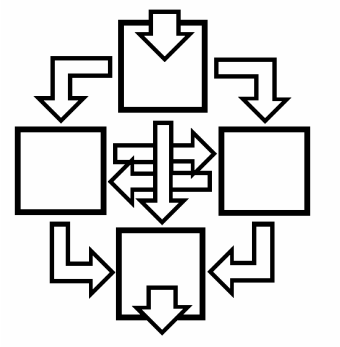
Twenty-ninth International Conference

INFORMATION AND COMMUNICATION TECHNOLOGIES AND PROGRAMMING

Varna, Bulgaria
June 24-26, 2004

ICT&P 2004



Proceedings

FOI-COMMERCE
SOFIA - 2004

The conference is organised by the Association for the Development of the Information Society and co-organised by the Department of Information Research at the Institute of Mathematics and Informatics.

Peter Barnev (editor)
Proceedings of the Twenty-ninth International Conference "Information and Communication Technologies and Programming"

Publisher: **FOI-COMMERCE, Sofia**
Sponsor: **Association for the Development of the Information Society**

First edition
Printed in Bulgaria by FOI-COMMERCE

ISBN 954-16-0028-X
C\o Jusautor, Sofia, 2004

TABLE OF CONTENTS

Second workshop on multimedia semantics

Information Materials

Preface

The Twenty–ninth Conference “Information and Communication Technologies and Programming – ICT&P’04 is taking place in the “St. St. Constantine and Elena“ resort near Varna.

The Conference is devoted to the development of the information society in all fields of human activity.

The Second special workshop on Multimedia Semantics is held within the framework of ICT&P’04.

This volume contains invited talks as well as regular papers reviewed and accepted by the International Program Committee.

The papers are ordered in the proceedings according to the schedule of presentation at the Conference.

The proceedings volume includes also information materials about the Association for the Development of the Information Society and about the International Journal "Information Theories and Applications".

I would like to express my gratitude to the invited speakers and to the authors of the others papers, presented at the Conference, as well as to the members of the Programme Committee. I would also like to thank all participants in ICT&P’04. Special thanks go to the Association for the Development of the Information Society and to the Institute of Mathematics and Informatics.

I would like to thank Prof. Peter Stanchev for organizing the Second special workshop on Multimedia Semantics.

I would also like to express my special thanks to Krassimir Markov the editor of the International Journal "Information Theories and Applications" and Krassimira Ivanova for their assistance in publishing this proceedings.

I would like to thank all the people who undertook the hard work of organizing this conference, and especially Detelina Stoilova – the Secretary of the Organizing Committee.

Petar Barnev

Invited Papers

COMPUTER DEMOCRACY – OUR NEXT STEP IN EUROPE

T. Vámos

Computer and Automation Research Institute, Hungarian Academy of Sciences
H- 1111 Budapest, Lágymányosi u. 11, Hungary
e-mail: vamos@sztaki.hu

Introduction

After about a quarter of a century of enlightened development and ongoing preparatory technological, scientific and political activities we are arrived at the realization period of the idea. The two major technological vehicles of progress are the World Wide Web, the most democratic international forum of information exchange and the advent of public key cryptography as a combined philosophical and practical device of individual integrity and collective responsibility.

The two major technological vehicles

A detailed explanation was given in detail in earlier papers and talks, several ones here in Bulgaria. A short summary to refresh your memory:

www:

- ► accessible all over the world, even if it is forbidden by (the authorities);
- ► difficult to trace the receiving user and even the dissemination origin;
- ► instant information regardless of distance;
- ► the creation of new groups within a global society;
- ► provides a stimulus for global standards of reasonable, acceptable, communication among different cultures, a real global human society.

Unfortunately as *Virtue* and *Evil* accompany all human related issues, we meet the *Evil* in:

- ► terrorism, crime, populist deceptive politicians, people spreading hate, misleading pseudo-science;

- ► an ocean of information without reliable and well oriented browsing facilities;
- ► the increasing orientation to business interests and not the original aim of free access to information.

To combat these problems, devoted professional people have developed new tools for elevating the *Virtues* and fighting against the Evil. We, the scientific brotherhood are active, too.

PKC: (public key cryptography)

- ► defends the individual against all kinds of mishandling of his/her personal data, ideas, views;
- ► elevates the responsibility and self-defence of the information issuers by preserving an unalterable document of the originally sent message;
- ► enables legally constituted public authorities to control malevolent information flow;
- ► enables all active constitutional, legal players in the information flow to control the legal conditions;

These two vehicles create not only technological tools for our global human efforts but a highly general instructive metaphor for future coexistence, the mutual responsibility of different people towards each other and their communities, i.e. a renewed and realistically establish-able *New Agora* in the Athenian, the first drafted but *never substantiated* democracy.

The current state in Europe

I suppose that you have some more detailed surveys on the subject here, at this conference or at any other meeting devoted to the subject. The best references are collected at the special homepage of e-government. I strongly recommend subscribing to this and join the group of these benevolent people who try to digest an immense and almost impossible amount of information. I am not an agent of this group but for the third year a happy consumer of their blessed activities and, organizational contacts. This helped us a lot in our successful joining process, which was celebrated on May 1 but prepared during the last quarter of a century.

Indicating the headlines only:

- The efforts are concentrated under the project name: E-2005

  that means a deadline —

  - for adopting standards for EU information; interoperation among the member countries. The IDA project of the EU is the main channel of communication, discussions and setting of standards based on these preparations;
  - for standards of realization on the principles reflected in the PKC ideology, i.e. protection of privacy balanced against the common interest of public democracy and defence of both the individual and communities against mistreatment by terrorism, fraud and political adventurers;
  - creating and distributing technology standards for all these purposes
  - helping people who are handicapped in education, social environment or some other condition to get appropriate support and equal opportunities.

- No EU country can serve as a general and completed standard due to the both highly different traditions of a democratic society and their technological status. In some countries, having a long positive experience in living in a people-serving and autonomous society, people see no problem with a more transparent system based primarily on a single natural identifier. This can be the normal data of domicile or birth register, too. In our kind of countries, people living a long time, i.e. centuries long, under foreign rule, consider the state authorities to be organizations against their civil interests, and traditionally regard underhand dealing as a national virtue. Therefore, they try to defend the individual at any cost against any imaginable governance intrusion.
- The interests of individual and community protection are different and their common view is relative to historical age and type of political system. We can state a positive viewpoint: in the last fifteen years these relations have changed a lot, sometimes also in the minds of the people, especially in that of the younger generations. Technology and experience of protection, and of course inimical actions, have changed even more.
- The EU boards have taken steps to reach a consensus, i.e.

appropriate standards according to the technological and psychological possibilities, putting in effect our common constitutional principles. We can learn a lot using the American technology and legislation experience but Europe has its own nearly three millennia experience of its own to reach from, sometimes more tragic, sometimes more human.

- The EU Constitution has accepted currently sets out the general principles outlined in the PKC ideology. Relevant additions should be the separation of personal identification and other data, having a virtual envelop and opening operation. All actions should be registered in a no erasable and no alterable way, monitored by legally elected, independent, responsible bodies. All data unifications should be erased after the action, except the result and the record of the action. All kinds of these data procedures should be permitted by the individuals concerned and communicated to them but the actions of legal authorities (prosecutors and courts) should be carried out under well-controlled legal conditions (e.g. communication of the action only and not the result, time limits for action and secrecy, notification of people for whom the action and the data should be opened or closed).

The Hungarian story and experience

The present Hungarian practice is one of the most rigorous in Europe forbidding all kinds of data unifications except those based on prosecutor's or court's decisions. The previous system of a personal identifier (an 11 digit string composed of gender and birth data and a four digit zip code) was abolished though hidden in some way until now by certain authorities. A set of three different and not unified codes was legalized: one for domicile registry (2 characters, 6 digits), one for taxation (10 digits) and one for social security (9 digits). All these happened nearly fifteen years ago, immediately after the fall of the uncontrolled police control system and at a time of very low-level civilian computer usage. The international state of legislative and cryptographic practice was lower by an order of magnitude and not only the US but the whole world lived before the drama of Sept. 11th and the massive experience of hacker and virus creating operations.

Electronic signature is generally not used though it is legislated. The reason, similar to the general European experience, is the exclusive

financial condition: groups receiving the authorization power would like to receive high profits and for justification they started or demanded immense investments for explosion safe buildings, hardware and software systems, all separate for different purposes. The obsolete legislative situation and the particular interests of the different political groups and authorities supported these exaggerated demands.

We have now arrived at a point of almost general consensus for a revision of the early nineties' views and the introduction of current algorithmic software tools. Possessing an excellent school of algorithmic procedures and probability theory we are ready to create a highly safe system. I refer to the schools of Rényi (our academic institute of mathematics recently adopted the name of Rényi) and Erdős.

Politics and science

Unfortunately, any kind of legislative action largely depends on mostly unintelligent, corrupt, malevolent, erratic politicians and their sycophants in dependent positions. In addition the situation in the daily press is submerging into a tabloid level, even the broadsheet newspapers are more and more interested in scandals and sensational news.

We proposed and partly realized a common effort of all sensible decision makers to unify our forces in a reasonable and given solution. Three branches of the government worked or shifted work and related financial responsibly to each other in the fuzzy channels of bureaucracy. The best educated and experienced, benevolent civil servants stood frustrated within the whirl of irresponsible politics.

The Academy of Sciences, being a partly independent and respectful body tries to convince the responsible decision making persons to consider national interest as a higher principle than their own financial and power involvement. The Committee, appointed by the President of the Academy includes outstanding personalities in the legal, social, computer related sciences, senior figures of our information history and the Ombudsman of data protection, who should be independent of political parties and elected by the Parliament. No active politician or government administrator is among the membership to maintain the Committee's political independence as a body. The Committee has no claim for any intervention but works with all state related people and organizations that are willing to do so.

I would like to mention that we found in a small minority of politicians

several devoted and able people, who joined political groups in the hope of improving the regrettable situation. However, they all are subdued by the overpowering, negative influence of the more aggressive unscrupulous powers. These positive actors, sitting on both sides of the political divide welcomed the initiative of the Academy and are meaningfully cooperating with us.

We have had to experience the disastrous influence of political splits in relevant non-political problems and the dysfunctional organization of the political system, in its personal selection constraints and in overburdening practice, extending political and administrative activity far beyond the really necessary principal tasks. The operation stimulated thinking about the revision of state administration practice, returning to much older ideas of democratic and professional governance by adoption of both new tasks and technologies.

According to our observations, similar problems arose in every developed country and organization, even in multinationals and other international bodies. Thus the problem is less an issue of unrealistic ethical philosophical judgement but much more a social, cultural and organizational issue, i.e. an information science related question of our age and our intellectual communities.

