/94.

[Boehm2000] Jongmoon, B., Boehm, B., 2000. Empirical Analysis of CASE Tool Effects on Software Development Effort. Technical Report University of South Carolina. Center for Software Engineering.

[Cuvillo2004a] Cuvillo, C. et al. 2004. Generation tool for DBMS focused applications. Applied Computing 2004. Lisbon. Portugal.

[Cuvillo2004b] Cuvillo, C. et al. 2004. Multiplatform web applications generated from relational data models. (In spanish). CISCI 2004. Florida, USA.

[Calejo2002] Calejo, M. et al. 2002. Web Application Maker - A model based approach to web database development. 6th International Conference on Enterprise Information Systems.

[Grønbæk 1991] Grønbæk, K.et al. 1991. ApplBuilder - an Object-Oriented Application Generator Supporting Rapid Prototyping. The 4th International Conference on Software Engineering & Its Applications.

[Schmidt2000] Schmidt, D., et al., 2000. POSA2: Pattern -Oriented Software Architecture: Patterns for Concurrent and Networked Objects.

Authors' Information

Hector Garcia - Contrated Professor. Technical University of Madrid. E.U. Informática. Ctra. de Valencia Km. 7. E28031 Madrid. e-mail: hgarcia@eui.upm.es

Carlos del Cuvillo - Associate Professor. Technical University of Madrid. E.U. Informática. Ctra. de Valencia Km. 7. E28031 Madrid. e-mail: ccuvillo@eui.upm.es

Diego Perez - Consultant. Technical University of Madrid. E.U. Informática. Ctra. de Valencia Km. 7. E28031 Madrid. e-mail: dperez@tdi.eui.upm.es

Borja Lazaro - Technician, Group leader. Technical University of Madrid. E.U. Informática. Ctra. de Valencia Km. 7. E28031 Madrid. e-mail: blazaro@eui.upm.es

Alfredo Bermudez - Project manager. National Institute for Statistics. c/ Rosario Pino 14-16. E28071 Madrid. e-mail: abc@ine.es

СИСТЕМА АВТОМАТИЗИРОВАННОЙ ПОДГОТОВКИ ВЫВОДОВ О ФИНАНСОВОМ СОСТОЯНИИ АКЦИОНЕРНЫХ ОБЩЕСТВ

Григорий Н. Гнатиенко, Николай Н. Маляр

Аннотация: *Описывается программное обеспечение автоматизированной подготовки экспересс-анализа финансового состояния предприятий. Система рассчитана на использование ее при осуществлении контрольно-аналитической деятельности по управлению государственными корпоративными правами. Приводятся области применения финансового анализа, структура и основные характеристики системы. Перечислены основные результаты применения описанного программного обеспечения.*

Ключевые слова: *анализ финансового состояния, государственные корпоративные прав, коэффициент, дивиденды.*

Введение

Для обоснованного принятия решений при управлении государственными корпоративными правами (ГКП) возникает необходимость в проведении оперативного анализа финансово-хозяйственной деятельности акционерных обществ. Важная роль финансового анализа при управлении ГКП объясняется тем, что в том числе и от качества предшествующего анализа зависит качество основанных на нем управленческих решений ([Гнатиенко, 1999а], [Гнатиенко, 1999б]).

Объективные показатели деятельности открытых акционерных обществ (ОАО), содержащих ГКП, целесообразно определять путем анализа финансовых коэфициентов. Эти коэффициенты рассчитываются на основе данных Балансов и Отчетов о фининсовых результатах и широко используются специалистами по экономической деятельности предприятий. Указанные показатели характеризуют финансовое состояние предприятия, его прибыли и убытки, изменения в расчетах с дебиторами и кредиторами, изменения в структуре активов и капитала, помогают определить тенденции, увидеть возможности развития и своевременно оценить потенциальные угрозы.

Области применения финансового анализа при управлении ГКП

Необходимость экспресс-анализа финансового состояния предприятия возникает во многих случаях. Такой анализ применяется, в частности ([Гнатиенко, 1999в], [Гнатиенко, 1999г]), при:

- подготовке макроэкономических решений на основании обобщения финансового состояния ОАО по отраслевым и региональным признакам;
- формировании условий и заданий по управлению ГКП органам государственной власти;
- разработке заданий при передаче полномочий по управлению уполномоченным лицам, определенным на конкурсной основе;
- подготовке к участию в общих собраниях акционеров ОАО с целью определения потенциальных финансовых возможностей обществ и формирования заданий на участие представителей государства в собрании акционеров;
- подготовке представителей государства к участию в работе наблюдательных советов акционерных обществ;

- подготовке представителей органов государственной власти к участию в работе ревизионных комиссий ОАО;
- заключении и перезаключении государственными органами контрактов с председателями правлений акционерных обществ, в уставных фондах которых государственная доля превышает 50 процентов, на осуществление ими функций по управлению обществом;
- рассмотрении финансового состояния ОАО при разрешении конфликтных ситуаций между субъектами корпоративного управления;
- подготовке материалов для проведения заседаний Комиссии по рассмотрению деятельности ОАО;
- проведении контрольных мероприятий с целью определения и проверки адекватности отчислений от чистой прибыли акционерного общества с целью начисления дивидендов на ГКП, наличии задолженности по заработной плате и бюджетных платежах.

Актуальность разработки системы

Особенности осуществления массовой приватизации в Украине стали причиной возникновения большого количества ГКП. В первые годы приватизации ГКП исчислялись тысячами. Динамика изменения количества и структуры ГКП в 2000-2006гг. приводится в таблице 1.

год	Всего ГКП	В том числе по размеру ГКП					
		0%-10%	10%-25%	25%-50%	50%-75%	75%-100%	100%
2000	624	163	81	263	52	39	26
2001	885	231	100	357	77	66	54
2002	1861	448	245	785	156	129	98
2003	1743	375	257	717	152	139	103
2004	1510	308	249	602	139	132	80
2005	1293	267	230	485	123	112	76
2006	1223	247	224	446	122	108	76

Таблица 1. Динамика изменения количества и структуры ГКП в 2000-2006гг.

Большое количество ГКП и необходимость оперативного предоставления достоверной, полной и удобной информации в Верховную Раду, Администрацию Президента, Кабинет Министров Украины обусловили необходимость создания и внедрения соответствующего программного обеспечения. Поэтому в Национальном агентстве Украины по управлению ГКП была разработана и внедрена автоматизированная система «Отчет» анализа финансовой отчетности ОАО.

Система «Отчет», как система поддержки принятия решений (СППР) ([Ларичев, 1996]) позволяет использовать данные, знания, объективные и субъективные модели для анализа и решения плохо структурированных проблем.

Структура системы «Отчет»

СППР «Отчет» состоит из нескольких подсистем (ПС), которые реализуют следующие функции.

ПС1. Оценка ситуации, выбор критериев и оценка их относительной важности. На начальном этапе разработки системы «Отчет» экспертным путем был осуществлен целенаправленный поиск финансовых коэффициентов для достижения целей экспресс-анализа. Из множества известных в экономическом анализе финансовых коэффициентов были выбраны те, которые являются критическими при управлении ГКП и могут служить индикаторами при принятии решений.

Финансовые коэффициенты являются относительными величинами, поэтому их анализ позволяет сравнивать акционерные общества различной величины и направлений деятельности. Для характеристики и экспресс-анализа деятельности предприятий были выбраны следующие коэффициенты: покрытия, быстрой ликвидности, обеспечения собственными средствами, общей рентабельности, финансовой стабильности, автономии. Для анализа вычислялись также следующие показатели: рабочий капитал, износ основных средств, дебиторская задолженность, затраты на единицу продукции.

ПС2. Генерация возможных решений: конкретизации направления запросов на проведение экспресс-анализа, их приоритет и т.п.

Результатом реализации этой ПС является множество *X*, состоящее из элементов $x \in X$, вида $x = \langle p,\ v,\ s,\ h \rangle$, где p – идентификатор акционерного общества; v – направление, по которому должен готовиться экспресс-анализ; s – дата проведения экспресс-анализа; h – вектор коэффициентов, характеризующих состояние акционерного общества.

Методика отбора акционерных обществ для экспресс-анализа является математическим объектом *m*, который определяется как тройка $m = \langle Q,\ D,\ R \rangle$, где Q – набор условий, при которых может применяться методика; D – набор данных из внутренних и внешних источников использования методики; R – алгоритм вычисления характеристик *x*.

ПС3. Ввод данных. Для информационной поддержки функционирования системы «Отчет» осуществляется регулярный ввод бухгалтерской отчетности, предоставляемой ОАО, других источников информации. При отсутствии необходимых данных информация об этом поступает в ПС3, в которой принимается решение об оперативном вводе недостающих данных.

Для обеспечения работы ПС3 в системе содержатся сведения о:

- правилах формирования начальной выборки ОАО;
- вводе необходимых данных;
- разбивке акционерных обществ на группы для анализа;
- правилах выбора методов классификации ОАО, шкалах критериев и т.д.

ПС4. Подготовка текстового отчета о финансовом состоянии ОАО. В системе «Отчет» программно реализованы следующие функции:

- генерация соответствующего текстового фрагмента $T1$, пригодного для дальнейшего использования специалистами, не имеющими экономического образования, на основе данных фининсовой отчетности;

- на основании анализа вычисленных коэффициентов автоматизированная формулировка выводов о состоянии ОАО и определение его финансового состояния (текстового фрагмента $T2$);
- окончательный выбор формулировок и подготовка окончательного отчета на основании текстовых фрагментов $T1$ и $T2$.

Для того, чтобы сгенерировать текстовый отчетный документ, необходимо значениям коэффициентов поставить в соответствие фразы, характеризующие полученный уровень показателей. Для этого для каждого коэффициента определены интервалы значений. Разбивка интервалов для каждого коэффициента осуществлялась экспертным путем. Анализ осуществлялся относительно нормативных значений, установленных для каждого коэффициента.

Различным комбинациям значений коэффициентов соответствуют несколько вариантов предложений, из которых генерируется окончательный отчет.Комбинации значений коэффициентов, которым поставлено в соответствие выражение «Ситуация невозможна», являются невозможными с точки зрения экономического анализа. Это свидетельствует об ошибках при вводе бухгалтерской отчетности, об ошибках при составлении отчетности на предприятиях или при передаче информации от различных источников.

С помощью ПС4 осуществляется вывод на экран монитора и печать справки о финансовом состоянии ОАО или результатов анализа группы акционерных обществ при формировании аналитических отчетов.

ПС5. Расчет коэффициентов, характеризующих деятельность ОАО, по которому готовится отчет и передача данных в архив.

ПС6. Обеспечение обмена информацией между ПС, согласование групповых решений между департаментами Национального агентства.

Эта ПС выполняет следующие основные функции:

- определение общего количества N акционерных обществ, которые могут быть проверены, исходя из наличных трудовых ресурсов;
- применение методик $m_{обяз}$ определения акционерных обществ, которые подлежат обязательному отбору для проведения экспресс-анализа;
- применение методов $m_{ТППР}$ теории поддержки принятия решений для агрегирования результатов решения задач экспресс-анализа состояния отдельных ОАО;
- применение методик $m_{стох}$ обоснованного выбора предприятий для включения их в перечень для составления плана документальных проверок организаций из числа не проверявшихся в течение заданного срока либо из других категорий ОАО, определяемых руководством Национального агентства.

ПС7. Архивирование результатов работы системы «Отчет» с целью динамического комплексного анализа возможных последствий принятых решений. Для объективности анализируется информация о деятельности предприятия за несколько отчетных периодов. Осуществляется архивирование значений вычисленных финансовых коэффициентов, информации о датах и адресатах предоставления отчетов, а также о реакции адресатов на предоставленные им отчеты.

Структура системы представлена в виде схемы на Рис.1.

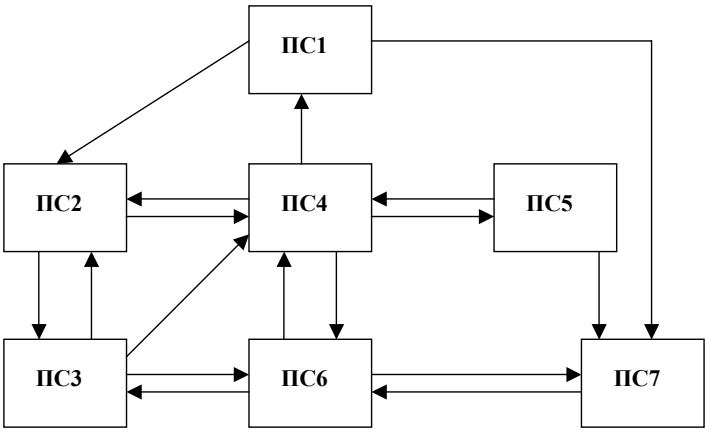


Рис.1. Структура системы “Отчет”

Основные характеристики системы «Отчет»

Основными характеристиками разработанного программного обеспечения являются:

- доступность в эксплуатации и дружественный интерфейс пользователя;
- обработка ошибок, возникающих при вводе бухгалтерской отчетности;
- целостность структуры данных;
- возможность экспорта выходных форм в формат MS Word;
- близость текста отчета о финансовом состоянии ОАО к естественному языку.

Система реализована в среде Access и предназначена для роботы под операционной системой MS Windows на базе протокола TCP/IP. Вибор оболочки Access обусловлен несколькими причинами:

- Access является штатной системой управления базами данных для компьютеров Microsoft, интегрирована с MS Office и является удобным инструментом;
- Работа в среде Access по стилю аналогична с другими инструментами из MS Office, что значительно упрощает проблемы освоения системы;
- Access является локальной СУБД с большим набором встроенных процедур.

Основные результаты применения системы «Отчет»

Внедрение автоматизированной обработки отчетности о финансово-хозяйственной деятельности ОАО позволило:

- осуществлять оперативный анализ финансово-хозяйственной отчетности акционерных обществ, имеющих ГКП;
- организовать единый централизованный архив отчетов о деятельности ОАО, их систематизацию, удобный доступ к информации;
- формализовать технологические процессы создания и обработки выводов о деятельности предприятий;

- улучшить средства контроля за ходом выполнения технологических процессов обработки отчетных документов, что существенно повысило продуктивность труда;
- предоставить возможность консолидации отчетности по отраслевым и региональным признакам;
- повысить возможности оперативного рассмотрения динамики деятельности акционерных обществ;
- добиться достижения высокого уровня независимости работы с отчетностью от личностных качеств персонала путем автоматизации выполнения большинства формальных функций.

Программное обеспечение подготовки принятия решений о финансовых возможностях акционерных обществ было внедрено и использовалось в Контрольно-аналитическом департаменте Национального агентства Украины по управлению ГКП и принесло значительный экономический эффект.

Концептуальный подход и оригинальная методика, применялись Национальным агентством при прогнозировании объемов поступления в Государственный бюджет дивидендов от управления корпоративными правами государства и определении региональной структуры дивидендных отчислений акционерных обществ. На основании осуществленных таким образом прогнозов формировалась дивидендная политика государства ([Гнатиенко, 1999д], [Гнатиенко, 1999е]).

Описанный программный комплекс использовался:

- для подготовки принятия решений об участии представителей государства в 1300 общих собраниях акционеров;
- при разработке свыше 800 условий и заданий органам государственной власти (министерствам, ведомствам, областным государственным администрациям) по управлению ГКП;
- при формировании свыше 100 заданий председателям правлений ОАО при заключении контрактов с ними;
- во многих других ситуациях принятия решений с области управления корпоративными правами государства.