A consensus of relevant thought in the legal profession has now been reached. Those who were pioneers of our present democratic constitutional order advocate the need for rational revision and that provides additional support of the need, as a priority, professional quality in all public affairs. We refer to the great Greek thinkers on city-governance (πολιτεια) especially *Aristotle* and the funeral speech of *Pericles*, reported later by *Thucydides* and to the *Founding Fathers* of the US through, their essays and papers in the *Federalist*. From the 19th century we have also had a wonderful tradition in Hungarian history, starting with the *Sage of the Country by,* Ferenc Deák.

The proposal of the Hungarian Academic Committee

The proposal is clear:

- ► for the equal opportunity of citizens the right for electronic signature on an equitable basis, i.e.
  - it should be given free of charge for those whom it is a financial burden and not expensive for anybody. (.e.g. in relation to the taxation system)
  - electronic signature should be the only required authorization for any kind of public activity. If possible, this should be extended to banking operations, too;
  - all public authorities should participate in the popularisation, education and, training of different layers of society for usage and for being conscious of one's rights;
  - the state and, all accountable public authorities related to the electronic signature issue should be responsible for the preservation of the Civil Rights of individual citizens and any of their respective legal groups.

The measures are detailed above and should follow the agreements of the EU. EU conformity is the basis of interoperation and is a constitutional requirement. According to our legal experts this requires no fundamental change in our legal system, only some further updating and corrected interpretations, and the constitutional empowerments for participation in the EU.

- ► Technological means should not be included in the legal regulations, the system must be flexibly open for any kind of realization, i.e. currently traditional authorized handwriting, smart card, SIM-card used in mobile systems, biometrical (fingerprint, fundus, DNS, etc.) data.
- ► The Law should take care of independent and open operational authorities prescribing algorithms, the code length for citizens and prosecution and, other safety conditions related to data and their handling personal.

Going together — Neumann and Athanasoff — Iliev

Bulgaria and Hungary have much common ties in our history, beginning with the Huns for those who believe the Hungarians are the successors of Attila and the ancient Bulgars who are really supposed to be the descendants, with lesser and greater Byzantine influence, with the tragedy of a certain city called Varna in 1444, with Turkish, German and Russian domination but most important of all should be the future, based on another lesson: of Neumann and Athanasoff.

Both were pioneers of the computer age, Neumann in mathematical and logic theory, Athanasoff more in technology. Neumann had to leave his country to avoid being a victim of the Holocaust, Athanasoff's family left for a better life, both, subsequently, had more possibility to develop their genius.

Now we enter a new age, based on our common three millennia old European history and, hopefully, our talent find a home within a more peaceful, less hatred-contaminated world, preparing a common home for our descendants. I remember here my friend Lubomir Iliev who passed away not too long ago and was not only a great mathematician and teacher of computer science but, at the same time, a representative of European cultural tradition and values. We always considered the two be inseparable by regarding these subjects as both metaphors and parallel realities.

These are the main lessons of that progress: preserving individual values within a cooperative, empathy driven human community. Let us hope that this comes true!

NEURAL NETWORKS: A DIAGNOSTIC TOOL FOR GASTRIC ELECTRICAL UNCOUPLING?

Catherine P. Gooi and Martin P. Mintchev

Department of Electrical and Computer Engineering,
University of Calgary,
Calgary, Alberta, Canada T2N 1N4

Address for correspondence:
Professor Martin P. Mintchev, Ph.D. P.Eng.
Department of Electrical and Computer Engineering, University of Calgary
2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4
Phone/Fax: (403)220-5309, e-mail: mintchev@enel.ucalgary.ca

Abstract: Neural Networks have been successfully employed in different biomedical settings. They have been useful for feature extractions from images and biomedical data in a variety of diagnostic applications. In this paper, they are applied as a diagnostic tool for classifying different levels of gastric electrical uncoupling in controlled acute experiments on dogs. Data was collected from 16 dogs using six bipolar electrodes inserted into their antral gastric wall. Each dog underwent three surgically induced conditions: (1) basal, (2) mild uncoupling, and (3) severe uncoupling. For each condition half-hour recordings were made. The neural network was implemented according to the Learning Vector Quantization model. This is a supervised learning model of the Kohonen Self-Organization Maps. Records of the data collected from the dogs were used for network training. Remaining records served as a testing tool to examine the validity of the training procedure. Approximately 90% of the dogs from the neural network training set were classified properly. However, only 31% of the dogs not included in the training process were accurately diagnosed. The poor neural-network based diagnosis of records that did not participate in the training process might have been caused by inappropriate representation of input data. Previous research has suggested characterizing signals according to certain features of the recorded data. This method, if employed, would reduce the noise and possibly improve the diagnostic abilities of the neural network.

Keywords: Neural Networks, Gastric Electrical Activity, Gastric Electrical Uncoupling

1. Introduction

Neural networks are useful tools in medical settings. They have been applied successfully in classifying various forms of Parkinson syndrome [1], in diagnostic electromyography [2], and in studying breast cancer disease [3]. These applications are usually implemented using Kohonen self-organizing neural networks [1-3]. Kohonen maps build clusters based on the similarities among input data. In this case it will be applied in diagnosing gastric electrical uncoupling.

1.1. Medical Condition

Gastric motor function is an important part of the digestive process, and entails the storing, mixing and grinding of food, as well its movement towards the intestines. This process requires the coordination of gastric smooth muscle contractions. Similarly to cardiac contractions, stomach contractions are preceded by electrical activity. These electrical events determine the frequency, velocity and direction of the contractions [4]. Accordingly, abnormal gastric function can occur when electrical signals are not synchronized, i.e. the electrical signals are uncoupled. To detect uncoupling, internal recordings of gastric electrical activities are made [5]. From these records it is important to be able to categorize the severity of uncoupling. This paper proposes the use of neural networks for such categorization.

1.2. What is a Neural Network?

Neural networks consist of interconnected simple computing cells, referred to as “neurons” [6]. The strengths of the interconnections between these neurons are called synaptic weights (Figure 1).

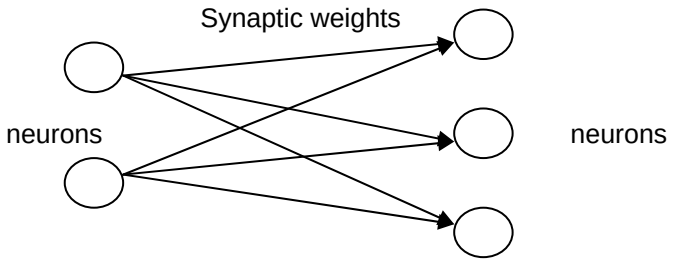


Figure 1: Simple Neural Network Architecture

Through modification of the synaptic weights, called learning or training, the neural network is able to store information.

1.3. Neural Network Training

In the first part of the training process the synaptic weights are initialized to small random values. Next, a training set of data is introduced to the network. There are two types of learning: (1) supervised learning, and (2) unsupervised learning. In supervised learning, a set of inputs along with target outputs is provided to the network. The network passes the inputs through the layers of neurons and modifies the synaptic weights according to a learning algorithm, which adjusts the outputs closer to those of the desired target outputs. In the case of unsupervised learning, no target outputs are provided. An example of unsupervised learning is found in the Kohonen map's competitive learning [7-8]. Kohonen maps cluster input data according to their similarities.

1.4. Kohonen Maps

Kohonen self-organizing maps are a type of neural network. They consist of two layers of neurons, the input neurons and the output neurons. Each input neuron is connected to every output neuron [6]. An example of Kohonen map architecture is shown in figure 2.

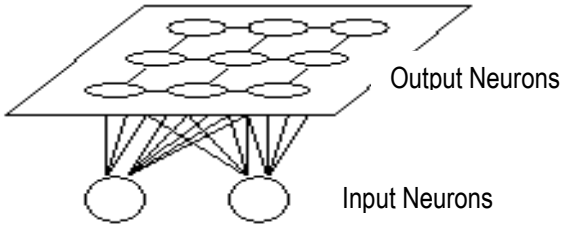


Figure 2: Kohonen Map

Kohonen maps learn in a competitive manner. First, the synaptic connections between the input and the output neurons are initialized, and each output neuron is characterized by a synaptic weight vector. Next, an input pattern (vector) is randomly selected. The Euclidean distance between the input vector and each synaptic weight vector is computed, and then the output neuron with the shortest Euclidean distance is declared the winning neuron. The synaptic weights of the winning neuron are adjusted, increasing

the similarity between its synaptic weight vector and the input vector. Similarly, the weight vectors of the neurons in the proximity of the winning neuron are adjusted, increasing their similarity, but to a lesser degree than that for the winning neuron [8]. The algorithm for weight adjustment is:

$$\omega_{ij}(t+1) = \omega_{ij}(t) + \eta(t)\left(x_i(t) - \omega_{ij}(t)\right) \quad (1)$$

where $\omega_{ij}(t)$ is the synaptic weight value from input neuron *i* to output neuron *j*; $x_i(t)$ is the input to neuron *i* at time *t*; and $\eta(t)$ $(0 < \eta(t) < 1)$ is the learning rate coefficient.

Due to its ability to group similar data, competitive networks are particularly useful for diagnosis, allowing similarly characterized inputs to be clustered together. This learning method, however, does not allow the user to control the categories into which the input will be classified. Learning vector quantization (LVQ) networks, on the other hand, allow the user to classify the input vectors into predetermined categories.

1.5. Learning Vector Quantization

LVQ is a supervised learning technique employed in combination with Kohonen maps [8]. As illustrated in figure 3, an LVQ network consists of an input layer, a competitive layer and a linear layer [9].

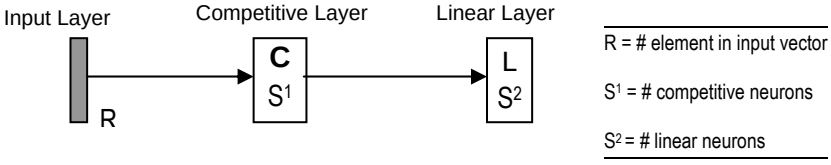


Figure 3: LVQ Network Architecture

The competitive layer classifies the inputs as described above, while the linear layer classifies the outputs from the competitive layer into target values. In other words, the outputs of the competitive layer are subclasses of the target layer. If the output of the input vector matches the target value, the weight vectors of the winning neuron, n_w, are modified with the following algorithm,

$$n_w(t+1) = n_w(t) + \eta(t)[x(t) - n_w(t)], \quad (2)$$

otherwise they are modified using:

$$n_w(t+1) = n_w(t) - \eta(t)[x(t) - n_w(t)], \quad (3)$$

Equation (2) moves the competitive neurons closer to vectors that belong in its same class, and equation (3) moves the competitive neurons farther from vectors that do not belong in its same class. In equation (2) and (3), $n_w(t)$ represents the winning neuron's present synaptic weight vector, i.e. at time t, $n_w(t+1)$ represents the winning neuron's modified synaptic weight vector, i.e. at time $t+1$, $\eta(t)$ $(0 < \eta(t) < 1)$ represents the learning rate coefficient, and $x_i(t)$ is the input to neuron i at time t.