Выводы

Система «Отчет» предоставила возможность автоматизировать процесс подготовки принятия решения центральным органом исполнительной власти, которым являлось Национальное агентство Украины по управлению ГКП. Положительной чертой комплекса является возможность использования результатов его работы лицам, которые не есть специалистами в обрасти компьютерных технологий, финансового и экономического анализа.

Об эффективности использования системы «Отчет» свидетельствует, в частности то, что план поступления в Государственный бюджет дивидендов от субъектов предпринимательской деятельности, в уставных фондах которых есть ГКП, в 1999 году был перевыполнен более чем на 30 процентов: поступило 53 млн.грн. вместо 40 млн.грн., запланированных постановлением Кабинета Министров Украины от 26 января 1999г. №88 «О мероприятиях по обеспечению своевременного и в полном объеме поступления доходов в Государственный бюджет Украины на 1999 год».

Авторы выражают благодарность проф. Волошину А.Ф. за постановку проблемы.

Библиография

[Гнатиенко, 1999а] Гнатиенко Г.Н., Лозовая Т.И. Деятельность уполномоченных лиц при управлении государ. корпоративными правами//Бизнес (бухгалтерия, право, налоги, консультации), 1999, №30, С.104-108.

[Гнатиенко, 1999б] Гнатиенко Г.Н., Лозовая Т.И. Деятельность уполномоченных лиц при управлении государственными корпоративными правами (продолжение)//Бизнес (бухгалтерия, право, налоги, консультации), 1999, №31, С.103-108.

[Гнатиенко, 1999в] Гнатієнко Г.М. Нормативно-правове забезпечення управління корпоративними правами держави/Матеріали міжнародної конференції "Управління державною власністю: правова модель та досвід інших країн"//Київ, 6 травня 1999р. - Київ, 1999.-С.44-50.

[Гнатиенко, 1999г] Гнатиенко Г.Н., Олейник Г.И. Управление государственными корпоративными правами // Бизнес (бухгалтерия, право, налоги, консультации), 1999, №35, С.112-114.

[Ларичев, 1996] Ларичев О.И., Мошкович Е.М. Качественные методы принятия решений. М.Наука. Физматлит. 1996.

[Гнатиенко, 1999д] Гнатієнко Г.М. Деякі особливості дивідендної політики при управлінні державнимии корпоративними правами // "Українська Інвестиційна Газета", 1 червня 1999, №21, NB. Тематичний випуск. Дивіденди. Нарахування та оподаткування, С.9-10.

[Гнатиенко, 1999е] Гнатієнко Г.М. Нарахування дивідендів при управлінні державним майном // Інформаційний бюлетень "Янус Нерухомість", 1999, №14, С.12-13.

Информация об авторах

Григорий Н. Гнатиенко – Киевский национальный университет им. Т. Шевченко, факультет кибернетики, докторант. Киев, Украина, e-mail: G.Gnatienko@veres.com.ua

Николай М. Маляр – Ужгородский национальный университет, математический факультет, декан

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ПРИ РАЗРАБОТКЕ ИНФОРМАЦИОННЫХ СИСТЕМ

Мария Е. Еремина

***Аннотация**: В работе предлагается подход к описанию БД, позволяющий при работе с ней не учитывать особенности физического хранения реляционных данных и тип управляющей СУБД. Описанный формально, в терминах графовой модели, этот подход позволил разработать некоторые алгоритмы, в частности алгоритмы верификации предметной области. Данная теория была реализована на практике, и в составе CASE-системы METAS была внедрена в опытную эксплуатацию.*

***Ключевые слова:** информационная система, база данных, метаданные, математическая модель, граф.*

Введение

В настоящее время возрастает потребность в создании крупных информационных систем, работающих в неоднородной среде. Под термином *информационная система* (ИС) понимается комплекс информационных ресурсов (то есть ИС – это информация и инструменты управления ею), технологий получения и обработки данных, поддержания их в актуальном и непротиворечивом состоянии. Это определение отражает точку зрения пользователя системы, а с точки зрения реализации, ИС – это сложный комплекс управленческих и технологических решений, компьютерной аппаратуры и программного обеспечения, а также информационного содержания (наполнения).

Основная часть любой ИС – *база данных* (БД), сложность и объем которой постоянно возрастает, что порождает все новые трудности при работе с ней. В частности, в современных БД возможна ситуация, когда разные узлы контролируются разными *системами управления базами данных* (СУБД). Такие ИС и БД принято называть *неоднородными* [2].

Данная работа посвящена рассмотрению подходов к описанию БД в крупных неоднородных ИС. В частности, рассматривается возможность работы с подобными БД, используя привычные для пользователя термины предметной области, и давая ему возможность не учитывать особенности физического хранения реляционных данных и тип управляющей СУБД.

Для обеспечения рассматриваемой возможности предложен подход, основанный на введении в ИС дополнительных метаданных о ней. Данный подход был описан формально, с помощью графовой модели этих метаданных. Использование этой модели позволило формализовать некоторые алгоритмы для работы с данными в ИС, например, в задачах верификации.

Разработка данного подхода ведется в рамках проекта создания CASE-средства METAS, позволяющего автоматизировать разработку ИС.

Технология METAS

CASE-технология METAS (METAdata System) – это основа для создания систем, управляемых метаданными. Данная технология предназначена для снижения трудоемкости разработки корпоративных информационных систем и повышения их гибкости, масштабируемости и адаптируемости в процессе эксплуатации. Ключевым моментом технологии является использование взаимосвязанных метаданных, описывающих информационную систему предприятия (учреждения, организации – любой бизнес-системы) [7].

Большинство существующих CASE-систем генерируют код на каком-либо языке программирования в соответствии с некоторыми спецификациям, описывающим предметную область. Основное отличие описываемой CASE-системы состоит в том, что она использует эти спецификации в виде метаданных во время своей работы. Таким образом, процесс создания приложения сводится к описанию необходимых метаданных.

Применение метаданных дает возможность гибкой настройки приложения и его функциональности, а также реструктуризации информационных объектов, описанных метаданными. Это создает хорошие предпосылки для создания «интеллектуальной» системы, которая сможет настраиваться на потребности пользователя и меняющиеся условия эксплуатации «самостоятельно», в ходе работы с ней. Кроме того, при таком подходе проект обладает высокой степенью обратной связи, так как разработчик, меняя метаданные, сразу видит соответствующие изменения в создаваемой ИС. При обычном же подходе,

реализованном в большинстве других CASE-систем, разработчик сначала описывает систему, затем запускает генерацию, и только после этого может оценить результат внесенных изменений [7].

Все метаданные в системе разделены на *модели* (или *уровни*), каждая из которых описывает определенную часть, аспект ИС. Некоторые модели могут описывать одни и те же части ИС, но с различных точек зрения. Все модели взаимосвязанны, одна модель может основываться на другой, и представлять собой более высокоуровневое описание ИС [8].

Данная работа посвящена рассмотрению логическая модели метаданных (ЛМ). Она непосредственно опирается на физическую модель (ФМ), которая описывает все объекты базы данных ИС: таблицы и их поля, связи между таблицами, индексы и многое другое. Подробнее описание ФМ можно найти в [8].

Помимо названных, в системе существуют еще ряд моделей метаданных. Например, презентационная модель описывает визуальный интерфейс пользователя при работе с объектами ИС [6], модель бизнес-процессов – бизнес-операции и бизнес-процессы, поддерживаемые ИС, модель репортинга описывает запросы, первичные документы и отчеты, формируемые системе, а модель защиты – метаданные, представляющие пользователей системы и их права на выполнение операций над объектами ИС или на доступ к моделям метаданных [8]. В данной работе все эти модели рассматриваться не будут.

Основные цели создания логической модели

При использовании любой реляционной БД возникает ряд трудностей, которые обуславливаются именно особенностями хранения реляционных таблиц и необходимостью нормализации. Структура хранимых данных в этом случае часто отличается от той, которую хотел бы видеть пользователь. Рассмотрим эту ситуацию более подробно.

Первая трудность возникает еще на этапе анализа предметной области и проектировании БД. Основные методы, используемые для этого – различные диаграммы типа «сущность-связь» (ERD – entity-relationship diagrams) или диаграммы классов в языках типа UML. Во всех этих подходах необходимо учитывать специфику реляционной модели данных. Человек, занимающийся анализом и проектированием, должен оперировать информацией на уровне физических таблиц и связей между ними, а также знать основные принципы нормализации реляционных БД. Вместе с тем, он должен достаточно глубоко разбираться в предметной области. Все это предъявляет достаточно серьезные требования к квалификации разработчика.

Другая трудность возникает при дальнейшей работе с созданной БД. При построении SQL-запросов к ней необходимо снова вспоминать всю специфику хранения информации в реляционных таблицах и учитывать ее.

В CASE-системе METAS для преодоления этих трудностей был использован следующий подход. На этапе проектирования по-прежнему строится диаграмма, состоящая из сущностей и связей между ними. Однако здесь сущности уже соответствует не одна, а несколько связанных между собой таблиц БД. Это приближает понятие сущности (как объекта реального мира) к тому, которое мы используем в жизни, что позволяет проектировать в терминах реальной предметной области.

Существенным преимуществом подхода является возможность дальнейшей работы в этих терминах, абстрагируясь от физического способа хранения информации для каждой сущности. В частности, это обеспечивает CASE-системе METAS свойство независимости от типа используемой СУБД и возможность работы с неоднородными БД.

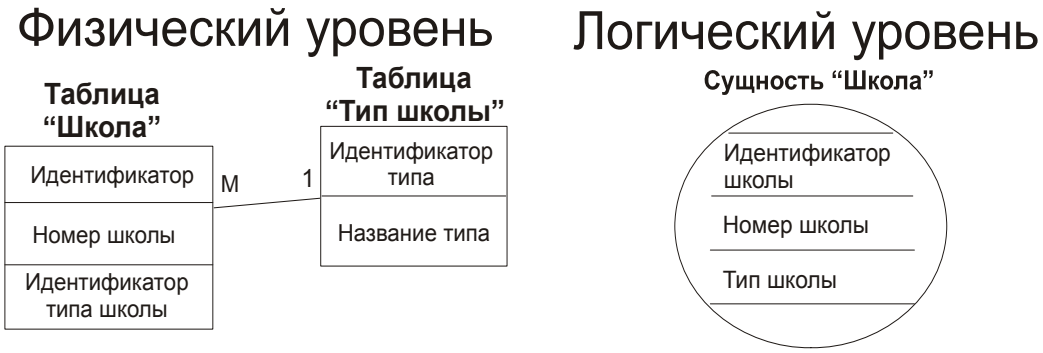
Для практической реализации изложенного выше подхода был создан компонент *«Логическая модель» (ЛМ)*, который опирается на метаданные логического уровня. Используя этот компонент, все остальные компоненты системы имеют возможность работы с БД не напрямую, а в терминах логического уровня, что значительно облегчает их написание. При этом существует достаточно простая возможность добавления в ЛМ новой функциональности, например, возможность генерации новых видов запросов.

В данной статье мы рассмотрим только основные принципы и понятия, используемые при построении ЛМ. Более подробное ее описание можно найти, например, в статьях [3,4].

Основные понятия логической модели

В ЛМ реально существующие объекты представляются набором характеристик (атрибутов). Эти атрибуты физически могут храниться в разных таблицах ФМ. Поэтому введем некоторую обобщающую логическую конструкцию – **сущность** – представляющую собой совокупность этих атрибутов.

Для иллюстрации рассмотрим фрагмент БД, хранящей информацию о школах, а именно, их номерах и типах (например, лицей, гимназия). Очевидно, что в обычной реляционной БД для представления этой информации необходимо создать две таблицы, одна из которых содержит внешний ключ другой. Это и будет «физический уровень» метаданных.



На логическом уровне для данного фрагмента можно выделить сущность «*Школа*» с атрибутами «*Номер*» и «*Тип школы*». При этом «*Тип школы*» становится таким же атрибутом, как и остальные, с тем только отличием, что для него устанавливается пометка, что его значения выбираются из справочника. Далее сущность сама обеспечивает всю работу с этим справочником прозрачно для пользователя. Таким образом, на физическом уровне каждой сущность соответствует набор таблиц. Одна таблица является главной, остальные связаны с ней соотношением «1:M», то есть являются справочниками для нее.

Основная задача сущности – обеспечивать пользователям возможность работы в терминах предметной области, не задумываясь над тем, как информация представлена на физическом уровне. То есть на логическом уровне уже не надо заботится о том, какое поле находится в какой таблице и как эти таблицы связаны. Теперь пользователь может оперировать в терминах сущности, то есть просто указывать ее название в сочетании с названием атрибута. Сущность сама построит SQL выражения, необходимые для работы с БД, и либо сама же их и выполнит (операции вставки, удаления, обновления информации в БД), либо вернет готовое SQL выражение пользователю (операция выборки).

Иногда возникают ситуации, когда необходимо в достаточно сжатом виде представить основную информацию о конкретном реальном объекте некоторой сущности. Например, ситуация, когда две сущности связаны связью «1:M». При этом сущность со стороны M имеет атрибут, представляющий

родительскую сущность (внешний атрибут). Пусть в сущности «*Ученик*» есть внешний атрибут «*Школа*». В нем необходимо представить не всю информацию о школе, где он учится, а только какую-то основную, достаточную для ее идентификации пользователем, например ее номер. Что именно для данной сущности является такой «основной информацией» и определяется ее презентационным выражением. Для школы это может быть город и номер, для ученика – ФИО и т.д.

Презентационное выражение представляет собой SQL-подобное выражение, в котором могут присутствовать любые атрибуты представляемой сущности. В нем допускается использование стандартных операций и функций, содержащихся в языке SQL. Все встречающиеся атрибуты должны быть записаны в квадратных скобках. Например, для ученика допустимо такое презентационное выражение: *[familia]+[imia]+[otchestvo]*.

Как уже было сказано выше, каждая сущность представляет собой совокупность **атрибутов**. Атрибуты у сущности бывают нескольких видов.

- *Собственный атрибут сущности*. Это любой атрибут из главной таблицы.
- *Атрибут справочника*. Для каждого справочника сущности в нее добавляется отдельный атрибут, который имеет ссылку на соответствующее ему информационное поле таблицы-справочника. При работе с этим атрибутом можно выбирать необходимые значения из списка значений в справочнике или вводить новое значение (тогда оно будет занесено в справочник).
- *Внешний атрибут*. Если данная сущность связана с другой отношением «M:1», то у нее появится дополнительный атрибут, содержащий информацию об этой связи. Например, в сущности «*Ученик*» есть внешний атрибут, представляющий родительскую сущность «*Школа*». Работа с внешними атрибутами аналогична работе с атрибутами справочника, но его значения можно только выбирать, а не изменять или добавлять.
- *Ключевой атрибут.* На самом деле это обычный собственный атрибут. Он содержит ссылку на ключевое поле главной таблицы, то есть однозначно определяет значения всех остальных атрибутов. Это понятие вводится исключительно для удобства работы.

С атрибутами сущности не следует путать так называемый *презентационный атрибут*. Он не является атрибутом в обычном смысле, его значение не хранится в БД. Это некоторое значение, вычисляющееся на основе презентационного выражения и представляющее основную информацию о сущности. С атрибутом его роднит то, что оно вместе со значениями остальных атрибутов хранится в каждом экземпляре сущности.

Введем понятие **связи между сущностями**. Оно является аналогией связей между таблицами, только на логическом уровне, то есть в терминологии предметной области. Например, сущность «*Школа*» может быть связана с сущностью «*Город*» связью M:1, определяющей ее местонахождение.