2. Aim

The aim of this paper is to apply Learning Vector Quantization neural networks in recognizing gastric electrical uncoupling from internal recordings of canine gastric electrical activity.

3. Experimental Design

3.1. Data Acquisition

In order to understand and recognize varying degrees of uncoupling, 16 anesthetized dogs underwent surgically induced gastric uncoupling [9]. Data were obtained from each dog in the three different states. Six pairs of electrodes were placed in the antral gastric wall of the dogs, three along the anterior wall and three along the posterior wall. These six pairs of electrodes provided 6 channels from which half-hour recordings of gastric electrical activity (GEA) were made for each state. During the first session the dogs were in basal state, in the second session the stomach was divided by a single circumferential cut of the entire gastric muscle between the distal and the middle electrode sets, dividing the organ into two electrically active regions each oscillating at different electrical frequency, thus producing mild uncoupling. Finally, a second circumferential cut surgically divided the stomach between the middle and the proximal electrode sets, dividing the organ into three electrically active regions, simulating severe uncoupling.

The gastric electrical activity (GEA) signals were filtered in a frequency band of 0.02 - 0.2 Hz and digitized with a 10 Hz sampling frequency. In total 18 000 samples were collected for each channel per session.

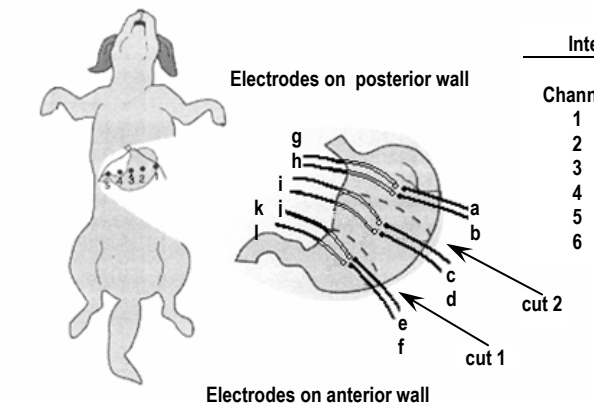


Figure 4: Data acquisition setup.
The locations of the circumferential cuts are also denoted.

3.2. Neural Network Modelling

The aim of the neural network is to categorize the condition of the dogs into one of the 3 states:

1. Basal
2. Mild Uncoupling: One circumferential cut
3. Severe Uncoupling: Two circumferential cuts

In view of the fact that the categories are predetermined, an LVQ network is chosen for the implementation. The network is created, trained and simulated using the Neural Network Toolbox from MatLab 6.0 (MathWorks, Natick, MA).

Since the neurons from the competitive layer form subclasses for the linear layer's target neurons, the number of neurons in the competitive layer should always be larger than the number of target neurons [9]. In addition, the number of neurons in the competitive layer should be smaller than the number of training examples, otherwise each training example would have a separate winning neuron in the competitive layer and the competitive layer would serve no purpose in the classification process. Given these limitations the number of neurons in the competitive layer was chosen to be six, two neurons belonging to each target class.

Next step in the implementation process was the training of the network.

In order to train the network, data were to be selected and represented in a vector form acceptable for the input neurons.

3.3. Data Selection and Representation

To determine which data samples should be used, the data from dogs 1 through 16 were displayed using locally designed gastrointestinal signal acquisition and analysis software, GAS v. 3.0. Visual inspection indicated that channel 2 was not functioning for dog 1 thus it was not used for training the neural network. In addition, data from dog 2 and 3 were collected utilizing filters with smaller bandwidths. This is a manner, inconsistent with the data acquisition parameters utilized for the other records, so these two recordings were also disregarded when training the neural network. It was also noted that sometimes signals were not adequately registered within the first few seconds of recording, thus data for training and validation were extracted from the middle of the recording time. 5000 recorded data samples from each channel were utilized. Each training session was therefore characterized by an input vector of 30 000 elements, 5000 from each of the six channels.

3.4. Training and Simulating

The LVQ network was trained by repeatedly feeding the network with data from dogs 4 through 13, and their corresponding target outputs. Each time the data was fed through the synaptic weights were modified according the learning algorithm described earlier. The first output neuron was designated for the basal state, the second for mild uncoupling, and lastly the third for severe uncoupling.

Data from dogs 4 through 16 were used for simulation and verification. The vectors for each case were input and the output was recorded and compared with desired outputs. Outputs from dogs 4 to 13 provided verification of the network ability to diagnose for cases it has seen before during the learning process and outputs from dogs 14 to 16 were used to demonstrate the network ability to generalize.

4. Results

4.1. Neural Network Training and Verification

Training was performed with thirty training vectors from Dog 4 to Dog 13 (three from each dog), and subsequently, seven simulations were executed with data from Dog 4 to Dog 16. The average result is shown in the Table 1.

	Percentage of correct diagnosis
Dog 4 – Dog 13	89.5%
Dog 14 – Dog 16	31.0%

Table 1: Verification results

The performance of the network for diagnosing dogs within the training set was fairly high, at 89.5%. However, its generalization ability was poor and had an accuracy of only 31%.

5. Conclusion

The network diagnosed well for the training data but was unable to provide accurate diagnosis for new cases. Performance of a neural network is directly related to the quality of its input data. Therefore it is necessary for it to contain sufficient information [10]. It is important to represent the input data in an appropriate fashion, eliminating noise where possible and capturing characterizing features. Similar study [2], suggested that segments of the signal can be characterized by seven parameters, duration, spike duration, amplitude, area, spike area, phase and turns. This study involved diagnosing neuromuscular disorders based on the electromyography (EMG) recordings of muscle electrical activity. The study resulted in an accurate diagnosis in the order of 80%. As a proposal for further study, a similar approach might be applied for characterizing the gastric electrical signals.

Acknowledgement

This study was sponsored in part by the Natural Sciences and Engineering Research Council of Canada

References

[1] Fritsch T., Kraus P.H., Pruntek H., Tran-Gia P.: "Classification of Parkinson Rating-Scale-Data Using a Self Organizing Neural Net", In: IEEE International Conference on Neural Networks, pp. 93-98, Mar28-April1 1993.

[2] Pattichis C.S., Schizas C.N., Middleton L.T.: "Neural Network Models in EMG Diagnosis", IEEE Transactions on Biomedical Engineering, v42, n5, pp. 486-496, Piscataway, NJ, May 1995.

[3] Allan R., Kinsner W: "A Study of Microscopic Images of Human Brest Disease Using Competitive Neural Networks", In: Canadian Conference on Electrical and Computer Engineering, v1, 2001.

[4] Sanmiguel C.P., Mintchev M.P., Bowes K.: "Electrogastrography: A noninvasive technique to evaluate gastric electrical activity", Canadian Journal of Gastroenterology, vol. 12, n. 6, September 1998.

[5] Mintchev M.P., Otto S.J., Bowes K.L.: "Electrogastrography Can Recognize Gastric Electrical Uncoupling in Dogs", Gastroenterology 112:2006-2011, 1997.

[6] Haykin S.: "Neural Networks a Comprehensive Foundation", New Jersey: Tom Robins, 1999.

[7] Kohonen T.: "Self-Organizing Maps", New York: Springer-Verlag Berlin Heidelberg New York, 1997.

[8] Beale R., Jackson T.: "Neural Computing an Introduction", Bristol and Philadelphia: Institute of Physics Publishing, 2001.

[9] Demuth H., Beale M.: "Neural Network Toolbox User's Guide", The MathWorks Inc., 1998.

[10] Noyes J.: "Training and Generalization", In: Handbook of Neural Computation, IOP Publishing Ltd and Oxford University Press, pp. B3: 5:4, 1997.

HANDWRITING IN FORENSIC INVESTIGATIONS

Georgi Gluhchev

Institute of Information Technologies – BAS
2, Acad. G. Bonchev Str., Sofia 1113
e-mail: gluhchev@iinf.bas.bg

Abstract: The process of automatic handwriting investigation in forensic science is described. The general scheme of a computer-based handwriting analysis system is used to point out at the basic problems of image enhancement and segmentation, feature extraction and decision-making. Factors that may compromise the accuracy of expert's conclusion are underlined and directions for future investigations are marked.

Keywords: Handwriting, Identification, Verification, Screening, Forensic Investigation, Image Enhancement, Feature Extraction, Decision-making.

1. Introduction

The individual's authentication becomes a serious problem concerning different areas of the social and economic relations in the world. Its importance increases as a factor in the prevention of terrorist actions and illegal access to important information. The problem attracts the attention of researchers all over the world and during the last years a few international projects were launched aimed at the development of reliable systems for authentication using biometric information.

Writer authentication is one of the broadly investigated modality in this aspect. Until now it was mainly used in forensic investigations dealing with handwritten document analysis or signature verification. Despite of the long history in that respect handwriting investigation still remains a difficult, time-consuming and subjective process, where qualified experts evaluate the similarity between letters, strokes and writing styles on the basis of their experience. In all cases of writer identification the objectiveness of the analysis and reliability of conclusion are of great importance. However, the inevitable variation in writing under different conditions and psychological state or when a significant time gap exists between the incriminated and reference documents may mislead the expert. Also, extremely difficult are cases where the handwriting is deliberately changed. In such situations

different experts may disagree as to who is the writer of a particular document and a wrong conclusion may be drawn.

Other problems in writer identification concern the expert's workload during the analysis and difficulties stemming from sometimes poor quality of handwritten materials.

To overcome the problems quantitative methods for objective handwriting analysis and adequate decision-making have to be developed and implemented.

To achieve this, serious scientific investigation is required in order to develop appropriate methods for feature measurement, selection of reliable sets of features, evaluation of the minimal number of handwriting elements which is necessary for the reliable decision-making, suggestion of robust classification algorithms dealing with mixtures of continuous and categorical variables. Major difficulties in this direction stem from the qualitative character of most of the handwriting parameters used by the experts.

Despite that the problem of writer identification is of great practical importance in forensic investigations a relatively small number of papers have been published until now [5,11,14,22]. The existing computer systems are aimed especially at the screening of similar handwritings from a large data base of handwritings. After that the identification problem is carried out manually by an expert on the basis of his one experience and subjective evaluation of the similarity between the handwritings under investigation. For this he compares visually strokes, letters or combinations of letters performing sometimes simple measurements.

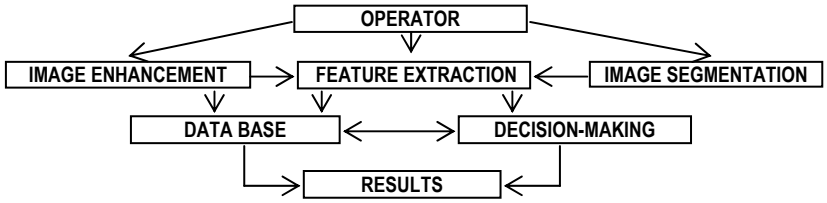


Figure 1. Block-diagram of a handwriting investigation system

The goal of the paper is to sketch the frame of a computer-based handwriting investigation system and discuss the problems of its major components (Fig. 1). It is organized in the following way: Section 2 concerns the improvement of image quality; Section 3 deals with feature extraction techniques; Section 4 describes decision-making and Section 5 points out at the unsolved problems.

2. Image Enhancement

The block-diagram in Fig. 1 shows that the development of a computerized handwriting investigation system follows the general methodology for the development of image processing and pattern recognition systems.

Since very often the handwritten materials are of poor quality, it is necessary to achieve some pre-processing. Its goal is two-fold: a) to improve image quality including contrast enhancement, random and structured noise reduction, and edge sharpening [2,3,9,17,19,21]. In such a way the image will offer better possibilities for the automatic analysis; b) to correct strokes and complex lines using morphological operations. This is especially important for the analysis of specific handwriting features, where the skeleton of the characters is used. Morphology allows to automatically connecting disrupted lines or cutting-off wrongly connected strokes.

3. Image Segmentation

Image segmentation is an essential step in the automatic document analysis. In handwriting analysis additional difficulties may arise due to the possibility of significant variation in rows', words' and letters' position.

The automatic segmentation includes background elimination as a first stage. Approaches based on histogram analysis (uniform background is supposed), locally adaptive binary trees and heuristic approaches (non-uniform background or presence of structured noise is assumed) are used.

The second stage concerns separation of rows. This operation is based on the analysis of histogram-like graphs obtained via horizontal projections of pixels, Hough transform or analysis of envelopes of continuous lines. While this operation could be easily achieved, special techniques are required for a proper detection of under-row and over-row elements of some characters. Also, the medial axis of the row may not be presented as a single straight line which requires piece-wise presentation.

The segmentation continues with the separation of the words. Different

cases of concise writing or writing where letters are not connected between them are a challenge. For the word separation vertical projections of pixels from the corresponding row are analyzed, distances between envelopes of continuous lines are used, separation lines parallel to the predominant slope of the vertically oriented strokes could be applied, as well.

However, the most difficult problem concerns segmentation of letters and strokes. Except some special cases, e.g. child's writing, their delineation may be quite difficult even for a human being. A proper solution of this problem could be achieved via a man-machine dialog. The operator has to identify some specific points like end-points or vertexes of a polygon that encompasses a handwriting element. After that lines could be automatically investigated for the detection of points of intersection or bifurcation, local extremums and like [12,13,20].

4. Feature Extraction

Feature extraction is the crucial problem that should be solved. While during the last decades a common methodology for the handwriting analysis has been set up, many of the suggested features are of qualitative character and are prone to different evaluation from different experts. Also, there are no strong recommendations as to what number of features is to be used for a reliable decision-making.

Different types of features may be investigated including graphometric, densitometric, categorical, model-based and topological invariants [2,5,6,7,8]. From the expert's point of view they are classified as general and specific features.

The general features are of categorical type and describe qualitative characteristics as: degree of connection between letters (usually three degrees are accepted: low, moderate and high), slope (right, left, upright), motion (rectilinear, curvilinear, angular or arched, loop-like, oval, wavy or spiral), elaboration (presence of ornaments), direction of movement (clock-wise or counter-clock-wise), quantity of movement (average number of strokes used to draw separate letters) and like [2,7,8,12,16]. They are difficult for automatic evaluation and are specified by the expert.

Specific features admit quantitative evaluation. They are known as graphometric and are aimed at the automatic or semi-automatic measurement of the following characteristics: distances between rows, height and width of letters, distances between letters, size of the above-row

and under-row elements, distances between words, predominant slope, geometric parameters of handwriting elements like strokes, fragments and/or combination of characters [2,11,22].

Figure 2. a) Original image, b) Areas of different pressure

While many of the above-mentioned features can be easily imitated, there are features that are not clearly seen and, therefore, difficult to falsify. In that respect a special attention is paid to the distribution of pressure alongside the strokes (Fig. 2). It could be analyzed in different ways looking for a reliable description, e.g. evaluation of the geometric parameters of areas of different pressure and their mutual disposition at different writing elements, or pressure change alongside the skeleton of the elements. In that respect promising results have been reported in [15].

Except the described features, which are reasonable and intuitively clear for the experts, other features that do not express a particular property of the handwriting may be used as well. These include topologically invariant points associated with a particular character. According to this approach characters are divided into specific segments that can be transformed and compared piece-wise.

Another approach is based on the presentation of characters as elastic changes of an ideal model. Thus, a transformation between the model and the real character can be evaluated and its parameters used for classification.

Very important is the problem concerning the reliability evaluation of different set of features. A number of approaches may be used for this based on the information theory, statistics and classification power.

5. Decision-Making

The overall estimation of the similarity between two handwritings must be obtained as a combination of decision-making classifiers.

The decision-making for the specific features is based on the evaluation of the similarity between particular elements from the handwritings under investigation. Since the overall estimation will be based first on the estimations of separate elements and second on groups of elements, multi-level classifiers are to be used. The first level will concern the comparison of basic elements like strokes, letters and signs of punctuation. At the output of these classifiers every element will be assigned a number that reflects the degree of similarity between the handwritings under investigation. Since a particular element may be detected in a few places in the text, an average similarity relative to this element will be calculated at the second stage. After the similarity is evaluated for all different elements, an overall evaluation is being obtained at the third level. One of the basic problems that has to be solved here concerns the weight factors of the elements, i.e. their classification power. Different types of decision rules could be used, including statistical, linear, heuristic, and NN-based [1,4,6,18].

The authenticity, where a forgery is expected, would be predominantly verified using stable features like pressure distribution. This is usually applied to small pieces of written text like particular words or signatures. Depending on how the pressure will be measured (areas of different pressure or a function alongside a skeleton line) different comparison techniques could be applied.

The categorical features are mainly used for the search of similar handwritings in a large database. Also, for the sake of one-to-one comparison mixed variables discriminant techniques could be used. A simple approach for the analysis of mixture of categorical and continuous data requires arbitrary scoring of all the categorical variables followed by the use of standard methods for multivariate continuous data, which in the case of classification means use of techniques such as linear or quadratic discriminant analysis.

6. Discussion and Future Work

Different aspects and major problems that are to be solved for the development of a computer-based handwriting investigation system are described.

While the pre-processing stage of digital images is thoroughly investigated during the last decades, the well-known approaches may not work properly due to possible damages, changed background and poor contrast of the images. This requires locally adaptive methods to be developed, reflecting the specificity of the investigated images.

A big challenge is the selection of a reliable set of features. A computerized system must include as much as possible features that experts are accustomed to, but at the same time, special attention must be paid to the measurement of some parameters that are difficult for expert's evaluation, and therefore difficult for imitation, e.g. curvature at characteristic points, line smoothness or pressure distribution. Also, the expert has to have the possibility to interfere and suggest his one selection of features. This requires a friendly man-machine interface to be available.

The decision-making seems to be the most expert-independent part of the problem, since various measures of similarity (parametric, non-parametric, clustering) have been developed in pattern recognition theory. The different levels of similarity estimation will require the development of multi-level hierarchical classifiers.

Age-due variations in handwriting or changes due to different diseases must be investigated as well. This will increase the possibility for reliable writer identification when a significant time-gap between the handwritten materials exists or in case of a psychiatric disease [10].

A successful solution to all of the discussed problems will allow developing of a reliable and user-friendly computer system for handwriting analysis that could be implemented in police departments, bank and notary offices. It must be noted that such a system will help the expert to do an objective analysis, not to replace it.

The obtained solutions to specific handwriting analysis problems could be easily incorporated in a more complicated access-permit systems or person authentication systems at check points.

Acknowledgements

This work is supported by the Ministry of Education and Sciences under contract # И 1302/2003.

References

Alexandre, Luís A., A. C. Campilho, M. Kamel. "Combining Unbiased and Independent Classifiers Using Geometric Mean", 11th Conf. of Portuguese Association on Pattern Recognition, 2000

Angelov, A., D. Nestorov, S. Benchev, G. Gluhchev, D. Kamenov, P. Veleva. "A system for the automated handwriting investigation", In: Information Bulletine of RIFSC, Sofia, 1997, pp. 127-133

Dimov, D. T., G. Y. Gluhchev. "A Locally Adaptive Binary-Tree Method for Binarization of Text Images", In: Progress in Handwriting Recognition, (Edts A.C. Downton, S. Impedovo), World Scientific, London, 1996, pp. 575-580

Gluhchev, G., V. Shapiro, D. Lalchev. "Probabilistic estimate of handwriting similarity", In: Theory and application of cybernetic systems, Bulgarian Academy of Sciences, Sofia,1991, 3,pp. 90-93

Gluhchev, G., V. Shapiro, S. Ogorelkov. "Automatic investigation of handwritten characters", In: Forensic science and criminology, Sofia, 1989, 5,pp. 135-142

Heutte, L., T. Paquet, J.V. Moreau, Y. Lecourtier and C. Olivier. "A Structural/Statistical Feature Based Vector for Handwritten Character Recognition", Pattern Recognition Letters, vol. 19, no. 7, pp. 629-641, 1998.

Heutte, L., J.V. Moreau, B. Plessis, J.L. Plagnaud and Y. Lecourtier. "Handwritten numeral recognition based on multiple feature extractors", 2nd IAPR International Conference on Document Analysis and Recognition, IAPR-ICDAR'93, Tsukuba, Japan, IEEE Computer Society Press, pp. 167-170, 1993.

Heutte, L., J.V. Moreau and Y. Lecourtier. "A new feature extraction strategy for handwritten pseudo-character recognition based on multiple feature extractors", 6th IGS International Conference on Handwriting and Drawing, IGS-ICOHD'93, Paris, France, pp. 186-188, 1993.

Hummel R.A. "Image enhancement by histogram transformation", Comput. Graphics and Image Processing, 6, 1976, pp. 184-197

Kalcheva, E., G. Gluhchev. "Evaluation of handwriting changes in case of neurological diseases", Proceedings of the CompSysTech '2000, Sofia, pp.V.2-1–V.2-5

Klement, V. “an Application System for the Computer Assisted Identification of Handwritings”, Proc. Int. Carnahan Conf. Security Technology, Zurich, 1982, pp. 75-79

Lickforman-Sulem, L. “Extraction d’elements graphiques dans les images de manuscripts, CIFED’98, Quebec, Canada, 1998, pp. 198-207

Lickforman-Sulem, L. and C. Faure. “A Hough Based Algorithm for Extracting Text Lines and Handwritten Documents”, ICDAR’95, Montreal, 1995, pp. 774-777

Lima, Rui M., A. J. C. Campilho. "A System for Image Analysis of Documents", 7th Conf. of Portuguese Association on Pattern Recognition, pp. 1.5.1-1.5.6, 1995.