Поддерживается два основных типа связи, каждый из которых имеет свои особенности.

- *1:M.* В этом случае связь соответствует физической связи «1:M» между главными таблицами сущностей.
- *M:M.* Данная связь соответствует физической связи «M:M», которая в реляционных СУБД реализуется с помощью промежуточной таблицы.

Для любой связи может конкретизироваться количество сущностей с каждой стороны. Для этого у каждой стороны связи указывается минимальное и максимальное количество сущностей, участвующих в ней. Например, в типе «1:M» возможны следующие подтипы: «1 : 2..3», «0..1 : 0..1», «0..1 : 1» и т.д.

Построение математической модели

Для построения алгоритмов работы CASE-системы METAS необходимо формализовать описание используемых в ней метаданных всех уровней. Приведем формальное определение систем метаданных физического и логического уровней в виде графовой модели. В качестве примера ее использования опишем один из алгоритмов верификации предметной области.

Физический уровень

Пусть *Fields* – некоторое абстрактное множество. Элементы этого множества назовем *полями*. Множество *Tables* взаимно не пересекающихся подмножеств полей назовем *множеством таблиц*, а его элементы – *таблицами*. Для каждой таблицы множество ее полей назовем $F(t) \subset Fields$.

Определим бинарное отношение $R \subset Fields \times Tables$. Это отношение назовем множеством связей между таблицами. Будем использовать следующую запись:

$$r \in R \Leftrightarrow r=(f,t),\ f \in Fields,\ t \in Tables.$$

Таким образом, каждая связь $r \in R$ проводится между каким-то полем $f \in Fields$, принадлежащим одной из таблиц, и другой таблицей целиком (то есть целому подмножеству множества *Fields*). При этом возможна ситуация, что $r \in R$, $r=(f,t)$, $t \in Tables$, $f \in F(t)$, то есть связь проходит от поля некоторой таблицы к этой же таблице.

Логический уровень

Пусть *Attr* – некоторое абстрактное множество. Элементы этого множества назовем *атрибутами*.

Разобьем все множество атрибутов на группы, каждую из которых назовем *сущностью*. Множество всех сущностей *Ent* должно быть *дизъюнктивным*, Это означает, что с одной стороны, каждый атрибут множества *Attr* должен принадлежать какой-либо сущности, с другой стороны, сущности не пересекаются. Если обозначить за $A(e) \subset Attr$ – множество всех атрибутов, принадлежащих сущности $e \in Ent$, то формально справедливы следующие соотношения: $\bigcap_{e \in Ent} A(e) = \varnothing,\ \bigcup_{e \in Ent} A(e) = Attr$.

В каждой сущности все атрибуты можно разбить на 3 непересекающихся множества. Для любой сущности $e \in Ent$ назовем множество $O(e) \subset A(e) \subset Attr$ множеством *собственных атрибутов сущности*, $S(e) \subset A(e) \subset Attr$ – множеством *атрибутов справочников*, $V(e) \subset A(e) \subset Attr$ – множеством *внешних атрибутов*. Таким образом, справедливы условия непересечения этих множеств:

$$\forall e \in Ent\ \ O(e) \cap S(e) = \varnothing,\ V(e) \cap S(e) = \varnothing,\ O(e) \cap V(e) = \varnothing,$$

и условие обязательной принадлежности каждого атрибута сущности одному из этих множеств:

$$\forall e \in Ent\ \ \forall a \in A(e) \Rightarrow a \in O(e) \vee a \in S(e) \vee a \in V(e).$$

Подмножество атрибутов сущности, используемое при построении презентационного атрибута, обозначим P(e). В нем могут присутствовать атрибуты из любых множеств O(e), S(e), V(e).Таким образом,

$$\forall e \in Ent\ \ P(e) \subset A(e),\ P(e) \neq \varnothing.$$

Отсюда получим более строгое определение понятия сущности. *Сущность* определяется пятеркой

$$e = (A, O, S, V, P),$$

где каждое из последних четырех множеств есть подмножество множества атрибутов с описанными выше свойствами.

На всех приводимых ниже рисунках будем использовать следующие графические обозначения введенных понятий. Графически все элементы множества *Attr* будем обозначать кружками, а каждый элемент множества *Ent* большим кругом. Если атрибут принадлежит какой-либо сущности, то он будет расположен на ее круге. Цвет закраски атрибута будет обозначать, к какому из множеств O(e), S(e) или V(e) он принадлежит. Соответственно, это будут цвета белый, серый и черный. Множество атрибутов, составляющих презентационный атрибут, будем обводить пунктирной линией.

Имея фиксированное множество атрибутов, можно различными способами выделить сущности так, чтоб они удовлетворяли условию дизъюнктивности. В каждой выделенной сущности можно различными способами выделить множества O(e), S(e), V(e), P(e). Построенной модели это противоречить не будет. Однако для практического использования модель должна соответствовать реальной предметной области, а в ней все разбиения атрибутов на сущности и атрибутов по типам определяется однозначно. Поэтому можно считать, что данные разбиения однозначно определены.

Обозначим Y – множество всех внешних атрибутов всех сущностей. Тогда $\mathrm{Y} = \bigcup_{\mathrm{e}\in\mathrm{Ent}} \mathrm{V(e)}$. Введем отображение *W*, которое каждому $v \in Y$ ставит во взаимно однозначное соответствие некоторую сущность $e \in Ent$. Другими словами:

$$w \in W \Leftrightarrow w=(v,e),\ v \in V(e_1),\ e,e_1 \in Ent.$$

Таким образом, каждая связь $w \in W$ проводится между одним из внешних атрибутов одной сущности и некоторой другой сущностью. При этом возможна ситуация, что $w \in W$, $w=(v,e)$, $e \in Ent$, $v \in V(e)$, то есть связь проходит от сущности к ней самой. Графически такая связь будет обозначаться направленной дугой от внешнего атрибута к сущности.

Второе бинарное отношение $M \subset Ent \times Ent$ назовем множеством связей «M:M» («многие-ко-многим») между сущностями. Так как в предметной области между двумя сущностями может быть несколько связей «M:M», ее модель будет представлена *мультиграфом или графом с параллельными ребрами*, то есть графом, в котором между вершинами может быть несколько параллельных дуг [9]. Для их идентификации предлагается каждой паре ставить в соответствие пометку, ее уникальное имя. Это имя в дальнейшем будет использоваться при реализации, например при построении дерева объектов, используемого для навигации пользователя по системе [6]. Следовательно, каждая связь «M:M» представляет собой тройку $(e_1,e_2,name)$, где e_1,e_2 – связываемые сущности, а name – уникальное имя этой связи. Будем использовать следующую запись:

$$m \in M \Leftrightarrow m=(e_1,e_2,name),\ e_1,e_2 \in Ent,\ name \in String.$$

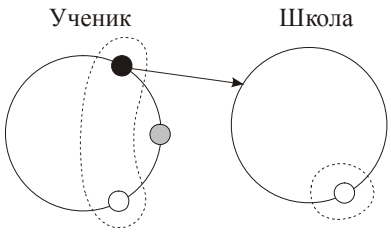
Таким образом, каждая связь $m \in M$ проводится между двумя сущностями целиком. Графически такие связи будем изображать ненаправленными дугами между сущностями.

Рассмотрим пример. Пусть у нас есть сущности «*Школа*» с собственным атрибутом *«Номер» и* «*Ученик*» с атрибутами «*ФИО*» (собственный атрибут), «*Пол*» (атрибут справочника), «*Школа ученика*» (внешний атрибут), со следующими атрибутами. Они связаны как «1:M» через атрибут «*Школа ученика*».

У ученика можно в качестве презентационного атрибута можно использовать выражение «*ФИО*»+’ ‘+«*Школа ученика*», а для школы – значение ее атрибута *«Номер».*

В этом примере множество *Attr* состоит из четырех элементов, *Ent* – из двух. При этом если $e_1 \in Ent$ – это сущность *«Ученик»*, $e_2 \in Ent$ – это сущность *«Школа»*, то $A(e_1)$ состоит из трех элементов, $A(e_2)$ – из одного. $O(e_1)$, $S(e_1)$, $V(e_1)$ содержат по одному атрибуту ученика, а множество $P(e_1)$ – два атрибута: *«ФИО»* и

«Школа ученика». Для школы множества $S(e_2)$ и $V(e_2)$ пусты, а множества $O(e_1)$ и $P(e_1)$ содержат единственный атрибут *«Номер»*.



Тройку L=(*Ent*, *W*, *M*) назовем *полным графом предметной области*. Под термином «полный» подразумевается, что в модель включена вся информация о предметной области. В дальнейшем мы будем рассматривать также упрощенные графы, полученные из полного с помощью какого-либо преобразования. Например, если в какой-то задаче нас не интересуют связи «M:M», то в полном графе необходимо заменить множество *M* на пустое множество. Полученный упрощенный граф может быть удобнее для решения этой задачи.

Из сказанного выше следует, что полный граф предметной области обладает следующими особенностями: это *смешанный* граф (то есть с ориентированными и неориентированными дугами), с параллельными ребрами и петлями.

В терминах рассмотренных графовых моделей можно создать алгоритмы, решающие различные задачи в описываемой CASE-системе. Например, это могут быть алгоритмы выделения подсхем данных, алгоритмы реинженеринга и миграции [1]. В следующем разделе будет рассмотрен другой пример применения этих моделей.

Использование графовой модели для верификации метаданных

Построенная графовая модель представляет метаданные системы, которые используются для генерации всей функциональности ИС. Если при построении метаданных разработчик сделает ошибку, то он может этого не заметить, так как система метаданных обычно достаточно сложная. В то же время в процессе построения ИС по неверным метаданным могут возникнуть нежелательные эффекты. Поэтому в системе должны быть реализованы алгоритмы, которые еще на этапе проектирования анализируют структурные статические свойства модели и позволяют выявить некоторые типы ошибок.

Для описания некоторых таких алгоритмов в терминах модели зачастую возможно использовать *упрощенные* графы предметной области. Они могут быть получены из полного с помощью отображений, различных для каждого алгоритма. Например, для некоторых алгоритмов из графа могут быть убраны связи «M:M», для других – атрибуты.

Приведем пример алгоритма проверки модели предметной области на наличие циклов.

В данном случае под термином *циклом* подразумевается следующее. Пусть у нас есть две сущности, в каждой из которых есть внешний атрибут, родительской сущностью для которого является другая сущность. Пусть оба этих атрибута входят в состав презентационного атрибута для своих сущностей. Тогда при попытке запросить информацию о первой сущности, необходимо узнать значение презентационного атрибута второй сущности. Но так как в него входит внешний атрибут, для этого надо запросить информацию о презентационном атрибуте первой сущности. В результате образуется циклическая ссылка, и система не сможет выполнить запрос. Естественно, что такой цикл может включать

и более двух сущностей и быть менее очевидными. Поэтому подобные циклы необходимо исключать еще на этапе проектирования.

Рассмотрим формальную модель этой ситуации.

В графе предметной области в этом случае нас не интересуют ни связи «M:M», ни собственные атрибуты, ни атрибуты справочника. Поэтому в упрощенном графе в множестве *Attr* будут присутствовать только внешние атрибуты. Сущности и множества их презентационных атрибутов P и внешних атрибутов V останутся те же. Не изменится и множество *W* (то есть связи «1:M»). При этом надо учесть, что нарушится условие $P(e)\neq\varnothing$, но это не является существенным для данной задачи.

Получим граф $L_1=(Ent_1,W,\varnothing)$, где $Ent_1=\{e_i\}$, $e_i=(Attr(e_i) \setminus O(e_i) \setminus S(e_i),\ \varnothing,\ \varnothing, V(e_i),\ P(e_i))$, то есть

$L_1=(Ent_1,W,\varnothing)$, $Ent_1=\{e_i\}$, $e_i=(V(e_i), \varnothing, \varnothing, V(e_i),\ P(e_i))$.

Введем формальное определение цикла.

Пусть $RV(e_1,e_2)$ – множество атрибутов, удовлетворяющих следующему соотношению:

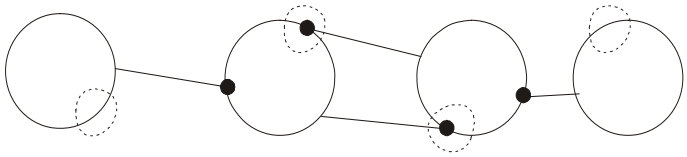
$$\forall a\in RV(e_1,e_2),\ a\in V(e_1),\ a\in P(e_1),\ \exists w_1\in W,\ w=(a,e_2),\ e_1,e_2\in Ent.$$

Пусть выполняется следующее условие:

$\exists a_1\in RV(e_1,e_2), \exists a_2\in RV(e_1,e_2), \dots, \exists a_{n-1}\in RV(e_{n-1},e_n), \exists a_n\in RV(e_n,e_1), e_i\in Ent$, i=1..n.

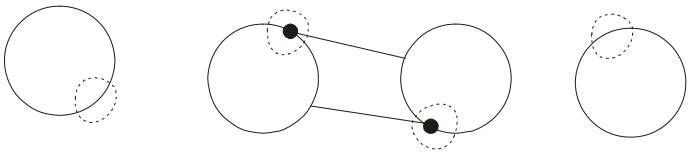
В этом случае будем говорить, что в графе предметной области присутствует *цикл*.

Пример графа предметной области, содержащий цикл, приведен на рисунке.



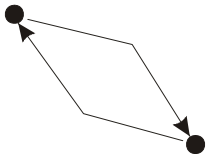
Очевидно, что если внешний атрибут не входит в состав презентационного, то на наличие циклов он не влияет. Поэтому наш граф можно еще упростить, убрав из него все подобные атрибуты. Соответственно, необходимо убрать и соответствующие им связи «1:M». Получим следующий граф:

$$L_2=(Ent_2,W_2,\varnothing),\ Ent_2=\{e_i\},\ W_2=(v_{ij},e),\ e\in Ent,\ \ e_i=(\{v_{ij}\}, \varnothing, \varnothing, \{v_{ij}\}, P(e_i)),\ v_{ij}\in V(e_i),\ v_{ij}\in P(e_i).$$



Далее этот граф можно свести к классическому направленному графу, если принять, что все внешние атрибуты – это вершины графа, а дуги от одной вершины к другой присутствуют тогда и только тогда, когда внешний атрибут, соответствующей первой вершине, связан с сущностью, содержащей внешний атрибут, соответствующий второй вершине:

$$G=(GV,GE),\ GV=\{v_i\},\ v_i\in Attr_2\ ,\ GD=\{(v_i,v_j)\},\ \forall v_j\in A_2(e),\ w=(v_i,e)\in W_2,\ e\in Ent_2$$



Здесь под $Attr_2$ подразумевается множество всех атрибутов графа L_2 , а под $A_2(e)$, $e\in Ent_2$ – его подмножество для сущности e.

Далее можно использовать любой алгоритм поиска циклов в графе.

Заключение

Таким образом, в работе предложен подход к описанию БД, позволяющий работать с ней в привычных для пользователя терминах предметной области. При этом не требуется учитывать особенности физического хранения реляционных данных и тип управляющей СУБД. Описанный формально, в терминах графовой модели, этот подход позволил разработать алгоритм верификации предметной области.

Изложенная теория получила практическое использование при создании ядра CASE-системы METAS. На основе последней были разработаны ИС «Образование Пермской области» и «Межведомственная ИС персонифицированного учета детей Пермской области», которые были успешно внедрены в опытную эксплуатацию в системе образования Пермского края России.