Nestorov, D., V. Shapiro, P. Veleva, G. Gluhchev, A. Angelov, I. Stoyanov. "Towards objectivity of handwriting pressure analysis for static images", 6th Int. Conf. on Handwritting and Drawing ICOHD '93, 5-7 July, Paris, pp. 216-218

Nosary, A., L. Heutte, T. Paquet and Y. Lecourtier. "Defining writer's invariants to adapt the recognition task", 5th International Conference on Document Analysis and Recognition, ICDAR'99, Bangalore, India, IEEE Computer Society Press, pp. 765-768, 1999.

Pizer, S.M., E.P. Amburn, J.D. Austin et al. “Adaptive hiastgram equalization and its variations”,Comput. Vision, Graphics and Image Proc., 39, 1987, 355-368

Plessis, B., A. Sicsu, L. Heutte, E. Menu, E. Lecolinet, O. Debon and J.V. Moreau. "A multi-classifier combination strategy for the recognition of handwritten cursive words", 2nd IAPR International Conference on Document Analysis and Recognition, IAPR-ICDAR'93, Tsukuba, Japan, IEEE Computer Society Press, pp. 642-645, 1993.

Pratt W. K. Digital Image Processing, 2nd edn, John Wiley & Sons, 1991

Shapiro, V., G. Gluhchev, V. Sgurev. "Handwritten document image segmentation and analysis", Pattern recognition letters, North Holland, 1993, 14,pp. 71-78

Shapiro, V., G. Gluhchev, V. Sgurev. "Preprocessing for automatic examination of handwritten documents", 7th Scandinavian conf. on image analysis, Denmark, 1991,v.2, pp. 790-797

Steinke, K. “Entwicklung von Mustererkennungsverfahren zur textunabhangigen Analyse von Handschriftenbildern”, Dissertation TH Aachen, 1981

Papers

PRACTICAL, COMPUTATION EFFICIENT HIGH-ORDER NEURAL NETWORK FOR ROTATION AND SHIFT INVARIANT PATTERN RECOGNITION

Evgeny Artyomov and Orly Yadid-Pecht

The VLSI Systems Center, Ben-Gurion University, Beer Sheva 84105, Israel.
E-mail: artemov@bgumail.bgu.ac.il

Abstract: In this paper, a modification for the high-order neural network (HONN) is presented. Third order networks are considered for achieving translation, rotation and scale invariant pattern recognition. They require however much storage and computation power for the task. The proposed modified HONN takes into account a priori knowledge of the binary patterns that have to be learned, achieving significant gain in computation time and memory requirements. This modification enables the efficient computation of HONNs for image fields of greater that 100 × 100 pixels without any loss of pattern information.

Keywords: HONN, higher-order networks, invariant pattern recognition.

1. Introduction

Invariant pattern recognition using neural networks was found to be attractive due to its similarity to biological systems. There are three different classes that use neural networks for invariant pattern recognition [1], that differ in the way invariance is achieved, i.e. Invariance by Training [2], Invariant Feature Spaces, or invariance by Structure, good examples are: the Neocognitron and HONN [3].

In third-order networks, which are a special case of the HONN, invariance is built into the network structure, which enables fast network learning with only one view of each pattern presented at the learning stage. However, an exponentially growing amount of interconnections in the network does not enable its usage for image fields larger than 18 x 18 pixels [3]. A few different solutions were proposed to minimize the number of the HONN interconnections. Weight sharing, by similar triangles [3]. Weight sharing by “approximately similar triangles” [4]-[5]. Coarse coding [6]. Non-fully interconnected HONN [7]. All these methods partially solve the problem of the HONN interconnections but do not help with larger images.

Consequently, the research community in the field of invariant pattern recognition largely abandoned the HONN method.

In this paper, a modification for the third-order network is described. The proposed modification takes into account a priori knowledge of the binary patterns that must be learned. By eliminating idle loops, the network achieves significant reductions in computation time as well as in memory requirements for network configuration and weight storage. Better recognition rates (compared to conventionally constructed networks with the same input image field) are attained by the introduction of a new "approximately equal triangles" scheme for weight sharing. The modified configuration enables efficient computation of image fields larger than 100 × 100 pixels without any loss of image information — an impossible task with any previously proposed algorithm.

2. HONN architecture

Following equation describes the output of a third-order network:

$$y_i = f\left(\sum_j \sum_k \sum_l w_{ijkl} x_j x_k x_l\right), \quad (1)$$

where *w* is the weight associated with a particular triangle, *y* is the output and *x* is a binary input, *j*, *k*, and *l* are the indices of the inputs.

A schematic description of this network is shown in Fig. 1.

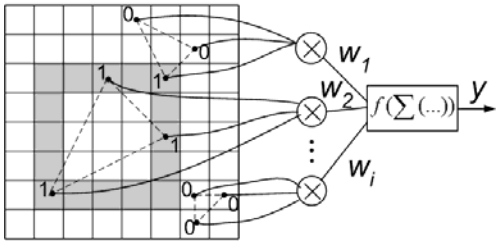


Fig.1. Schematic description of a third-order network

In the training phase, the perceptron-like rule is used:

$$\Delta w_{ijkl} = \eta(t_i - y_i) x_j x_k x_l , \quad (2)$$

where *t* is the expected training output, *y* is the actual output, η is the

learning rate and *x* is a binary input.

The number of triangles (*NoT*) can be calculated with the following equation:

$$NoT = \frac{IN!}{(IN-3)!3!}, \quad (3)$$

where *IN* is the number of input nodes.

For image fields 100 x 100 and 256 x 256 the number of triangles will be 1.6662×10^{11} and 4.6910×10^{13} accordingly. As can be seen, the number of triangles grows very fast off the limits of any current hardware. A few techniques to reduce the number of weights have been proposed in the literature (as described in section 1), but they do not reduce computation time.

The problem of large computational demands arises since the network is constructed in the pre-processing stage before the learning phase. At this stage, all possible triangles are computed and pointers to the weights are saved [8]. In addition to the pointers, the weight array is also stored. At least two memory bytes are required for each pointer. If, for example, an input field of 100 x 100 pixels is given, the total number of bytes required to store the entire vector of pointers is 3.3324×1011 bytes. The memory and computation requirements are enormous. To work with large input patterns, significant network modifications are required.

3. The proposed modified HONN method

As noted before, the input pattern is binary: edge or contour pixel has the value “1” and all other pixels have the value “0”. As can be seen from equation (1), each product with pixel value “0” will give “0” as a result. This means that only active triangles (in which all pixels belong to an object contour) will influence the result. In addition, the weights that belong to the inactive triangles will not be updated and will keep “zero” value during the learning process.

Following this observation, the network can be modified and all inactive triangles can be disregarded during the construction phase, which eliminates the idle loops from the computation. With this modification, the network configuration strictly depends on the input patterns that have to be learned.

In addition, to improve network performance regarding rotation, distortion and a number of learned classes we introduce an “approximately

equal triangles" scheme for network construction. This scheme, in addition to the "approximately similar triangles" scheme presented in [4] for weights sharing, adds triangle area equality. This means that "approximately similar triangles" with "approximately equal" areas will share the same weight.

3.1 The proposed network construction

The modified algorithm for network construction can be described as follows: 1. Load all patterns that must be learned. 2. Run through each image and save the coordinates of the contour (boundary) pattern pixels to the different arrays. A set of such arrays is shown in Fig. 2. 3. Compute angles of all presented triangles and classify them in order to associate with a particular weight.

Indices X_{im} ,Yim and n_{ij} correspond to pattern number (n), pixel number (m), weight index (j) and puttern number (i). The variable n_{ij} is the number of triangles from the particular pattern that correspond to the particular triangle weight index (class).

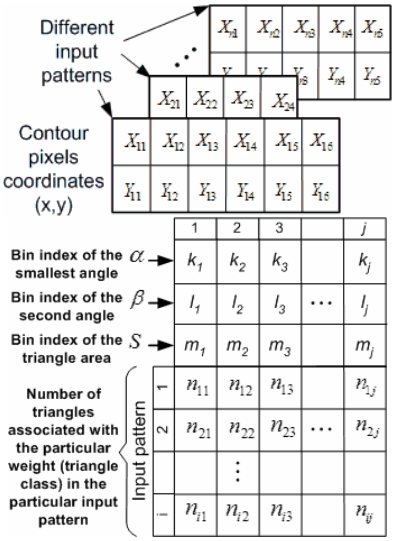


Fig.2. Arrays of the pixel coordinates of the object contours

Fig.3. General array for classifying triangles by weight index.

The presented method of classification is based on "approximately equal

triangles". For the association of a triangle with a particular weight, the sets of possible values of the two smallest triangle angles (α, β) and the triangle area (S) are partitioned into bins defined by:

$$(k-1)\,w \le \alpha < kw,\ \ (l-1)\,w \le \beta < lw,\ \ (m-1)\,s \le S < ms, \qquad (4)$$

where w is an angular tolerance, α and β are the smallest angles of the triangle so that $\alpha < \beta$, s is an triangle area tolerance, S is an area of the computed triangle, k, l and m are the bin indices associated with two angles and triangle area, respectively.

During the classification step, each triangle class is associated with a corresponding weight and is represented by three variables *k, l,* and *m.* The array of triangle classes is constructed as shown in Fig.3. After construction, only the array of triangle classes presented in Fig.3 must be stored in memory.

3.2 Network training

The previously constructed array of triangle classes (Fig. 3) is used as the basis for learning in the training phase. In addition, a zero matrix of weights (*W*) with the size of *NoP x NoW* is constructed. Where *NoW* is the number of individual weights, *NoP* is the number of training patterns.

Output computation takes into account only information presented in the weights array (*W*) and in the triangle array (*N*) (Fig. 3). It follows the next equation for a particular input image:

$$y_i = f\left(\sum_j w_{ij} n_{kj}\right), \qquad (5)$$

where *i* is the output index, *j* is the weight index, *k* is the pattern index, *w* is the weight, and *n* is the number of triangles that correspond to the particular triangle class (i.e., the particular weight).

All weights are updated only once after each iteration, according to the next equation:

$$\Delta w_{il} = \eta(t_i - y_i)\ \text{, if } N_{kl} > 0\ ;\ \Delta w_{il} = 0\ \text{, if } N_{kl} = 0, \qquad (6)$$

where t is the expected training output, y is the actual output, η is the learning rate.

After the training phase is complete, only the array of learned weights and the corresponding coefficients *k, l,* and *m* that represent the equivalence class (from the upper three rows of the array in Fig. 3) must be saved.

3.3 Recognition

The algorithm for recognition can be described as follows: 1. Load pattern intended for recognition. 2. Construct an array of coordinates of contour pixels (as in the construction stage). 3. Construct a zero matrix (*N*) with the size equal to *1 x NoW*. This is a counter for triangles in the image, which correspond to the particular weight. 4. Run through the coordinate array and compute coefficients *k*, *l* and *m* for all possible triangles as was described in 3.1. After each computation, compare the newly found *k*, *l* and *m* with the ones previously saved (upper part of the array from Fig.3). If a matching class for the triangle is found, the counter corresponding to that triangle class position is increased by one ($n_{1j} = n_{1j} + 1$) Thus, during classification the nonzero one-dimensional matrix of counters (*N*) is built. 4. Compute outputs according to equation (5), using the weights array (*W*) built during construction phase and the triangle counters (*N*) built in the beginning of recognition phase.

4. Experimental results

To study the performance of the modified network and compare computational resources with the conventional network, seven different object classes with 60 x 60 and 170 x 170 pixels were prepared. One object from each class was used in the training phase and 14 rotated patterns of each class were used in the recognition phase. Pattern examples are shown in Fig.4.

Fig.4. Pattern examples

The comparison for computational resource demands for 60 x 60 and 170 x 170 input fields are presented in Table 1.

As can be seen from the table, the gain achieved with the modified network in computational steps amount is four orders of magnitude for an input field 60 x 60 and five orders of magnitude for an input field of 170 x 170. This gain will be more significant with image size increase. In addition, the memory resources are minimized also.

Table 1: Comparison of the computational resources demands
("approximately similar triangles" scheme is used alone,
the network was trained for first five pattern classes, w = π/180, m - not used).

Input field size	60 x 60		170 x 170	
Network type	Conven-tional	Modi-fied	Conven-tional	Modi-fied
Computational steps (number)	7.8×10^{9}	13.8×10^{5}	4.02×10^{12}	4.5×10^{7}
Total memory requirements (bytes)	15.5×10^{9}	81340	8.04×10^{12}	81340

Table 2: Recognition rate for a varying number of trained classes,
angular and area similarities. Input pattern: 60 x 60 pixels.

Tolerance		Number of trained classes						Weight number
Angular (*w*)	Area (*S*)	2	3	4	5	6	7	
π/60	10	100	95	94	84	80	80	17286
π/60	20	100	95	95	85	80	80	8935
π/20	10	100	95	91	87	82	80	2502
π/20	20	100	95	88	80	-	-	1217

Table 3: Recognition rates of the net with the "approximately similar
triangles" scheme alone. Input pattern: 60 x 60 pixels.

Angular (w) tolerance	Number of trained classes				Weight number
	2	3	4	5	
π/225	100	80	75	60	8533

For comparison with the "approximately similar triangles" scheme, a few results are provided in Table 3. Results for the best configuration are shown only, but even this shows much worse recognition rates. The cause for this is that similar triangles with very large difference in size are associated with the same triangle class, as a result, some object classes will be associated with the same triangle class, preventing from the objects to have an individual triangle set associated with it.

From the experimental data provided, it can be seen that our method enables the possibility of large input field computation without significant resource demands. Translation invariance is built into the network, thus

100% translation invariance is achieved. All experimental data are provided for this particular data set. For other data sets, where object classes differ significantly in size and in form, much better recognition results can be achieved.

5. Conclusions

A modified High-Order Neural Network for efficient invariant object recognition has been presented. The proposed modification achieves significant gain in computation time and memory requirements. The gain in computation time is achieved by eliminating the idle loops, by taking a priori knowledge of training patterns. With the proposed modified HONN, large input patterns can be processed without large computation demands. Performance of the network is improved also significantly, by using the "approximately equal triangles" scheme.

References

Barnard E., D. Casasent. 1991. Invariance and neural nets. IEEE Transactions on Neural Networks, vol. 2, no. 5, pp. 498-507.

Wood J. 1996. Invariant pattern recognition: a review. Pattern Recognition, vol. 29, no.1, pp.1-17.

Spirkovska L., M.B. Reid. 1992. Robust position, scale, and rotation invariant object recognition using higher-order neural networks. Pattern Recognition, Vol.25, No. 9, pp. 975-985.

Perantonis S.J., P.J.G. Lisboa. 1992. Translation, Rotation, and Scale Invariant Pattern Recognition by Higher-Order Neural Networks and Moment Classifiers. IEEE Transactions on Neural Networks, Vol.3, No. 2, pp. 241-251.

He Z., M.Y. Siyal. 1999. Improvement on Higher-Order Neural Networks for Invariant Object Recognition. Neural Processing Letters, Vol. 10, pp 49-55.

Spirkovska L., M.B. Reid. 1993. Coarse-Coded Higher–Order Neural Networks for PSRI Object Recognition. IEEE Transactions on Neural Networks, Vol. 4, No. 2, pp. 276-283.

Spirkovska L., M.B. Reid. 1990. Connectivity Strategies for Higher-order Neural Networks applied to Pattern Recognition. Proceedings of IJCNN, Vol. 1, San Diego, pp. 21 - 26.

He Z. 1999. Invariant Pattern Recognition with Higher-Order Neural Networks. Master Thesis, School of Electrical and Computer Engineering, Nanyang Technological University, Singapore.

ONLINE GENETIC ALGORITHMS

Alfredo Milani

Department of Mathematics and Computer Science, University of Perugia
Via Vanvitelli, 1, 06100 Perugia Italy

Abstract: This paper present a technique based on genetic algorithms for generating online adaptive services.

Online adaptive systems provide flexible services to a mass of clients/users for maximising some system goals, they dynamically adapt the form and the content of the issued services while the population of clients evolve over time.

The idea of online genetic algorithms (online GAs) is to use the online clients response behaviour as a fitness function in order to produce the next generation of services. The principle implemented in online GAs, "the application environment is the fitness", allow modelling highly evolutionary domains where both services providers and clients change and evolve over time.

The flexibility and the adaptive behaviour of this approach seems to be very relevant and promising for applications characterised by highly dynamical features such as in the web domain (online newspapers, e-markets, websites and advertising engines). Nevertheless the proposed technique has a more general aim for application environments characterised by a massive number of anonymous clients/users which require personalised services, such as in the case of many new IT applications.

Keywords: genetic algorithms, adaptive web, evolutionary computation

1. Introduction

The research on the topic of adaptive systems has mainly focused on architectures based on knowledge representation and reasoning [1], fuzzy reasoning [2][3] and probabilistic models [4]. These approaches are often able to give an adequate account of uncertainty and dynamical aspects of the domain, but they also require a great effort in building a detailed model of the problem. Despite of the good qualitative response, they often reflect too rigidly the domain constraints at modelling time. When the environment, i.e. the constraint of the domain, evolves, the system performance tend to

decrease until the model needs to be modified or redesigned.

The increasing diffusion of mass services based on new information technologies (ITs) poses new requirements and goals on adaptive systems which are seemingly contradictory, such as the problem of providing adaptive personalised services to a mass of anonymous users [4]. Sometimes models of user behaviour [1] for the new services does not even exist, and, in addition, services and technologies appear and disappear very quickly thus vanishing the effort of building accurate models.

The growing interest in self adaptive and self modelling systems is partially motivated by these reasons.

The two leading approaches to self adaptation, i.e. genetic algorithms [5][6] and neural networks [7][8] are characterised by somewhat symmetrical features which are worth to be pointed out: *neural networks (NNs) tends to be online systems while GAs operate offline*. GA usually operates offlinein the sense that they can be seen as building a simulated application environment in which they evolve and select the best solution among all the generations, under the well known Darwinian principle of "survival of the fittest".

Some works [9][10][11] have introduced "real world" issues into the GA loop, in the interactive GAs approach [12] the user is inserted in the algorithm with the role of providing the fitness functions by interacting with the GA, in other works still following the offline approach [13][14] about machine learning by GA, historical real data are used as fitness function.

Despite of their offline nature GA are able of a highly dynamical behaviour. The main reason is that the knowledge about "reasoning" structure of GA is embedded in the population chromosomes: when the population evolves the structure evolves as well. GA concepts such as *cross over* and *mutation* have no counterpart in NNs approach, but they are a powerful tools which can allow a GA to make fast hill climbing of local minimum and plateau in optimisation problems [6].

The idea of bringing these adaptive features in the *online system* scenario is made more challenging from the facts that the population of clients asking for services is evolving over time, then their response to services changes.

In this paper we propose a new approach, *online genetic algorithms* (online GAs) which tries to combine timely responses with the adaptive behaviour of GAs. The basic idea of online GAs is to evolve populations by using the application world as a fitness function, under the principle “the real world is the fitness”.

The goal of systems based on online GAs is to give a timely response to a massive set of clients requesting services, and to be able to adapt services to clients, both changing over time in unknown and unpredictable way.

As noted in the beginning, it is not realistic to rely on the hypothesis of detailed user models [1][15]. The increasing consciousness of privacy issues, legal limitations on personal info [16] and the growth of mobile and pervasive interfaces accessible from casual users, often make the user model impossible to collect. The anonymity of users/clients is then a structural constraint in mass adaptive services.

In the next paragraphs we will motivate the online GAs approach by analysing the features of a sample dynamical scenario regarding an online newspaper management system.

The principles and the architectural scheme of the online genetic algorithms approach will be presented, an example application and experimental results will be discussed.

2. The Online Adaptive Scenario: Web Newspaper

Let us consider as a typical scenario for online genetic algorithms: the problem of managing the generation of an online newspaper with the goals of maximising customers, i.e. readers, contacts.

The problem, well known to journal editors, is to build a newspaper in order to publish news according to the newspaper politics and mission, and selling it at its best. Selling news in this contexts means the goal of capturing readers attention for reading the articles, and for, possibly, satisfying the newspaper advertisers. Online readers browse time by time the newspaper web site and read the news which interests them. It is assumed that a good journal will collect a great number of contacts and many users will spend time in reading it. Managing editors of online newspapers have a great advantage with respect to their hardpaper colleagues: while a conventional paper journal is limited and bounded to a single daily edition (except the cases of extraordinary events), an online

editors, instead, can make timely adaptation of the newspaper to the latest news, thus maximising the impact of the newspaper on the readers.

Online media have the likely feature that can be produced and delivered instantaneously such that, in principle, each user can read his own single, personalised and different copy of the journal.

The main issues, and source of difficulties, in the newspaper scenario are the lack of information about the users and the unpredictable dynamical evolution of all the elements which characterised it, in particular:

- anonymity of clients
- dynamical evolution of potential services
- dynamical evolution of clients
- dynamical evolution of client goals

these evolutionary features are shared by a wide class of online problems.

2.1. Anonymity of clients

Anonymity of clients means that no hypotheses can be made about profiles of the users of online services. As discussed in the introduction the typical assumption for online newspaper is that the information available to the system comes from anonymous user sessions, where users cannot be identified, nor recognised from previous sessions [16][17].