Результаты работы допускают обобщение на случай распределенной БД, когда информация хранится на разных узлах сети. В данный момент ведутся разработки в этом направлении [5].

Список литературы

[1] Борисова Д.А. Компонент реструктуризации CASE-системы METAS. // Межвуз. сб. науч. трудов / Математика программных систем. Пермь: Перм. ун-т., 2003. с. 34-42.

[2] Дейт К., Дж. Введение в системы баз данных, 7-е издание.: пер. с англ. М.: Издательский дом «Вильямс», 2001. – 1072 с.

[3] Еремина М.Е. Генерация SQL-выражений на основе метаданных. // Межвуз. сб. науч. Трудов / Математика программных систем. Пермь: Перм. ун-т., 2003. с. 43-50.

[4] Еремина М.Е. Использование метаданных для генерации SQL-запросов к базе данных // Современные проблемы механики и прикладной математики: Сборник трудов международной школы-семинара. Воронеж: ВГУ, 2004. с. 216-219.

[5] Еремина М.Е. Реализация запросов в распределенных неоднородных системах, управляемых метаданными // 5-ая Всероссийская научно-практическая конференция молодых ученых, аспирантов и студентов. Молодежь. Образование. Экономика. Сборник научных статей участников конференции. Часть 4. Ярославль: Ремдер, 2004. с. 65-71.

[6] Куделько Е.Ю. Генерация и настройка экранных форм на основе метаданных. // Межвуз. сб. науч. трудов / Математика программных систем. Пермь: Перм. ун-т., 2003. с. 51-59.

[7] Лядова Л.Н., Рыжков С.А. CASE-технология METAS // Математика программных систем / Межвуз. сб. науч. трудов. Перм. ун-т. Пермь, 2003. С. 4-18.

[8] Рыжков С.А. Концепция метаданных в разработке информационных систем // Математика программных систем / Межвуз. сб. науч. трудов. Перм. ун-т. Пермь, 2002. С. 36-44.

[9] Свами М., Тхуласираман К. Графы, сети и алгоритмы. М.: Мир, 1984. – 454 с.

Информация об авторе

Еремина Мария Евгеньевна – Пермский Государственный Университет; 614990, Россия, г. Пермь ул. Букирева 15; e-mail: erm6@mail.ru

Networks and Telecommunications

DIMENSIONING OF TELECOMMUNICATION NETWORK BASED ON QUALITY OF SERVICES DEMAND AND DETAILED BEHAVIOUR OF USERS

Emiliya Saranova

***Abstract**: The aim of this paper is to be determined the network capacity (number of internal switching lines) based on detailed users' behaviour and demanded quality of service parameters in an overall telecommunication system. We consider detailed conceptual and its corresponded analytical traffic model of telecommunication system with (virtual) circuit switching, in stationary state with generalized input flow, repeated calls, limited number of homogeneous terminals and losses due to abandoned and interrupted dialing, blocked and interrupted switching, not available intent terminal, blocked and abandoned ringing (absent called user) and abandoned conversation.*

We propose an analytical - numerical solution for finding the number of internal switching lines and values of the some basic traffic parameters as a function of telecommunication system state. These parameters are requisite for maintenance demand level of network quality of service (QoS). Dependencies, based on the numerical-analytical results are shown graphically.

For proposed conceptual and its corresponding analytical model a network dimensioning task (NDT) is formulated, solvability of the NDT and the necessary conditions for analytical solution are researched as well. It is proposed a rule (algorithm) and computer program for calculation of the corresponded number of the internal switching lines, as well as corresponded values of traffic parameters, making the management of QoS easily.

***Keywords:** Telecommunication Network, Circuit Switching, Network Traffic, Terminal Traffic, Human Factors, Network Dimensioning.*

1. Introduction

The purpose of the teletraffic theory is to find relation between quality of services and equipment cost [Iversen 2004]. This is very important for a good planning and controlling of telecommunication networks.

The Quality of service (QoS) concept is defined in the ITU-T Recommendation E-800 as: "The collective effect of service performance, which determines the degree of satisfaction of a user of the service".

QoS parameters are administratively specified in Service Level Agreement (SLA) between users and operators. These QoS parameters (from a contract of SLA) are reflecting on GoS parameters.

Network dimensioning is necessary for designing and control of network and its level of quality of services (QoS), in an advance determined level.

Based on a given set of QoS requirements, a set of GoS (Grade of service) parameters are selected and determined as functions of human behaviour characteristics.

2. Conceptual Model

In this paper we consider detailed conceptual and its corresponded analytical traffic model [Poryazov 2005b] of telecommunication system with channel switching, in stationary state, with BPP (Bernoulli-Poisson-Pascal) input flow, repeated calls, limited number of homogeneous terminals and losses due to abandoned and interrupted dialing, blocked and interrupted switching, not available intent terminal, blocked and abandoned ringing and abandoned conversation.

The conceptual model of the telecommunication system includes the paths of the calls, generated from (and occupying) the A-terminals in the proposed network traffic model and its environment (shown on Fig. 1).

The names of the virtual devices used are constructed according to the device position in the model.

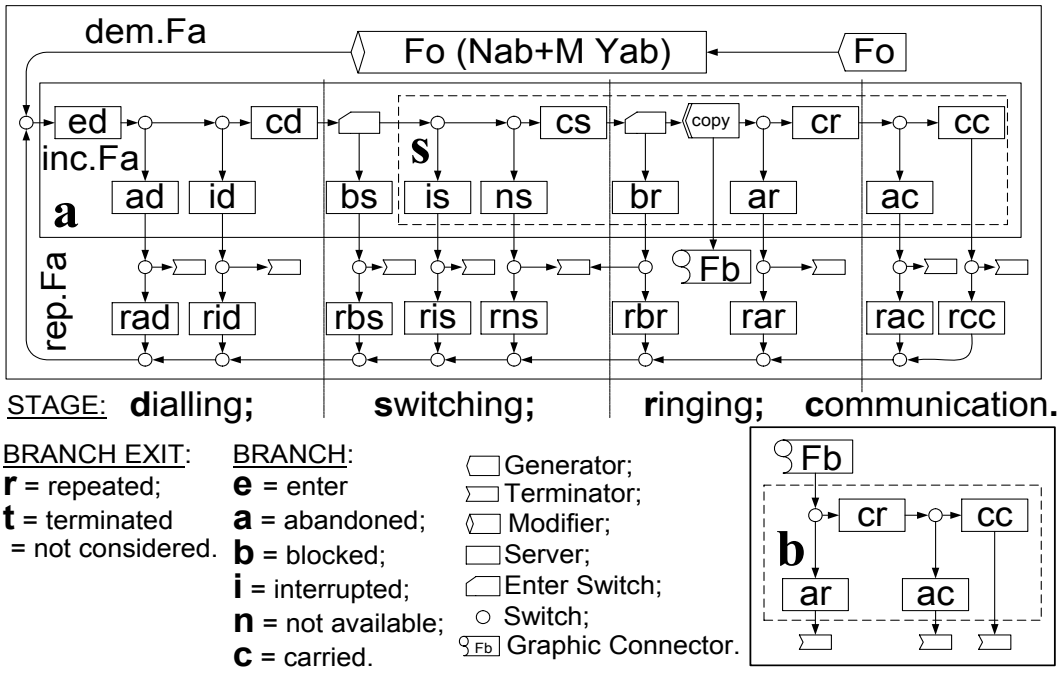


Fig. 1. Normalized conceptual model of the telecommunication system and its environment and the paths of the calls, occupying A-terminals (**a**-device), switching system (**s**-device) and B-terminals (**b**-device); base virtual device types, with their names and graphic notation.

2.1. The Comprising Virtual Devices

The following important virtual devices on Fig.1 are shown and considered:

a = comprises all the A-terminals (calling) in the system (shown with continuous line box).

b = comprises all the B-terminals (called) in the system (box with dashed line).

ab = comprises all the terminals (calling and called) in the system (not shown on Fig.1);

s = virtual device corresponding to the switching system. It is shown with dashed line box into the **a**-device. ***Ns*** stand for the capacity (number of equivalent internal switching lines) of the switching system.

2.2. Stages and Branches in the Conceptual Model:

Service *stages*: **d**ialing, **s**witching, **r**inging and **c**ommunication.

Every service stage has *branches*: **e**nter, **a**bandoned, **b**locked, **i**nterrupted, **n**ot available, **c**arried (correspondingly to the modeled possible cases of ends of the calls' service in the branch considered).

Every branch has two *exits*: **r**epeated, **t**erminated (which show what happens with the calls after they leave the telecommunication system). Users may make a new bid (repeated call), or to stop attempts (terminated call).

2.3. Device Parameters and its Notations in the Conceptual Model:

Letter *F* stands for intensity of the flow [calls/sec.], *P* = probability for directing the calls of the external flow to the device considered, *T* = mean service time, in the device, of a served call [sec.], *Y* = intensity of the device traffic [Erl], *N* = number of service places (lines, servers) in the virtual device (capacity of the device). In the normalized models [Poryazov 2001], used in this paper, every virtual device, except switches, has no more than one entrance and/or one exit. Switches have one entrance and two exits. For characterizing the intensity of the flow, we are using the following notation: *inc.F* for incoming flow, *dem.F*, *ofr.F* and *rep.F* for demand, offered and repeated flows respectively (ITU E.600). The same characterization is used for traffic intensity (*Y*).

Fo is the intent intensity of calls of one idle terminal; *inc.Fa* = *Fa* is intensity of incoming flow of A-terminals and *M* is a constant, characterizing the BPP flow of demand calls (*dem.Fa*). If *M* = -1, the intensity of demand flow corresponds to Bernoulli (Engset) flow model, if *M* = 0 - to the Poisson (Erlang), and if *M* = +1 - to the Pascal (Negative Binomial) flow model. In our analytical model every value of *M* in the interval [-1, +1] is allowed. The BPP-traffic model is very applicable [Iversen 2004], but in the numerical examples, presented here, *M* = *0,* because the conclusions made are independent of the input flow model.

2.4. The Main Assumptions of the Model:

For creating a simple analytical model, we make the following system of fourteen (A-1 – A-14) assumptions [Poryazov 2005b]:

A-1. (Closed System Structure) We consider a closed telecommunication system with functional structure shown in Fig. 1;

A-2. (Device Capacity) All base virtual devices in the model have unlimited capacity. Comprising devices are limited: ab-device contains all the active $Nab \in [2, \infty)$ terminals; switching system (s) has capacity of *Ns* calls (every internal switching line may carry only one call); every terminal has capacity of one call, common for both incoming and outgoing calls;

A-3. (A-Terminal Occupation) Every call, from the flow incoming in the telecommunication system (*inc.Fa*), falls only on a free terminal. This terminal becomes a busy A-terminal;

A-4. (Stationarity) The system is in stationary state. This means that for every virtual device in the model (including comprising devices like switching system), the intensity of input flow F(0, t), call holding time T(0, t) and traffic intensity Y(0, t) in the observed interval (0, t) converge to the correspondent finite numbers F, T and Y, when $t \to \infty$. In this case we may apply the theorem of Little (1961) and for every device: $Y = FT$;

A-5. (Calls' Capacity) Every call occupies one place in a base virtual device, independently from the other devices (e.g. a call may occupy one internal switching line, if it find free one, independently from the state of the intent B-terminal (busy or free));

A-6. (Environment) The calls in the communication systems' environment (outside the blocks a and b in Fig. 1) don't occupy any telecommunication systems' device and therefore they don’t create communication systems' load. (For example, unsuccessful calls, waiting for the next attempt, are in "the head of" the user only. The calls and devices in the environment form the intent and repeated calls flows). Calls leave the environment (and the model) in the instance they enter a Terminator virtual device;

A-7. (Parameters' undependability) We consider probabilities for direction of calls to, and holding times in the base virtual devices as independent of each other and from intensity Fa = inc.Fa of incoming flow of calls. Values

of these parameters are determined by users' behavior and technical characteristics of the communication system. (Obviously, this is not applicable to the devices of type Enter Switch, correspondingly to Pbs and Pbr);

A-8. (Randomness) All variables in the analytical model may be random and we are working with their mean values, following the Theorem of Little.

A-9. (B-Terminal Occupation) Probabilities of direction of calls to, and duration of occupation of devices ar, cr, ac and cc are the same for A and B-calls;

A-10. (Channel Switching) Every call occupies simultaneously places in all the base virtual devices in the telecommunication system (comprised of devices a or b) it passed through, including the base device where it is in the moment of observation. Every call releases all its occupied places in all base virtual devices of the communication system, in the instant it leaves comprising devices a or b.

A-11. (Terminals' Homogeneity) All terminals are homogeneous, e.g. all relevant characteristics are equal for every terminal;

A-12. (A-Calls Directions) Every A-terminal directs uniformly all its calls only to the other terminals, not to itself;

A-13. (B-flow ordinariness) The flow directed to B-terminals (Fb) is ordinary. (The importance of A-13 is limited only to the case when two or more calls may reach simultaneously a free B-terminal. A-13 may be acquitted from results like in (Burk 1956) and (Vere-Jones 1968);

A-14. (B-Blocking Probability for Repeated attempts) The mean probability (Pbr) of a call to find the same B-terminal busy at the first and at the all following repeated attempts is one and the same.

3. Analytical Model

3.1. Some General Equations

For the proposed conceptual model we derived the following system of equations (Poryazov, Saranova 2005):

$$Yab = Fa[S_1 - S_2(1 - Pbs)\ Pbr - S_3\ Pbs] \tag{1.1}$$

$$Fa\ =\ dem.Fa\ +\ rep.Fa \tag{1.2}$$

$$dem.Fa\ =\ Fo\ (Nab\ +\ M\ Yab) \tag{1.3}$$

$$rep.Fa = Fa\ [R_1 - R_2\ Pbr\ (1 - Pbs) - R_3\ Pbs] \tag{1.4}$$

$$Pbr\ = \begin{cases} \dfrac{Yab\text{-}1}{Nab\text{-}1} & in\ \ case\ of\ \ \ 1 \le Yab \le Nab, \\ 0 & in\ case\ of\ \ \ \ 0\ \le Yab < 1. \end{cases} \tag{1.5}$$

$$Ts\ = S_{1Z} - S_{2Z}\ Pbr \tag{1.6}$$

$$ofr.Fs\ =\ Fa\ (1\text{-}Pad)(1\text{-}Pid) \tag{1.7}$$

$$ofr.Ys\ =\ ofr.Fs\ \ Ts \tag{1.8}$$

$$Pbs\ =\ Erl\_b\ (Ns,\ ofr.Ys) \tag{1.9}$$

$$crr.Ys\ =\ (1 - Pbs)\ ofr.Ys \tag{1.10}$$

The following notations are used:

$$S_1 = Ted + Pad\,Tad + (1-Pad)[Pid\,Tid + (1-Pid)[Tcd + Pis\,Tis + (1-Pis)[Pns\,Tns + (1-Pns)[Tcs + 2Tb]]]] \quad (2.1)$$

$$S_2 = (1-Pad)(1-Pid)(1-Pis)(1-Pns)[2Tb - Tbr] \quad (2.2)$$

$$S_3 = (1-Pad)(1-Pid)[PisTis + (1-Pis)[Pns\,Tns + (1-Pns)[Tcs + 2Tb]]] - (1-Pad)(1-Pid)\,Tbs \quad (2.3)$$

$$S_{1z} = Pis\,Tis + (1-Pis)[PnsTns + (1-Pns)(Tb + Tcs)] \quad (2.4)$$

$$S_{2z} = (1-Pis)(1-Pns)(Tb + Tcs) \quad (2.5)$$

$$R_1 = Pad\ \ Prad\ + (1-Pad)\,(Pid\ \ Prid\ + (1-Pid)\ Pis\ \ Pris\ + (1-Pis)(Pns\ \ Prns\ + (1-Pns)\ Q) \quad (2.6)$$

$$R_2 = (1-Pad)(1-Pid)(1-Pis)(1-Pns)(Prbr - Q) \quad (2.7)$$

$$R_3 = (1-Pad)(1-Pid)\{Pis\,Pris + (1-Pis)[Pns\,Prms + (1-Pns)Q] - Prbs\} \quad (2.8)$$

$$Q = Par\,Prar + (1-Par)[Pac\,Prac + (1-Pac)\,Prcc] \quad (2.9)$$

An important assumption for proposed analytical model is:

The intent intensity of calls of one idle terminal is $Fo \geq 0$.