2.2. Dynamical evolution of potential services

The purpose of online systems is to provide the best of their currently available services for maximising the client impact [18], the situation is made more complex since *the services that are issued by the providers can vary over time in unpredictable way*.

News, seen as services, are characterised by a lifetime cycle (i.e. they appear, disappear and are archived), and the news flow is by its nature unpredictable. Thus the news editor task is to select according to the editorial line, which news best interest and impress their readers, among the available ones.

2.3. Dynamical evolution of clients

The set clients connected with the online system evolves over time in unpredictable way. The set of connected clients are not always the same, since new clients come and previous sessions disconnect.

In the case of online newspaper there can be made some general assumption about the target users. Users are assumed to have somewhat

homogeneous features like in the case of readers of newspaper specialised in economics, politics, sports etc. Nevertheless the instantaneous audience profile of online newspaper can vary over time. For instance students can connect mainly in the afternoon, while corporate workers can connect in different time range. In addition, external factors and unpredictable events, such as holidays or exceptional events, can make different classes of readers to connect in unexpected time/dates.

Even assuming that we have a way of determining the ideal journal for the current audience given the currently available news, the newspaper edition will be no more adequate after some time, since the audience will change unpredictably.

2.4. Dynamical evolution of client goals and attitudes

Goals and attitudes of the single clients can vary and depend on time.

As we as pointed out before, external events of general interest can make the journal audience vary, but can also make the interests of the audience to vary. Economical or political events can induce a shift in the typical interest of the readers. Moreover even assuming to have a fixed audience, with fixed goals, is not possible to produce a fixed "ideal newspapers", since people expect that newspaper vary: it would be unlikely to read every day the same identical news; typical users of online newspapers connect to the system many times a day, expecting to read more news on topics of their interest.

2.5. Model of Service Impact Factor

A model of the impact factor of service cannot be easily defined and require classification effort.

The goal of the newspaper editor is to catch the attention of most of its readers by selecting the appropriate news and preparing a suitable edition according to the newspaper editorial line, i.e. mission, policy and cultural goals.

The typical tools available to an editor to maximise the impact of the service he provides (i.e. the news) are: *selections* of the news among the continuous flow (deciding which news are currently published and which news go to archive); *location* of the news in the grid of the newspaper layout (the position of the news usually reflect is evidence or priority in editor's intention); *presentation form* of the news, which regards aspects such are selecting a *title* for the news, and or selecting a possible *picture* accompanying it, and sometime also long or short versions of the article.

These tasks are usually regarded to as an "art" which the newspaper editor performs by the help of his/her experience.

It is worth noticing that some factors, such as the news position in the layout, are not necessarily determining the readers' priority. A well prepared journal, for example, usually offers a mix of different news (i.e. not many news on the same topic). The visibility strictly depend not only in the position but also in the context in which news are presented. Sometimes hot emerging topics require breaking these rules and, when it happens most part of the journal news are devoted to a single topic.

The next paragraph will describe a framework based on genetic algorithms for providing adaptive services in highly evolutionary environment to a massive audience of anonymous users, such as in the newspapers scenario.

3. Online GA Schema

GA have been classically proposed for use in an *offline schema*. In the offline approach populations of solutions are evolved offline for a given number of generations in order to produce the best evolved solution (usually determined in the last generation) which given as system output. For instance in classical optimisation problems [6] GA are used for exploring a search space of solutions and the best minimum/maximum value found over all generation is produced. In GA applied to learning problems, such as discovering stock market rules [14], real data about stock market are used to evolve the population, but again, the best solution is computed offline, and it is used in the *real market* afterwards. A different approach is that of Interactive Genetic Algorithms [9] [12] [19] where the real world is included into the GAs loop under the principle that "the user is the fitness", i.e. the user participates to a cooperative optimisation process. In some interactive GAs applications to robot learning [13], the real world is used to evolve the solution, but GAs uses real world in an offline phase of training.

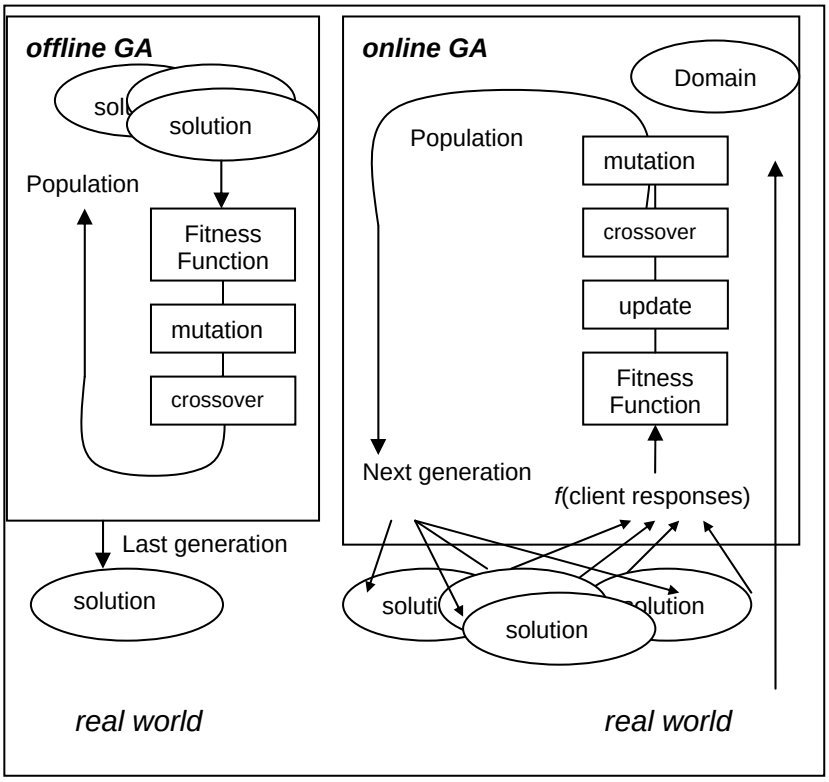


Fig.1 Online GA

In the online GAs approach we propose to literally implement the evolutionary metaphor which originally motivates GAs. In our proposal the basic elements and concepts of GAs such as population, chromosomes, crossover, mutation and selection mechanism still exists but, they are extended with the innovative but simple assumption that "*the real world is the fitness*", i.e. the application world, representing the environment, is used to selects the best surviving fittest solutions, moreover all generated solutions are output to the system and no interactive cooperation is required to users/clients.

The basic scheme of an online GAs versus offline GAs is depicted in the fig.1.

In *online GAs* a population of solutions is evolved with usual genetic

mechanisms, with the difference that the solutions in the population are actually "executed", and the client/users behaviours/responses upon solutions are used as a fitness to evolve the next generation of online solutions. In other words the fitness function resides in the real world and it is expected to give timely response to the evolution of clients' population and domain modification.

Updating Phase

Since in online GAs the problem domain also evolves unpredictably (for instance the provided services changes, i.e. the flow of incoming/outcoming news), then there is the need for a novel phase of *updating*, not present in classical GA. The purpose of the *updating* phase is to establish and implementing a policy about how to adapt the current population of solutions to domain changes (such as how to replace an article which is disappeared from the journal, because expired or delete after explicit editor's decision).

Online GAs can be used when it is required to dynamically adapt to evolution and changes in the problem domain, moreover application domains best suitable for the adoption of online GAs are characterised by:

- a solution space with "many" valid solutions to explore, i.e. the solution space with not unique or few valid solutions
- a set of clients which require solutions to be used immediately
- an optimisation function which measures the efficacy of a solution given in output, which can be "sensed" by the system in the external world as a response/result produced by clients

It is worth noticing that online genetic algorithms would not be a real possibility without the new ITs. A massive diffusion of the internet, mobile phones services, on demand phone services has made possible application servers where a huge number of anonymous clients (with no distinction among final users or software agents) are concurrently requiring services in an automated framework which directly connects consumers to service providers. The services providers are usually optimising very simple functions which are completely inside their "sensing" scope such as *time-spent*, *services bought*, *money charged*, *advertising clicked*.

4. An Online GA

The pervasive dynamical and unpredictable evolution of all the key elements in the newspaper scenario represents a difficult challenge for adaptive systems, which should provide adequate services in answers to clients' requests.

News to be offered in the newspaper is continuously flowing in from news agency and journalists. Different classes of anonymous individual readers continuously connect and disconnect in order to read interesting news The goals and interests of the individuals vary in an unpredictable way (people get bored of old news). The impact of news upon users depends of the form, the position and the context in which the news is given, and it is hard to be deterministically modelled [9][20]. Finally the editor policy represents a pervasive constraint to be respected throughout the journal editing.

In the following we present the architecture of a sample online GAs applied to the newspaper evolutionary scenario.

4.1. Domain and Constraints

A newspaper has a typical layout and structure in term of sections of topics, which are part of the recognizable corporate image. No editors are available to modify it, moreover the editor usually want to have control over the proposed news in order to implement the editorial policy.

In order to reflect these constraints the structure and layout of the journal do not evolve, and the editor decides which news include/exclude in/from the newspaper and how to assign (or remove) them to sections, let the sections, for instance: *TopStories, National, International, Sports, Health and Technology*. A limited set of headlines (for example 4 headlines) is reported in the front page for each section; the sections occupy fixed layout positions; the section headlines are chosen among the articles available inside the sections.

For each single article we will assume that the newspaper editor provide a set of possible alternative formats for each article, i.e. alternative titles, texts and pictures to be used for presentation.

The task of the editor is to decide how to update the set of news and formats, while the online GA actually build the newspapers deciding which articles will be inserted into the sections headlines and which alternative formats will be used in the articles presentation.

4.2. Population and Individuals

The individuals which compose the population of the current generation consist of the different versions of the newspaper which have been issued to the currently requesting online readers.

4.3. Time Intervals

In order to make the online GA having a sufficient number of individuals in the population, and a sufficient time to evaluate user response, i.e. fitness of the individual, it is needed to fix a time interval value, i.e. the duration of the minimal interval of time from one generation to the next one. If, for instance, a newspaper has 6000 contacts per hour, a time granularity of 1 minute guarantee, guarantee an average population of 100 individuals, but doe not allows evaluating responses whose duration is greater than one minute.

4.4. Fitness Function

The fitness function measures the adequacy of the solution in term of client response.