3.2. General Blocking Probability

Based on the conceptual model we define general blocking probability as follows:

Definition: General blocking probability (Pbl):

$$Pbl = \{Pbr \oplus Pbs, Pbr \in (0,1), Pbs \in (0,1) : (1-Pad)(1-Pid)\,Pbs + (1-Pbs)(1-Pis)(1-Pns)\,Pbr\} \quad (2.10)$$

Pad, Pid, Pis, Pbs, Pns, Pbr, Par, Pac and Pcc are known probabilities (see the conceptual model).

3.3. Probabilities of Blocking Switching (*Pbs*) and of Finding B-Terminal Busy (*Pbr*).

If $Pbr \in [0,1], Pbs \in (0,1]$ then each duple (*Pbr*, *Pbs*) define a value of *Pbl* throw (2.10) and back, each value of *Pbl* define a set of duples (*Pbr*, *Pbs*).

As GoS- parameter we consider general blocking probability *Pbl* based on (2.10).

Analogously, *adm.Pbl* (administratively determined value of *Pbl* in SLA in advance) defines set of duples (*adm.Pbr*, *adm.Pbs*) and back.

We consider general blocking probability (*adm.Pbl*) as a main QoS parameter, administratively determined in advance in SLA.

4. Network Dimensioning Task

4.1. Formulation of a Network Dimensioning Task (NDT):

1. To be dimensioned a network (to be found necessary number of internal switching lines), when in advance level of QoS is administratively determined and the values of known parameters are dimensioned and/ or calculated.
2. To be found the values of the unknown parameters, describing the system state in the upper case. For example, a system parameter, describing macrostate of the system (through the value of *Yab*), a terminal capacity of the system (the maximal number of active terminals *Nab*), intensity of demanded and repeated call attempts (respectively *dem.Fa* and *rep.Fa*), offered to the switching system traffic intensity (*ofr.Ys*) and others.

Parameters in the Network Dimensioning Task:

Administrative determined parameters:

adm.Pbl and *M* (3.1)

Known parameters:

Fo, Tb, *S*1, *S*2, *S*3, *S*1Z, *S*2Z, *R*1, *R*2, *R*3 (3.2)

Aim: To determine the number of switching lines *Ns*; and the following unknown parameters:

Yab, Fa, dem.Fa, rep.Fa, ofr.Fs, Ts, ofr.Ys (3.3)

Condition:

Pbl (*Pbr*, *Pbs*) ≤ *adm.Pbl* (3.4)

4.2. Solvability of the NDT:

The traffic intensity *Yab* characterizes the macrostate of the system. In Poryazov, Saranova (2005) is shown that

$$Yab=\frac{F_0(S_1-S_3Pbs)-(F_0(S_1-S_3Pbs)+F_0S_2(1-Pbs))Pbr+F_0S_2(1-Pbs)Pbr^2}{F_0(S_1-S_3Pbs)-(F_0M(S_1-S_3Pbs)+F_0S_2(1-Pbs)-1+R_1-R_3Pbs)Pbr+(1-Pbs)(F_0MS_2+R_2)Pbr^2} \quad (3.5)$$

Theorem 1: If $Pbr \neq 0$ and $Fo \neq 0$, then analytical presentation (3.8) of *Yab* in the NDT exist.

Proof: Considering the system equations (1.1) - (1.10) when $Pbr \neq 0$ and $Fo \neq 0$ from (1.5) and (1.3) follows

$$dem.Fa=\frac{Fo}{Pbr}[\,Pbr-1+(M\;Pbr+1)\,Yab] \quad (3.6)$$

From (1.2) and (1.4) follows

$$dem.Fa=Fa\;\{1-R_1+R_2\;Pbr+(R_3-R_2\;Pbr)\,Pbs\} \quad (3.7)$$

Then (3.6), (3.7) and (1.2) gives

$$Yab=\frac{F_0(1-Pbr)\{S_1-S_2Pbr-(S_3-S_2Pbr)Pbs\}}{F_0(1+MPbr)\{S_1-S_2Pbr-(S_3-S_2Pbr)Pbs\}-Pbr\{1-R_1+R_2Pbr+(R_3-R_2Pbr)Pbs\}} \quad (3.8)$$

(3.8) is new simplified expression of the (3.5).

If *Fo* = 0, then obviously *Fa* = 0, *dem.Fa* =0 and *rep.Fa* = 0.

Therefore, when $Pbr \neq 0$ in NDT, on the base of administrative determined values of parameters *Pbs*, *Pbr*, *M* and the known parameters (3.2), traffic intensity *Yab* is derivable. The other system parameters in the NDT are depending on the system state (respectively on *Yab*).

We will prove that the values of unknown parameters (3.3) in the NDT can be derived (evaluated) through *Yab* and known parameters (3.2) in correspondence of determined conditions.

Theorem 2: If

$$Pbr \neq 0 \ and \ \ Pbs \neq \frac{S_1 - S_2 Pbr}{S_3 - S_2 Pbr} \tag{3.9}$$

in the NDT, then for each unknown parameter of (3.3), an analytical expression for its evaluation exists.

Proof: Using the system (1.1) – (1.10) and from (1.1) and (3.5) by (*S*1 – *S*2 *Pbr*) - (*S*3 – *S*2 *Pbr*) *Pbs* ≠ 0, follows

$$Fa = \frac{Yab}{S_1 - S_2 Pbr - (S_3 - S_2 Pbr) Pbs} \tag{3.10}$$

For *dem.Fa* from (1.3) and (1.5) is received (3.6).

It is resulted from (1.4) and (3.9):

$$rep.Fa = \frac{Yab\{R_1 - R_2 Pbr - (R_3 - R_2 Pbr) Pbs\}}{S_1 - S_2 Pbr - (S_3 - S_2 Pbr) Pbs} \ . \tag{3.11}$$

From (1.6) and (3.9) follows:

$$ofr.Fs = \frac{Yab \ (1 - Pad)(1 - Pid)}{S_1 - S_2 Pbr - (S_3 - S_2 Pbr) Pbs} \tag{3.12}$$

The parameter *Ts* can be calculated from (1.7), and from (1.3) and (1.5) follows:

$$ofr.Ys = \frac{(1 - Pad)(1 - Pid)(S_1 z - S_2 z \, Pbr) Yab}{S_1 - S_2 Pbr - (S_3 - S_2 Pbr) Pbs} \tag{3.13}$$

Therefore, the values of the unknown parameters (3.3) in the NDT can be expressed and calculated by the conditions of the Theorem 1 and Theorem 2.

For the network dimensioning, when the level of QoS is determined administratively in advance (for example blocking probability Pbs), Erlangs'B - formula may be used:

$$Pbs = Erl\_b(Ns, ofr.Ys) \tag{3.14}$$

$$Erl\_b(Ns, ofr.Ys) = \frac{\frac{(ofr.Ys)^{Ns}}{Ns!}}{\sum_{j=0}^{Ns} \frac{(ofr.Ys)^j}{j!}} \ . \tag{3.15}$$

The number of switching lines *Ns* and the values of *ofr.Ys* are calculated by the conditions of the Theorem 1 and Theorem 2.

Remark 1-2: *ofr.Ys,* being evaluated on the base of the Theorem 1 and Theorem 2 for *adm.Pbs* and *adm.Pbr*, is resulted in a fixed value. Then *Pbs* = *Pbs* (*Ns,ofr.Ys*) is a function of *Ns* only and *Pbs* = *Pbs* (*Ns*).

Theorem 3: The function *Pbs* = *Pbs* (*Ns,ofr.Ys*), defined through (3.15) in the NDT is strictly monotone decreasing according to $Ns \geq 1$, when *ofr.Ys* >0 is a fixed value.

Proof: It can be proved that *Pbs* (*Ns*+1, *ofr.Ys*) < *Pbs* (*Ns, ofr.Ys*). Obviously (see (3.15) $Pbs(0, ofr.Ys) = 1$.

Using the recursion Erlangs'B - formula [Iversen 2004]:

$$Pbs(Ns, ofr.Ys) = \frac{ofr.Ys\, Pbs(Ns-1, ofr.Ys)}{Ns + Pbs(Ns-1, ofr.Ys)}. \quad (3.16)$$

But Erl_b(Ns,ofr.Ys) > 0 when ofr.Ys >0 and $Ns \geq 1$, $Ys\ Erl\_b(Ns,ofr.Ys)$+Ns+1> 0 and

$$Pbs(Ns+1, ofr.Ys) - Pbs(Ns, ofr.Ys) = Erl\_b(Ns+1, ofr.Ys) - Erl\_b(Ns, ofr.Ys) =$$
$$= Erl\_b(Ns, ofr.Ys)\frac{ofr.Ys\,[1 - Erl\_b(Ns, ofr.Ys)] - (Ns+1)}{ofr.Ys\ Erl\_b(Ns, ofr.Ys) + (Ns+1)} =$$
$$Erl\_b(Ns, ofr.Ys)\frac{crr.Ys - (Ns+1)}{ofr.Ys\ Erl\_b(Ns, ofr.Ys) + (Ns+1)}.$$

Because *crr.Ys* ≤ *Ns* follows *crr.Ys* – (*Ns*+1) < 0.

Therefore, *Pbs* (*Ns*+1, *ofr.Ys*) – *Pbs* (*Ns, ofr.Ys*) < 0 and the function *Pbs* = *Pbs* (*Ns,ofr.Ys*), defined through (3.14) is strictly monotone decreasing, when *ofr.Ys* > 0 is fixed value.

5. Analytical Solution

Based on the Assumption A-8 we are working with mean values of the parameters. Various techniques for analyzing complex teletraffic systems require a formulation of the Erlang function that is continuous in the parameter *Ns*. This is done via the integral representation [Berezner 1998].

Theorem 4: There is only one solution in the NDT through the equation

$$Erl\_b\ (Ns, ofr.Ys) = adm.Pbs, \quad (5.1)$$

according to the number of switching lines *Ns*.

Adm.Pbs ∈(0; 1] is in advance administratively determined value of blocking probability, providing of QoS.

Proof: Existence: It was proved, that the function *Pbs* = *Pbs* (*Ns,ofr.Ys*), defined through (3.14) in the NDT, is strictly monotonic decreasing, when *ofr.Ys* >0 is fixed value. The number 1 is absolute maximum and 0 is absolute minimum of the function. There is only one solution of the equation (3.15) for *adm. Pbs* ∈(0; 1], relying of the Intermediate Value Theorem (Dirschmidt, H. Yorg, 1992).

Uniqueness: Admitting that there are two different solutions *Ns1* ≠ *Ns2* of the equation (11.19) for *adm. Pbs* ∈(0; 1], therefore they are simultaneously fulfilled *Erl_b*(*Ns1, ofr.Ys*) = *adm.Pbs* and *Erl_b*(*Ns2, ofr.Ys*) = *adm.Pbs,* is contradicting to Theorem 3.

It is proved that only one solution of *Ns* exist, fulfilling the equation (3.15) and corresponding to the determined administratively in advance value of the blocking probability *adm.Pbs* ∈(0; 1].

6. Algorithm for Calculating the Values of the Parameters in the NDT:

1. From SLA and ITU-Recommendation are specified and determined administratively blocking probability *adm.Pbl*, respectively *adm. Pbs* ∈(0; 1] and *adm.Pbr* ∈[0; 1]:

$$\forall adm.Pbl \Rightarrow \exists (adm.Pbr, adm.Pbs):$$
$$adm.Pbr \oplus adm.Pbs = adm.Pbl : adm.Pbr \in [0,1], adm.Pbs \in (0,1] \quad (6.1)$$

2. The unknown parameters (3.3) in the NDT are evaluated on the base of Theorem 1 and Theorem 2, known parameters (3.2), especially *adm.ofr.Ys* (*adm.Pbr*, *adm.Pbs*).

3. On the basis of each calculated value *adm.ofr.Ys*, we evaluate

$$\widetilde{Ns} \in R_+ : \{\forall (adm.ofr.Ys, \widetilde{Ns}) :\ Pbs(adm.ofr.Ys, \widetilde{Ns}) = adm.Pbs\} \quad (6.2)$$

4. If $adm.Ns = \sup \widetilde{Ns}$, then

$$Ns = [adm.Ns] + 1 : \; Pbl \le adm.Pbl. \tag{6.3}$$

5. For finding of the number of internal switching lines *Ns,* a computer program is created on the base of the recursion Erlangs'B – formula (6.15) [Iversen 2004]. From numerical point of view, the following linear form is the most stable:

$$I(Ns, ofr.Ys) = 1 + \frac{Ns}{ofr.Ys} I(Ns-1, ofr.Ys), \qquad I(0, ofr.Ys) = 1, \tag{6.4}$$

where $I(Ns, ofr.Ys) = 1/Pbs(Ns, ofr.Ys)$. This recursion formula is exact, and for large values of (*Ns*, *ofr.Ys*) there are no round of errors.

6. The received results for numerical inversion of the Erlang's formula (for finding the number of switching lines *Ns*) were confirmed with results of others commercial computer programs.

Therefore, it is proved that if *Pbr* ≠ 0 and *Pbs* ≠ (*S*1 – *S*2 *Pbr*) / (*S*3 – *S*2 *Pbr*), then the NDT is solvable and there is proposed algorithm for its solution.

When *Pbr* = 0 the network loading is rather low and it is not of great practical interest, but in this case a mathematical research is made also.

7. Numerical Results

Among the easy computable QoS - parameters in the system (resulted from QoS- strategy of the network operators) is blocking probability *Pbl* in *pie-form model* [Poryazov 2000]. The sum of the loss probabilities due to abandoned and interrupted dialing, blocked and interrupted switching, not available intent terminal, blocked and abandoned ringing and abandoned conversation in *pie- form model* is 1.

For finding of the main teletraffic characteristics in proposed conceptual and its corresponding analytical model, the so called *normal-form model* (see Fig. 1) is used for presentation of blocking switching probability (*Pbs*) and probability of finding B-terminal busy (*Pbr*).

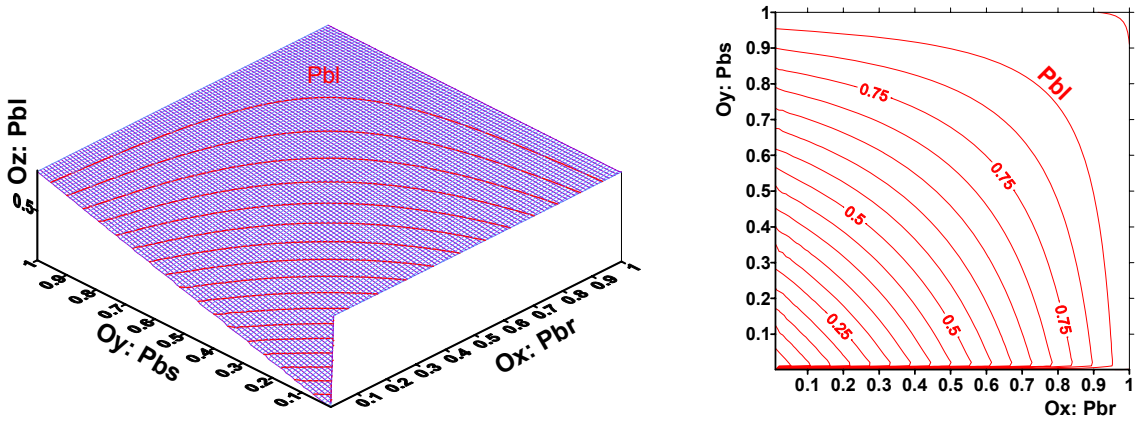


Fig. 2. General blocking probability *Pbl* in *pie-form model* is presented as function of probability of finding B-terminal busy *Pbr* and probability of blocking switching *Pbs* in *normal-form Pbr* and *Pbs* in 3D and contour - map presentation.

Based on the conceptual and its corresponding analytical model (1.1)- (2.9), defined general blocking probability *Pbl* is presented in *pie-form model* in (2.10) as function of the *Pbr* and *Pbs* (which are presented in *normal- form model* in the same equation).

On the Fig. 2 blocking probability *Pbl* is shown in *pie- form model*, depending on probability of finding B-terminal busy *Pbr* (Ox – axis) and probability of blocking switching *Pbs* (Oy – axis) in *normal- form model*. *Pbl* increases when *Pbr* and *Pbs* increase. Therefore, when *adm.Pbl* is predetermined as level of QoS administratively then *adm.Pbr* and *adm.Pbs* can be determined (evaluated) correspondingly.

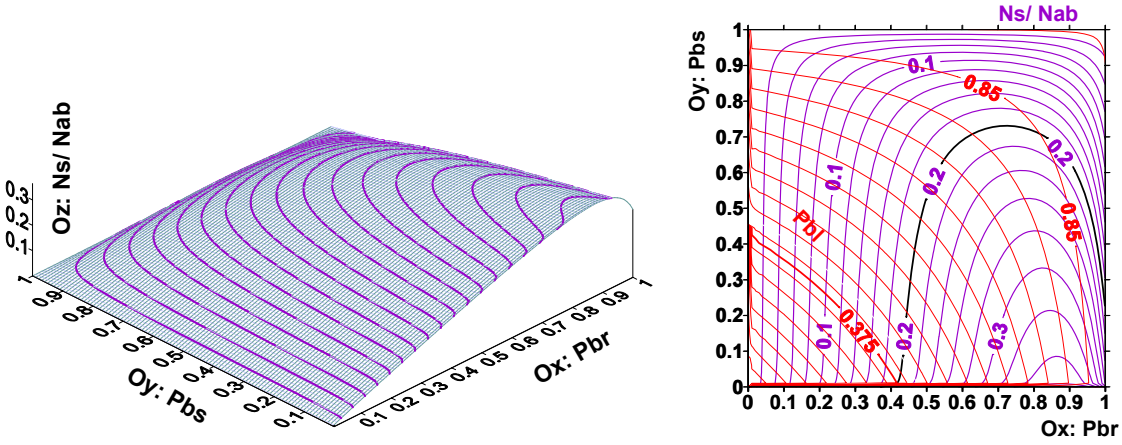


Fig. 3. The number of equivalent internal switching lines *Ns* (as percentage of number of active terminals *Nab*, where *Nab*= 7000 terminals) and general blocking probability *Pbl* are shown as functions of *Pbr* (Ox – axis) and *Pbs* (Oy – axis) in *normal-form model*, as well.

Conclusions of the numerical experiments:

According *Pbr*, *Pbs* and *Pbl*:

If $Pbr \in [0;1]$ and $Pbs \in [2\times10^{-9};0.999917]$ then

1. $Pbl \in [0;0.900896]$.
2. $0.000143 \le \frac{Ns}{Nab} \le 0.387857$, $Ns \in [1;2715]$, when $Nab = 7000$;
3. $0.77\times10^{-5} \le \frac{ofr.Ys}{Nab} \le 1.728311$, $ofr.Ys \in [0.473782;12098.18]$, when $Nab = 7000$.

   Ofr.Ys may exseed *Nab* by 73% approximately. This is “unproductiveness occupying of resources”.
4. Absolute maximum for *ofr.Ys*:

Maximum *ofr.Ys* = 12098.18 and this value is about 4.9 times greater than switching system capacity *Ns* = 2715.

Absolute maximum for *Ns*:

Ns = 2715 when *Nab*=7000 terminals, *Ns* = 38.79% of *Nab*. This is possible if *Pbl* = 0.900882 ≈ 90% (maximum theoretical value of *Pbl*), *Pbr*= 0.876623 ≈ 87.7% and *Pbs* = 9.28×10^{-9}, *Yab*= 6136.487 Erl ≈ 87.66 % of *Nab* and *ofr.Ys* = 2452.021 Erl = *crr.Ys* ≈ 35.0289 % of *Nab*.

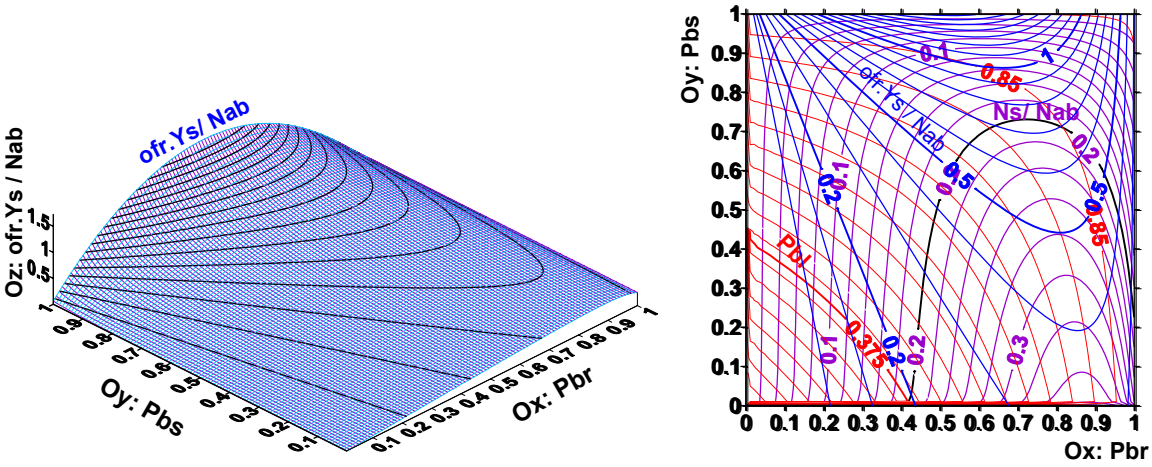


Fig. 4. The offered traffic *ofr.Ys*, the number of internal switching lines Ns and general blocking probability *Pbl* are presented as function of probability of finding B-terminal busy *Pbr* and probability of blocking switching *Pbs* in *normal – form model* in 3D and contour- map presentation.

8. Conclusions

1. Detailed normalized conceptual model, of an overall (virtual) circuit switching telecommunication system (like PSTN and GSM) is used. The model is relatively close to the real-life communication systems with homogeneous terminals.

2. General blocking probability *Pbl* as GoS parameter and *adm.Pbl* as QoS – parameter in *pie - form model* are formulated. The offered traffic *ofr.Ys*, the number of internal switching lines *Ns* and general blocking probability *Pbl* are derived as functions of probability of finding B-terminal busy *Pbr* and probability of blocking switching *Pbs* in *normal – form model*.

3. The network dimensioning task (NDT) is formulated on the base of preassigned values of QoS parameter *adm.Pbl* and its corresponding GoS - parameters - *adm.Pbr* and *adm.Pbs*; The NDT is formulated on condition that $Pbl \le adm.Pbl$.

4. The conditions for existence and uniqueness of a solution of the NDT are researched and an analytical solution of the NDT is found;

5. An algorithm and a computer program for a calculation the values of the offered (*ofr.Ys*), carried (*crr.Ys*) traffic and the number of equivalent switching lines *Ns*, are proposed. The results of numerical solution are derived and graphically shown;

6. The received results, in NDT, make the network dimensioning, based on QoS requirements easily;

7. The described approach is applicable directly for every (virtual) circuit switching telecommunication system (like GSM and PSTN) and may help considerably for ISDN, BISDN and most of core and access networks dimensioning. For packet switching systems, like Internet, proposed approach may be used as a comparison basis especially when they work in circuit switching mode.

References

Berezner S.A., 1998, - On the inverse of Erlang's Function- J. Appl. Prob.35, 246- 252

Dirschmidt, Hans Yorg, 1992 - Mathematische Grundlagen der Elektrotechnik-Braunschweig, Wiesbaden, pp.55

Engset, T., 1918. The Probability Calculation to Determine the Number of Switches in Automatic Telephone Exchanges. English translation by Mr. Eliot Jensen, Telektronikk, juni 1991, pp 1-5, ISSN 0085-7130. (Thore Olaus Engset (1865-1943). "Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wählerzahl in automatischen Fernsprechämtern", Elektrotechnische zeitschrift, 1918, Heft 31.)

ITU E.501. ITU-T Recommendation E.501: Estimation of traffic offered in the network. (Previously CCITT - Recommendation, revised 26. May 1997)

ITU E.600, ITU-T Recommendation E.600: Terms and Definitions of Traffic Engineering (Melbourne, 1988; revised at Helsinki, 1993).

ITU E.800. ITU-T Recommendation E.800: Terms and Definitions related to Quality of Service and Network Performance, including Dependability. (Helsinki, March 1-12, 1993, revised August 12, 1994).

Iversen V. B., 2004. Teletraffic Engineering and Network Planning, Technical University of Denmark, pp.125

Little J. D. C., 1961. A Proof of the Queueing Formula L=λW. Operations Research, 9, 1961, 383-387.

Poryazov S. A, Saranova E. T., 2002. On the Minimal Traffic Measurements for Determining the Number of Used Terminals in Telecommunication Systems with Channel Switching. In: "Modeling And Simulation Environment for Satellite and Terrestrial Communication Networks - Proceedings of the European COST Telecommunications Symposium", Kluwer Academic Publishers, 2002, pp. 135-144;

Poryazov S. A, Saranova E. T. 2005. Some General Terminal and Network Teletraffic Equations in Virtual Circuit Switching Systems. Symposium "Modeling and Simulation Tools for Emerging Telecommunications Networks: Needs, Trends, Challenges, Solutions", Munich, Germany, 8 - 9 September 2005, Institut für Technische Informatik, Universität der Bundeswehr München (in printing in LNCS, Springer).

Poryazov S. A. 2005a. Can Terminal Teletraffic Theory Be Liberated from the Main Illusions? In: Proceedings of the International Workshop "Distributed Computer and Communication Networks", Sofia, Bulgaria 24-25 April, 2005, Editors: V. Vishnevski and Hr. Daskalova, Technosphera publisher, Moscow, Russia, 2005, ISBN 5-94836-048-2, pp. 126-138.; COST-285 TD/285/05/04; COST 290 TD(05)009.

Poryazov S. A. 2005b. What is Offered Traffic in a Real Telecommunication Network? COST 285 TD/285/05/05; 19th International Teletraffic Congress, Beijing, China, August 29- September 2, 2005, accepted paper No 32-104A.

Poryazov S. 1991. Determination of the Probability of Finding B-Subscriber Busy in Telephone Systems with Decentralized Control. Comptes Rendus de l'Academie Bulgare des Sciences – Sofia, 1991, Tome 44, No.3, pp. 37-39.

Poryazov S. A. 2000. Потребительский и терминальный телетрафик- классификация и обозначения , Conference "Telecom'2000" - Varna, Bulgaria, October 2000 – pp. 58-59.

Poryazov S. A. 2001. On the Two Basic Structures of the Teletraffic Models. Conference "Telecom'2001" - Varna, Bulgaria, 10-12 October 2001 – pp. 435-450.

Poryazov, S. A. 2004. The B-Terminal Busy Probability Prediction. IJ Information Theories & Applications, Vol.11/2004, Number 4, pp. 409-415;

Author's Information

Emiliya Saranova – Bulgarian Academy of Science, Sofia, Bulgaria;
High College of Telecommunication and Posts, Sofia, Bulgaria;

e-mail: saranova@hctp.acad.bg

IMPLICATIONS OF RECENT TRENDS IN TELECOMMUNICATIONS ON MODELING AND SIMULATION FOR THE TELECOMMUNICATION INDUSTRY

Gerta Köster, Stoyan Poryazov

Abstract: *With this paper we would like to trigger a discussion on future needs of modeling and simulation techniques and tools for the telecommunication industry. We claim that the telecommunication market has undergone severe changes that affect the need for and type of simulations in industrial research. We suggest some approaches how to address these new challenges. We believe that there is need for intensive research in the area.*

Keywords: *Telecommunications, Market Evolution, Modeling and Simulation Challenges.*

Introduction

Models and simulations in telecommunications fulfill a number of tasks. We build models and run simulations to check the design of emerging products, we model the environment in which such a product is employed and we evaluate and optimize the performance of telecommunication equipment.

In the telecommunications industry the ultimate goals behind these tasks is to save costs and to make money by better or faster design and efficient development support.

We claim that the world of telecommunication has lately undergone severe changes that affect the need for and type of simulations in industrial research. In the course of this paper, we will outline the main trends we observe and discuss the implications on modeling and simulations.

1. Fractalization of the "old" incumbent telecommunication market.

2. Shift of operators' interest from providing a network to providing services and applications.

3. Shortening of development cycles.

4. Telecommunication and telecommunication problems pervade other sectors of private and business life.

Evolution of the Telecommunication Market

2.1. Fractal markets: We observe that a large part of the market has become fractal as opposed to monolithic in the past. The large state owned companies have been replaced by private enterprises. This is true for both, the Western industries, where the telecommunication business has been privatized and the former Soviet block where telecommunication has been denationalized.

This means that a large number of companies are producing and offering new - often small - products, services and devices, among which the customers, operators and consumers, may choose. Often these products complement each other enhancing each other's value. This means that there are no longer single solutions out of one hand.

An example might be the offering of ring-tones that enlivens the world of mobile communications as a complementary business. Another one is the home entertainment sector where networks are closely intertwined with the services, such as TV over broadband, that are offered through the networks and even the content that is offered.

2.2. Enlarged value chain: Operators and manufacturers seek to get a bigger portion of the value chain by offering applications and services or at least middleware to enable applications and services. This can be observed, for example, in the home entertainment market. Broadband internet access opens a new channel to the end consumers of content, e.g. movies, and thus offers an opportunity for direct cooperation between content

manufacturers and operators such as the big national telecommunication providers. We may, at some point, see a merger of network provisioning and services.

2.3. Shortened development cycles: Industrial research is always coupled with a certain product that the company wishes to launch. When products become smaller and follow, to a certain extend, the fashion of the day, development cycles shorten and design becomes more volatile. In industry, we need to cope with these shortened development cycles. Otherwise people may cease to use simulations as a decision basis. As a matter of fact, we have the impression that there is already a decline in the use of simulations. We think, that there is a need for "rapid modeling", may be even accepting a degradation of simulation quality.

2.4. Telecommunication pervades other industrial and private sectors: Telecommunication, at the very beginning, dealt with the communication between two people at some distance. Today not only people communicate, but people with devices and devices with devices. We see, for example, that the control units in a truck (for motor, breaks, gearing) exchange information across a mini communication system. Tags are attached to new cloths that communicate with the security system of a store through a small radio system. Infotainment systems in cars communicate with some traffic control network outside the car or with other cars.

There are many more examples. But do we carry over the knowledge of how to build a good telecommunication system to these areas?

3. State of the Modeling and Simulation Art in Industry

3.1. Adaptive development processes: In “old” product design, requirements were supposed to be complete and clear before product development began. With today’s explorative technology and rapidly changing markets (e.g. in telecommunications today), modeling and simulation targets are difficult to fix. A new development process must enable quick requirement adaptation, often without sufficient prior modeling and simulation. Some of the changes in the telecommunication market are discussed in Andersen (2002).

3.2. Volatility of modeling requirements: The lack of stable modeling and simulation requirements makes it difficult to validate and verify product models. Product "verification" usually refers to testing whether a product meets certain specifications. "Validation" seeks to ensure that the customer is satisfied and that the correct specifications were incorporated into the product.

3.3. Availability of software tools: The design and manufacturing processes may be presented as activities, both series and parallel, at several levels of detail over time during the development of a product. As depicted in NRC (2004), software tools are not available for 60% of the required product development activities. For other activities software tools may be emerging or even commonly available. However they rarely interoperate and their use is often inefficient.

3.4. Tool interoperability: Only when tools are available and fully interoperable, designers and engineers can use and link various data and models for a given activity as well as across different activities required for product design and realization. It yet needs to be demonstrated whether modeling and simulations tools can be integrated across multiple domains including geometric modeling, performance analysis, life-cycle analysis, cost analysis, and manufacturing.

3.5. Need for econometric evaluation of modeling and simulation: The critical decisions in the design and manufacturing of a product are taken in the early stages of its life cycle, based on the products’ modeling and simulation. But the economic usefulness of product modeling and simulation is still difficult to judge. We need methods and tools for the econometric evaluation of product modeling and simulation (Sargent et al 2000). The usefulness of models is also discussed in (Andersen 2002) also.

3.6. Multilayer modeling merging the product, management and market levels: Modeling and simulation play an important role on the product design level, on the management level (Ericsson AB 2005) and on the market level (Andersen 2002). Merging these levels implies multilayer modeling, incorporating different paradigms, languages

and methods. Such a methodology is emerging in physics (Fishwick et al. 1992) but a modeling methodology that integrates the languages of physics, management and economics is still a matter of the future.

4. Challenges and Conclusion

On the one hand, the incumbent telecommunication market seems to undergo severe changes towards a fast-living volatile world and on the other hand "classic" telecommunications modeling and simulation problems pop up in new and unexpected areas of life.

We think that modeling and simulation can respond to these changes by tackling the following challenges:

- Develop new paradigms for modeling and simulating emerging New Generation Networks, applications and services that take into account the volatility of product requirements.
- Integrate dynamic models with different levels of abstraction, different paradigms, and different modeling languages.
- Development of meta-modeling methodology and tools to evaluate the technical and econometric usefulness of modeling and simulation in every stage of product life cycle.
- Establish sets of clear criteria for modeling and simulation needs for industrial applications, such as the necessary levels of sensitivity and accuracy and the ability to adapt to requirement changes.
- Establish suitable methods and criteria to verify, validate, and certify model trustworthiness for emerging systems, devices and their environments, especially when we develop adaptable methods.
- Increase the reusability of model components as well as of data of the real systems measurements and of simulation results.
- Integrate the methods and tools for design, creation, management and control of telecommunication products and systems (for a preliminary approach see Caughlin 2000).

5. References

Andersen 2002. Outlook of the development of technologies and markets for the European Audio-visual sector up to 2010. http://europa.eu.int/comm/avpolicy/stat/tvoutlook/tvoutlook_finalreport.pdf

Caughlin 2000. Don Caughlin. An Integrated Approach to Verification, Validation, and Accredition of Models and Simulations. Proceedings of the 2000 Winter Simulation Conference, J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

Ericsson AB 2005. Enhancing Telecom Management. White Paper. http://www.ericsson.com/technology/whitepapers/telecom_management.pdf

Fishwick et al. 1992. Paul A. Fishwick and Bernard P. Zeigler. A Multimodel Methodology for Qualitative Model Engineering. ACM Transactions on Modeling and Computer Simulation, 2(1):52- 81, 1992.

NRC 2004. National Research Council of the National Academy of Sciences, 2004. Retooling Manufacturing: Bridging Design, Materials, and Production. http://www.nap.edu/catalog/11049.html)

Sargent et al 2000. Robert G. Sargent, Priscilla A. Glasow, Jack P.C. Kleijnen, Averill M. Law, Ian McGregor, Simone Youngblood. Strategic directions in verification, validation, and accreditation research. http://citeseer.ist.psu.edu/cache/papers/cs/17452/http:zSzzSzcwis.kub.nlzSz~few5zSzcenterzSzstaffzSzkleijnenzSzsargentr.pdf/strategic-directions-in-verification.pdf (Access 23.05.2006)

Authors' Information

Gerta Köster – Siemens AG, e-mail: gerta.koester@siemens.com

Stoyan Poryazov – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, e-mail: stoyan@cc.bas.bg

ОПТИМИЗАЦИЯ СТРУКТУРЫ СЕТЕЙ С ТЕХНОЛОГИЕЙ MPLS ПО ОГРАНИЧЕНИЯМ НА ПОКАЗАТЕЛИ ЖИВУЧЕСТИ

Юрий Зайченко, Мухаммедреза Моссавари

***Аннотация**: В докладе сформулирована проблема оптимизации структуры сетей с технологией MPLS по ограничениям на заданный уровень показателей живучести. Построена математическая модель задачи и предложен алгоритм ее решения, позволяющий оптимизировать сеть по критерию стоимости при ограничениях на установленные показатели живучести для различных категорий сервиса.*

***Abstract**: The problem of MPLS computer network structure optimization under constraint on survivability is considered in this paper. The mathematical model is built and corresponding algorithm is suggested allowing to optimize network structure by cost criterion under constraints on the established level of survivability for different classes of service.*

Введение

По мере значительного расширения мультимедийных приложений и телеконференций, требующих все более высоких скоростей передачи и более широкой полосы пропускания возникла необходимость в создании новой технологии, которая бы предоставила единый транспортный механизм поверх самых разнообразных технологий, таких как Ethernet, Frame Relay, ATM, SONET и обеспечивала высокоскоростную передачу информации с требуемым качеством обслуживания (QoS). Известные коммуникационные технологии такие, как IP, Ethernet, Token Ring не позволяют обеспечить заданное качество обслуживания и требуемый уровень QoS. Такой технологией является технология многопротокольной коммутации меток MPLS (Multiprotocol Label Switching), которая пришла на смену технологии ATM.

MPLS является универсальным решением проблем обеспечения заданного уровня QoS, стоящих перед современными сетевыми технологиями, она обеспечивает высокую скорость передачи, масштабируемость, контроль, оптимизацию распределения трафика, а также маршрутизацию [1, 2].

Одной из важных задач, которые приходится решать в процессе проектирования сетей с технологией MPLS является задача обеспечения заданного уровня ее живучести. В работе [4] была рассмотрена задача анализа живучести сетей MPLS, введены показатели живучести сетей и предложен алгоритм их оценки для различных классов потоков. Целью настоящей работы является решение задачи оптимизации сетей с технологией MPLS по установленным ограничениям на показатели живучести сети по критерию минимизации стоимости.

Постановка и математическая модель задачи

Задано множество узлов сети $X = \{x_j\}$ $j = \overline{1,n}$ - маршрутизаторов MPLS (так называемых LRS – Label Switching Routers), их размещение по территории региона, набор пропускных способностей каналов связи $D = \{d_1, d_2, ..., d_k\}$ из которых ведется синтез и их удельных стоимостей на длины $C = \{c_1, c_2, ..., c_k\}$, определены классы обслуживания CoS (Class of Service), известны матрицы входящих требований для *k*-го класса $H(k) = \|h_{ij}(k)\|$ $i, j = \overline{1,n}$; $k = 1, 2, ..., K$, где $h_{ij}(k)$ – интенсивность *k*-го класса, который необходимо передавать из узла *i* в узел *j* в единицу времени (*Кбит/с*).

Кроме того, введены ограничения на показатели качества QoS для каждого класса *k* в виде ограничения на среднюю задержку $T_{зад,k}$, $k=\overline{1,K}$ и установлены требования на уровни показателей живучести каждого класса $P_{k,зад}[y]$.

Требуется найти структуру сети в виде набора каналов связи (КС) $E=\{(r,s)\}$, выбрать пропускные способности (ПК) каналов связи $\{\mu_{rs}\}$, таким образом, чтобы обеспечить передачу требований всех классов $H(k)$ в полном объеме и при этом бы выполнялись ограничения на установленные показатели живучести, а стоимость сети была бы минимальной.

Составим математическую модель данной задачи синтеза.

Требуется найти:

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

при следующих ограничениях на показатели живучести

$$\begin{cases} P\left\{H_{\Sigma}^{\Phi}(1)\geq aH_{\Sigma зад}(1)\right\}\geq P_{1,зад} \\ P\left\{H_{\Sigma}^{\Phi}(2)\geq aH_{\Sigma зад}(2)\right\}\geq P_{2,зад} \\ ----------- \\ P\left\{H_{\Sigma}^{\Phi}(k)\geq a\%H_{\Sigma зад}(k)\right\}\geq P_{r,зад} \end{cases} \tag{2}$$

где $H_{\Sigma}^{\Phi}(k)$ - фактическая величина потока к-го класса, передаваемого в сети при отказах; ее элементов (каналов и узлов связи).,

$H_{\Sigma}^{(0)}$ - величина номинального потока к-го класса,в безотказном состоянии.

В работе [5] для потока *k*-го класса; при условии что обслуживание в классах происходит с относительными приоритетами ρ_k в порядке убывания номера класса (т.е. $\rho_1>\rho_2>\ldots>\rho_k$) при заданном наборе ПС каналов $\{\mu_{rs}\}$ и распределении потоков (РП) – $F(k)=[f_{rs}^{(k)}]$ было получено следующее выражение для средней задержки $T_{ср,k}$:

$$T_{ср,k}(\{\mu_{rs}\},F)=\frac{1}{H_{\Sigma}^{(k)}}\sum_{(r,s)\in E}\frac{f_{rs}^{(k)}\sum_{i=1}^{K}f_{rs}^{(i)}}{(\mu_{rs}-\sum_{i=1}^{K-1}f_{rs}^{(i)})(\mu_{rs}-\sum_{i=1}^{K}f_{rs}^{(i)})} \tag{3}$$

при условии, что суммарная величина потока в канале (r,s) удовлетворяет условию $\sum_{i=1}^{K}f_{rs}^{(i)}=f_{rs}<\mu_{rs}$

где $f_{rs}^{(i)}$ - величина потока класса i в КС (r,s).

В работе [4] была рассмотрена задача анализа живучести компьютерных сетей с технологией MPLS,, введены показатели живучести и предложен алгоритм оценки показателей живучести (ПЖ) сетей для разных классов сервиса при отказах её элементов – каналов и узлов связи (КС).

В данной работе ставится задача оптимизации структуры сетей по заданным значениям ПЖ.

Описание алгоритма оптимизации структуры при ограничениях на заданные уровни ПЖ

Пусть одним из известных методов, например предложенным в работе [3] получена базовая структура сети E_1 в виде набора каналов (r,s) заданной пропускной способности, обеспечивающих передачу всех категорий требований в полном объеме в безотказовом состоянии.

Цель алгоритма – оптимизация структуры при ограничениях на заданные уровни ПЖ $P_{k,зад}$.приотказах ее элементов.

Перед началом этапа (0 шаг) рассчитываем начальные значения ПЖ для всех классов: требований

$$P\left\{H_{\Sigma}^{\Phi}(k) \geq a\% H_{\Sigma}^{0}\right\} \quad k = \overline{1,K} \quad a \in \left[50 \div 100\right]\%$$

Для этого используем алгоритм оценки ПЖ, предложенный в [4].

Далее проверяем выполнение ограничений:

$$P\left\{H_{\Sigma}^{\Phi}(k) \geq a\% H_{\Sigma}^{0}\right\} \geq P_{k,зад}^{a} \tag{4}$$

Если условие (4) выполняются для всех классов $k = \overline{1,K}$ и всех уровней $a \in \left[50 \div 100\right]\%$, то конец работы алгоритма, иначе переход на 1-ю итерацию.

Цель каждой итерации состоит в повышении текущих значений ПЖ, путем резервирования соответствующих каналов и узлов.

Пусть заданы надежностные характеристики каналов (КС) и узлов связи (УС) – коэффициенты готовности КС (r,s), $k_{\Gamma rs}$ - коэффициент готовности i-го УС $k_{\Gamma i} \quad i = \overline{1,n}$.

Пусть z_i - отказовое состояние, состоящее в отказе КС (r_i, s_i).

Тогда вероятность его появления определится согласно [4] так

$$P(z_i) = (1 - k_{\Gamma r_i s_i}) \prod_{(r,s) \neq (r_i, s_i)} k_{\Gamma rs} \prod_{i=1}^{n} k_{\Gamma_i} \tag{5}$$

Допустим, что мы резервируем КС (r_i, s_i). Тогда вероятность состояния z_i после резервирования определяется так:

$$P_{рез}(z_i) = (1 - k_{\Gamma r_i s_i})^2 \prod_{(r,s) \neq (r_i, s_i)} k_{\Gamma rs} \prod_{i=1}^{n} k_{\Gamma_i} \tag{6}$$

Изменение этой величины равно

$$\Delta P(z_0) = P_{рез}(z_i) - P(z_i) = -k_{\Gamma r_i s_i}(1 - k_{\Gamma r_i s_i}) \prod_{(r,s) \neq (r_i, s_i)} k_{\Gamma rs} \prod_{i=1}^{n} k_{\Gamma_i} = -k_{\Gamma r_i s_i} P(z_i) \tag{7}$$

Будем оценивать эффект от резервирования КС (r_i, s_i) так

$$\alpha_{r_i s_i} = -\frac{\Delta P(z_i)}{C_{рез}(r_i, s_i)} \tag{8}$$

где $C_{рез}(r_i, s_i)$ - стоимость резервирования КС (r_i, s_i).

Разбиваем все множество отказовых состояний на подмножества $Z_1 = \{z_i\}$, $Z_2 = \{z_j\}$, $Z_3 = \{z_s\}$, $Z_4 = \{z_r\}$ и $Z_5 = \{z_t\}$, где $z_i \in Z_1$, если максимальный поток в состоянии z_i $H_\Sigma(z_i) < 90\% H_\Sigma^0$; $z_j \in Z_2$, если $H_\Sigma(z_j) < 80\% H_\Sigma^0$; $z_s \in Z_3$, если $H_\Sigma(z_s) < 70\% H_\Sigma^0$; $z_r \in Z_4$, если $H_\Sigma(z_r) < 60\% H_\Sigma^0$; и наконец $z_t \in Z_5$, если $H_\Sigma(z_t) < 50\% H_\Sigma^0$.

Очевидно между данными подмножествами имеется следующее отношение включения: $Z_5 \subseteq Z_4 \subseteq Z_3 \subseteq Z_2 \subseteq Z_1$.

Поэтому для резервирования выбираем в первую очередь состояния $z_k \in Z_5$ (т.е. наибольшие критические отказовые состояния).

1-я итерация

Рассматриваем состояние $z_i \in Z_5$, представляющее отказ КС (r_i, s_i).

1. Для всех КС (r_i, s_i) вычисляем эффект от резервирования $\alpha_{r_i s_i}$ согласно (8).

2. Выбираем КС (r_i^*, s_i^*) с максимальным показателем эффективности.

$$\alpha_{r_i^* s_i^*} = \max \alpha_{r_i s_i} \tag{9}$$

3. Резервируем КС (r_i^*, s_i^*) и пересчитываем ПЖ сети после резервирования

$$P^{(н)}\left\{H_\Sigma^{Ф}(k) \geq a\% H_\Sigma^0(k)\right\} = P\left\{H_\Sigma^{Ф}(k) \geq a\% H_\Sigma^0(k)\right\} + \left|\Delta P(z_i^*)\right| \quad \text{для всех } k = \overline{1, K} \tag{10}$$

где z_i^* - состояние отказа КС (r_i^*, s_i^*).

4. Проверка ограничений на показатели живучести:

$$P^{(н)}\left\{H_\Sigma^{Ф}(k) \geq a\% H_\Sigma^0(k)\right\} \geq P_{k,зад} \quad \text{для всех } k = \overline{1, K} \tag{11}$$

Если условия (11) выполняются для всех К, то STOP конец работы алгоритма, иначе переход к следующей итерации. Указанные итерации повторяем до тех пор, пока условия (11) не начнут выполнятся для всех К.

Конец работы алгоритма.

В докладе приводятся результаты экспериментальных исследований разработанного алгоритма и его применение для оптимизации структуры глобальной компьютерной сети с технологией MPLS.

Заключение

1.В работе сформулирована задача структурной оптимизации сетей MPLS по ограничениям на показатели живучести сети.

2. Предложен алгоритм структурного синтеза сетей MPLS, позволяющий оптимизировать топологию сети при ограничениях на заданные показатели живучести.

Литература

1.Гольдштейн А. Б., Гольдштейн Б. С. Технология и протоколы MPLS. СПб.: БХВ. Санкт-Петербург, 2005. 304 с.

2.Олвейн Вивьен. Структура и реализация современной технологии MPLS. Перевод с английского. Изд. дом «Вильямс», 2004. 480 с.

3.Зайченко Е. Ю. Сети АТМ: Моделирование, анализ и оптимизация. Киев, 2003. 216 с.

4.Зайченко Ю.П., Мухаммедреза Моссавари. Анализ показателей живучести компьютерной сети с технологией MPLS. Вісник національного технічного університету «КПІ». Сер. Інформатика, управління та обчислювальна техніка. N43 ВЕК+ 2005. с. 73 - 80.

5.Зайченко Ю.П., Шарадка Ахмед. Задача распределения потоков различных классов в сети с технологией MPLS. Вісник національного технічного університету «КПІ». Інформатики, управління та обчислювальна техніка №43. 2005. с.113 –123

Authors' Information

Зайченко Юрий Петрович – НТУУ Киевский политехнический институт, профессор, Киев-56, Проспект Победы 37, тел 38-044-2418693, e-mail: zaych@i.com.ua

Мухаммедреза Моссавари (Иран) – НТУУ Киевский политехнический институт, аспирант, Киев-56, Проспект Победы 37, тел.8-067-7099053, e-mail: olgamax@mmsa.ntu-kpi.kiev.ua

ОПТИМИЗАЦИЯ ХАРАКТЕРИСТИК СЕТЕЙ MPLS ПРИ ОГРАНИЧЕНИЯХ НА ЗАДАННЫЕ ПОКАЗАТЕЛИ КАЧЕСТВА ОБСЛУЖИВАНИЯ

Юрий П. Зайченко, Ахмед А. М. Шарадка

***Аннотация**: В докладе сформулирована задача оптимизации характеристик сетей с технологией MPLS, включающая оптимальный выбор пропускных способностей и распределение потоков (ВПС РП). Предложен алгоритм ВПС РП, позволяющий оптимизировать пропускные способности каналов связи и найти распределение потоков всех категорий по критерию минимизации стоимости сети при ограничениях на установленные значения показателей качества обслуживания. Приводятся результаты экспериментальных исследований разработанного алгоритма.*

***Abstract**: The problem of MPLS networks analysis and optimization – including optimal capacities assignment and flows distribution is considered in this paper. The corresponding algorithm of its solution is suggested enabling to choose optimal channel capacities and distribute flows of all classes minizing the network cost under constraints on quality of service (QoS). The experimental investigations of the suggested algorithm are presented.*

***Keywords:** MPLS networks, optimization, capacities assignment, flows distribution.*

Введение

К современным телекоммуникационным (сетевым) технологиям предъявляются требования передачи разных видов информации (аудио, видео и данных) по общим каналам связи с помощью унифицированного транспортного механизма и обеспечения при этом заданного качества обслуживания (Quality of Service) – а именно средней задержки $T_{ср}$, и её вариации. Существующие сетевые технологии такие, как IP, Ethernet, Frame Relay, Token Ring не в состоянии обеспечить требуемое качество обслуживания. Технологий ATM, которая была разработана для решения указанной проблемы качества, оказалась дорогостоящей и не смогла выдержать конкуренции с технологией Gigabit Ethernet и IP.

Поэтому на смену ей в конце 90-х годов была создана новая технология многопротокольной коммутации меток (Multiprotocol Label Switching-MPLS). Её отличительными способностями являются: 1) введение различных категорий потоков классов обслуживания (Class of Service); 2) возможность обеспечения заданного качества обслуживания QoS для разных категорий; 3) предоставление единого транспортного механизма для передачи разных видов информации и наконец возможность работы с различными сетевыми технологиями и протоколами (Frame Relay, Ethernet, IP, ATM). [1]

Важными задачами которые приходится решать в процессе построения сетей MPLS являются задача анализа и оптимизации их характеристик, и в частности, оптимальный выбор пропускных способностей каналов связи и распределение потоков различных классов по каналам (РП).

В работе [2] была рассмотрена задача оптимального выбора ПС каналов связи (ВПС) при ограничениях на установленные значения показателей качества обслуживания (QoS), а именно средней задержки для различных классов потоков и описан алгоритм её решения.

В работе [3] исследована задача оптимального распределения потоков различных каналов в сети MPLS при ограничениях на показатели QoS и предложен соответствующий алгоритм РП. Целью настоящей работы является обобщение полученных результатов, постановка и формализация комбинированной задачи ВПС РП, и разработка алгоритма её решения.

Постановка и математическая модель задачи

Задана сеть MPLS со структурой $G = (X, E)$, где $X = \{x_j\}$ $j = \overline{1,n}$ – множество узлов сети (УС); $E = \{(r,s)\}$ – множество КС; задан набор пропускных способностей $D = \{d_1, d_2, ..., d_K\}$ и их удельных стоимостей $C = \{c_1, c_2, ..., c_K\}$.

Определено число классов потоков $k = \overline{1,K}$, заданы матрицы требований входящих потоков $H(k) = \|h_{ij}(k)\|$ $i, j = \overline{1,n}$, где $h_{ij}(k)$ – интенсивность потока, который необходимо передавать из УС x_i в x_j (Мбит/с). Установлены требования по обеспечению заданного показателя качества (QoS) – $T_{ср,k}$. Требуется выбрать такие ПС всех КС $\{\mu_{rs}^0\}$ и найти распределение потоков всех классов $F(k)[f_{rs}(k)]$ при которых стоимость сети будет минимальна, а средняя задержка для каждого класса не будет превышать заданную величину $T_{ср,зад}$.

Составим математическую модель данной задачи.

Требуется найти $$\min C_{\Sigma} = \sum_{(r,s)\in E} c_{rs}(\mu_{rs}) \quad (1)$$

при условиях

$$T_{cp}^{(k)}(\{\mu_{rs}\}) = \frac{1}{H_{\Sigma}^{(k)}} \sum_{(r,s)\in E} \frac{f_{rs}^{(k)} \cdot \sum_{i=1}^{k} f_{rs}^{(i)}}{\left(\mu_{rs} - \sum_{i=1}^{k-1} f_{rs}^{(i)}\right)\cdot\left(\mu_{rs} - \sum_{i=1}^{k} f_{rs}^{(i)}\right)} \le T_{зад,к} \quad (2)$$

$$\sum_{i=1}^{k} f_{rs}^{(i)} < \mu_{rs} \text{ для всех } (r,s) \in E \quad k = \overline{1,k} \quad (3)$$

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

где $f_{rs}^{(i)}$ - многопродуктовый поток i-го класса, протекающий по КС (r,s), предполагается что обслуживание потоков в узлах сети (коммутаторах) с относительными приоритетами, приоритеты ρ_i убывают с ростом номера класса, т.е. $\rho_1 > \rho_2 > ... > \rho_k$.

Описание алгоритма решения

Алгоритм состоит из предварительного этапа и конечного числа однотипных итераций.

На предварительном этапе находятся начальные ПС каналов связи ($\{\mu_{rs}(0)\}$) и начальное распределение потоков всех классов $F_k(0)$, $k = \overline{1,k}$.

Цель последующих итераций – оптимизация ВПС и РП по критерию стоимости.

Пусть проведено *r* итераций и найдены ПС $\{\mu_{rs}(r)\}$, распределение потоков $F(k)\left[f_{rs}(r)\right]$ и стоимость сети $C_{\Sigma}(r)$.

(r+1)^{я}-итерация

1. Решаем задачу РП по критерию $\min T_{cp1}$ используя алгоритм предложенный в [1] и находим $F_k(r+1)$.

2. Решаем задачу ВПС по критерию $\min C_{\Sigma}$ для РП $F_k(r+1)$ при ограничениях $T_{cp}(k) \le T_{зад,k}$ и находим $\{\mu_{rs}(r+1)\}$.

3. Вычисляем величину критерия $C_{\Sigma}(r+1)$ и проверка условия $[C_{\Sigma}(r) - C_{\Sigma}(r+1)] < \varepsilon$. Если да, то STOP распределение потоков $F_k(r+1)$ и ПС $\{\mu_{rs}(r+1)\}$ – искомые, иначе $r = r+1$ и на шаг 1 следующей итерации.

Экспериментальные исследования алгоритма ВПС РП

Предложенный алгоритм был реализован программно и были проведены эксперименты по анализу его эффективности.

В экспериментах рассматривалась сеть с $n = 10$ узлов и $m = 16$ каналов связи. Структура сети задавалась случайным набором каналов связи:

{(1,6); (1,9); (2,3); (2,4); (2,6); (3,4); (4,5); (4,6); (4,7); (5,6); (5,10); (6,9); (7,8); (7,10); (8,9); (9;10)}.

Число классов сервиса к=6.

Набор базовых ПС каналов связи Д задается в таблице 1, а их удельная стоимость – в таблице 2.

Таблица 1 Набор пропускных способностей

ТИП КС	d_1	d_2	d_3	d_4	d_5	d_6	d_7
ПС Кбит/с	8000	10000	15000	20000	22000	25000	30000

Таблица 2 Удельная стоимость канала (на ед. длины)

	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Удельная стоимость (в у.е)	800	1000	1500	2000	2200	2500	3000

Были заданы матрицы требований по всем классам сервиса H_1, H_2, H_3, H_4, H_5 и H_6.

Суммарный объем исходящих требований по классам сервиса составляет (в Кбит/с): $H_{\Sigma}^{(1)} = 132$, $H_{\Sigma}^{(2)} = 1492$, $H_{\Sigma}^{(3)} = 4936$, $H_{\Sigma}^{(4)} = 7362$, $H_{\Sigma}^{(5)} = 12126$, $H_{\Sigma}^{(6)} = 18176$.

Заданы следующие ограничения по средним задержкам $T_{зад,к}$ по всем классам сервиса, которые приведены в таблице 3.

Таблица 3

к	1	2	3	4	5	6
$T^{(k)}_{зад,к}$	0,0002	0,0003	0,0005	0,0008	0,001	0,01

В результате решения задачи ВПС РП были найдены искомые ПС всех каналов связи и распределение потоков всех категорий $F(k)$ $k = \overline{1,6}$, а также общая стоимость сети: C_{Σ}=48300 у.е.

Искомые ПС представлены в таблице 5, а суммарные распределения потоков в таблице 4.

Таблица 4 Суммарное распределение потоков всех классов $F(k)$

0	0	0	0	0	6594,75	0	0	3246	0
0	0	5484	3006	0	9675,75	0	0	0	0
0	5484	0	4461	0	0	0	0	0	0
0	3006	4461	0	8571	4503	0	0	0	0
0	0	0	8571	0	8702,25	0	0	0	12070,5
6594,75	9675,75	0	4503	8702,25	0	0	0	11127,75	0
0	0	0	0	0	0	0	3369,75	0	9004,5
0	0	0	0	0	0	3369,75	0	8568	0
3246	0	0	0	0	11127,75	0	8568	0	2325
0	0	0	0	12070,5	0	9004,5	0	2325	0

Таблица 5 Оптимальные пропускные способности каналов связи (Кбит/с)

0	0	0	0	0	10000	0	0	10000	0
0	0	10000	10000	0	15000	0	0	0	0
0	10000	0	10000	0	0	0	0	0	0
0	10000	10000	0	20000	15000	0	0	0	0
0	0	0	20000	0	20000	0	0	0	22000
15000	20000	0	15000	20000	0	0	0	220000	0
0	0	0	0	0	0	0	15000	0	20000
0	0	0	0	0	0	15000	0	20000	0
15000	0	0	0	0	22000	0	20000	0	15000
0	0	0	0	22000	0	20000	0	15000	0

В последующих экспериментах изменялись матрицы требований $H(k)$ путем умножения на коэффициент r_i $H_i(k) = r_i \cdot H(k)$, где i – номер эксперимента и решалась задача ВПС РП в результате чего была найдена зависимость $C_\Sigma = f(r)$. Соответствующие результаты приводятся в таблице 6.

Таблица 6

r_i	1	1,25	1,5	2	2,5
C_Σ	48300	53200	59400	65000	72000

Заключение

1.В работе сформулирована задача оптимального выбора пропускных способностей и распределения потоков (ВПСРП) для сетей с технологией MPLS при ограничениях на среднюю задержку и построена её математическая модель.

2. Предложен алгоритм ВПСРП для сетей MPLS, позволяющий оптимизировать характеристики сети, пропускные способности и распределение потоков.

3. Проведены экспериментальные исследования предложенного алгоритма.

Литература

1. 1.Гольдштейн А.Б., Гольдштейн Б.С. Технология и протоколы MPLS. СПб. БХВ «Питер» 2005. 304 с.
2. Зайченко Ю.П., Хаммуди Мухаммед Али-Азам. Оптимальный выбор пропускных способностей каналов связи в сети с технологией АТМ. Вісник Національного технічного університету України «КПІ», сер. Інформатика, управління та обчислювальна техніка, №43. 2005. с.196-201
3. 3.Зайченко Ю.П., Шарадка Ахмед. Задача распределения потоков различных классов в сети с технологией MPLS. Вісник національного технічного університету «КПІ». Інформатики, управління та обчислювальна техніка №43. 2005. с.113–23

Authors' Information

Юрий Петрович Зайченко – НТУУ Киевский политехнический институт, профессор, Киев-56, проспект Победы 37, тел. 38-0442418693; e-mail: zaych@i.com.ua

Ахмед А. М. Шарадка (Иордания) – НТУУ Киевский политехнический институт, аспирант, Киев-56 , проспект Победы 37, тел. 38-0667655099; e-mail: Sharadqqh_78@yahoo.com

INDEX OF AUTHORS