According to the anonymity hypothesis the system is able to "sense" user sessions, but not to recognise user from previously started session. Sensing data are easily collected from the web server log files [Etzioni2000]. In the newspaper problem the fitness of a given solution k (i.e. the individual version of the newspaper) is defined as

$$F(k) = w_s t_s + w_{ca} n_{ca} + w_{cn} n_{cn} + w_{int} (\Sigma_{(i=1..n_{cn})} t_{si}/t_r) / n_{cn} - w_{nohl} c_{nohl}$$

The listed parameters reflect the general criteria that reward as positive, in particular: t_s is the *total time spen*t on the newspaper (measured as the time between the first and the last browser request); n_{ca} the *number of clicks on* newspaper *advertising*s; n_{cn} *number of clicks on news* (i.e. how many news have been red; the Σ term computes the *average interest of news*, where the interest is measured as the time spent t_{si} on a single news with respect to the time t_r needed to read the news (skipping rapidly a news means little interest versus carefully reading it); the minus terms c_{nohl} in $F(v)$ penalises the situations in which the readers find *no interest in headlines* and go straight to sections to read particular news, i.e. in other words it penalise at a certain extent the journal versions in which the content is interesting while presentation is not. Weights w_s, w_{ca}, w_{cn}, w_{int} and w_{nohl} are used to tune the contributions of the respective terms to the

global fitness.

4.5. Chromosomes

The individuals, i.e., the single newspaper versions, are encoded by a set of *sections vectors* each one encoding a section of the newspaper.

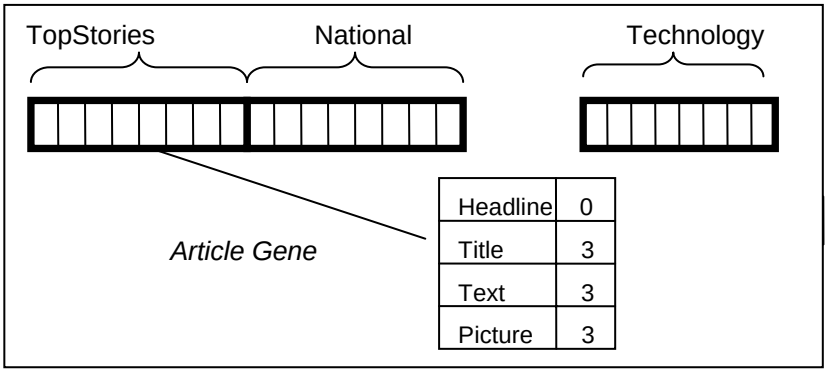


Fig.2 Individual chromosomes

Each element in the journal chromosome specifies a single news in term of its position in the section headlines (0 means not in headlines), and its presentation i.e. values indicating which title, text, and picture, the newspaper edition will contain for the given article among the different available versions.

4.6. Selection

A standard *proportional to fitness* selection method is used in order to determine the intermediate population used for crossover and mutation. The more the fitness is high more chances are given to individuals to survive. On the intermediate population thus determined crossover and mutation are applied.

4.7. Crossover

The purpose of crossover is to generate a new journal version from two individual chromosomes. The two offspring replace the parents. Again a proportional to fitness reproduction criteria is used.

The crossover is operated *section by section* on the whole chromosome. For each section a linear crossover point is determined (see dashed line in the figure below) for splitting the section subvector. The respective

subsection of the two parents is then combined.

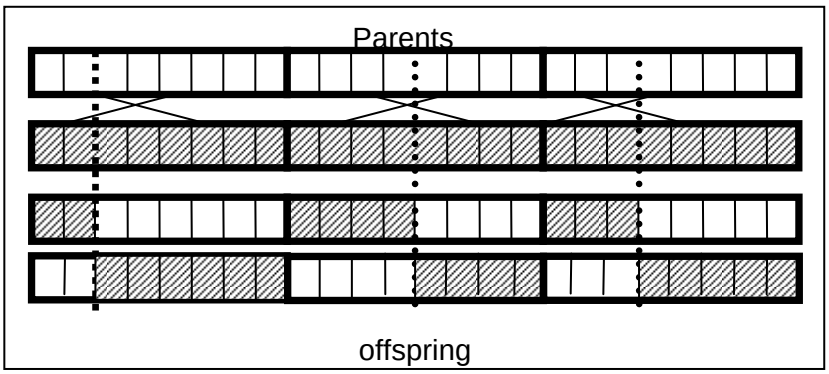


Fig.3 Segmented crossover

Restoring valid solutions can be necessary after crossover reproduction. Suppose that a given section is allowed h headlines; the split point position can divide the section segments such that one offspring segment contains more than h headlines while the other has less than h, i.e. the solution is not valid. In this case in order to restore a valid solution we move headlines from the longest to the shortest one selecting them randomly. Another case of invalid solution is when two headlines in a section points to the same position (another headline position must be empty), in this case the tie is broken randomly. Note this criteria guarantee that all headlines in the parents will be again headlines in the offspring.

4.8. Mutation

Mutations are operated at different levels with different priorities.

- *headline mutation*, is the operation which moves a news from sections into headlines and vice versa, since an headline mutation is a dramatic change in a newspaper version, the probability P_h of headline mutation is kept relatively low, on the other hand and additional factor P_{new} is considered, P_{new}, is giving more probability to become section headline to new articles versus old ones;
- *format mutation*, this mutation tends to adapt the form in which the single news are given, i.e. order, titles, alternative texts and accompanying pictures, the probability P_f of this mutation is slightly high than the previous one.

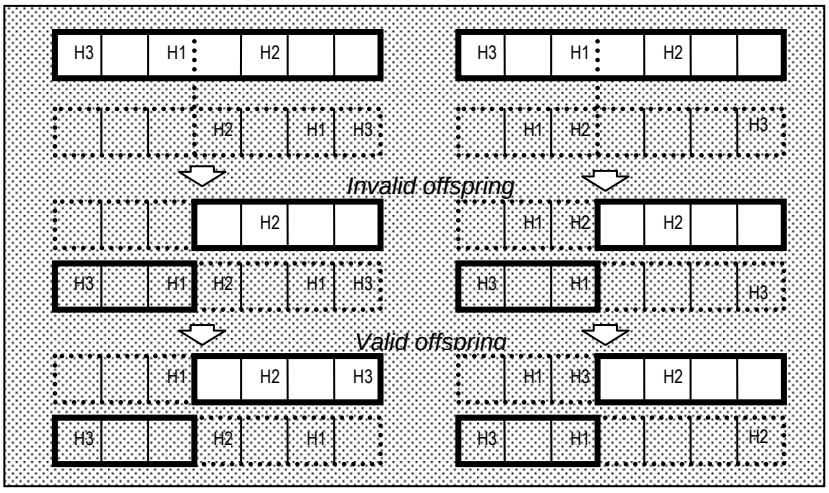


Fig. 4 Restoring Valid Offspring after crossover

A *format mutation* is realised by choosing randomly a format component (*order*, *title*, *text* and *picture*) and a feasible random value in the domain of the format component (e.g. one title over three available candidates). Format mutation of *ordering* is only applied to headline news and consists in swapping an article with another one randomly selected among the headlines.

Headline mutation is realised by randomly selecting the incoming article (taking into account of P_{new} to give priority to new articles), selecting the outcoming article and swapping them among headline and section.

4.9. Update adaptation phase

The adaptation phase concerns the problem of adapting the population of solution which were made invalid by external modifications. For example when the news editor decides that an old article has to be archived and/or a new one has to be inserted into the journal, some individual in the current population could be no more valid. In restoring the validity of the solution we use the following criteria:

- incoming news are added to the respective section with maximum P_{new}
- outcoming news not appearing in section hot headlines are simply deleted from the section
- outcoming news which are on the section headlines are replaced by shifting up the section headlines, and operating a format mutation on

the last position where simple insertion replace swapping

The array representation is updated accordingly.

5. Experimental results

The sample online GA for newspaper management described in the previous paragraph has been implemented in a simplified online version in order to manage the "What's new?" list box in our University home page, corresponding to a single section of a newspaper. The "What's new?" box is located right in the center of the home page and it contains a set of headlines which link to announcements of University activities and events. The WebMaster policy limits the number of headlines to a maximum of 8, but much more departments, administrative offices and other institutions are competing for having their announcements on the home page box. There are averages of 30 candidate announcements per day, some 22 of which are forced to lie in the internal "What's new?" section.

The online GA was operated on our university home page for two weeks. It was managed as a single section of the sample newspaper architecture presented in section 4. The fitness function was simply measured as the number of hits in one of the headlines in the "What's new?" box, plus a factor which take into account of the time spent in the headline (the more time, the more interest) the time spent in the home page before clicking on the headline link (the less the time, the more the interest), and the interest in the announcement. For each announce three possible headlines were provided to the system. The system was implemented in php in connection with an Apache server on a Linux platform.

A performance comparison have been made with the "old" static management of the "What's new?" by configuring the web server such that 50% percent of home page where served in the static version. The log files corresponding to the static pages have been evaluated by the GA fitness function.

The empirical results shown in the table below are encouraging.

Total contacts	Static WN#clicks	Static Total Fitness	OGA WN#clicks	GA Total Fitness
62538 (50% Static, 50% GA)	802 (2.56%)	1203	1679 (5.37%)	2530

Table 1. Experimental Results for WN box

The Static and GA total fitness values are the sum of the fitness values of the sessions which generates clicks on "What's new?" links. The results show that online GA performs about two times better than static management of the "What's new?" box. The similar increase from number of clicks to fitness shows that there is not a significant variation versus to interest in the two approaches: in the average 60% of clicking individuals, when they click on an announcement they are also interested in it (i.e. they read all the announcement).

Unfortunately the absolute number of hits to "What's news?" is very low since the home page is mainly used as a root to the University web site for reaching already known services or administrative information rather than being used as a source of "news".

Two further versions are under implementation: a newspaper manager, based on the content management system Nuke [21], and a simulation experiments which aim at compare the online GA with respect to simulated user response.

6. Conclusions

Online GAs represent a new approach to systems which provide adaptive services to a large number of anonymous clients/users which evolves over time in unpredictable way.

The basic idea of online genetic algorithms is that “the world is the fitness”, i.e. the fitness function resides in the application environment and it can be evaluated by sensing the environment i.e. by evaluating clients/users response to the current solutions. A phase of adaptation is added to usual GA schema for restoring validity to solutions made invalid by evolution in the problem domain.

Online GAs are related with interactive GAs methods [12][9][10], in

which the real world appear in GA in the form of the user cooperation to the selection process, or in the form of environment guided training [13]. The main difference between online GAs and interactive GAs is that in interactive GAs, GAs are used in a sort of offline simulation in order to select a *final* optimised solution or behaviour, used by the application. Instead online GAs based applications made immediate use of the solutions population.

The main issues which motivate the adoption of online GAs have been discussed in the framework of the newspaper scenario. Online GAs represents an answer in all those situation in which adaptation is required, while few o no data are available about users' profiles and attitudes [17].

The increasing diffusion of massive distributed services based on the new ITs, the increasing consciousness and laws about the privacy issues, motivates the apparently contradictory request of providing adaptive services to unknown users in dynamical domains.

Preliminary experimental results on a simplified version of the newspaper application confirm the validity of the online GAs approach.

Open theoretical and practical issues need to be further investigated in the framework of online GAs such as the problem of time granularity with respect to the time needed for fitness evaluation; defining effective methods for tuning GA parameters and weights, and discussing typical GA issues such as co-evolution [17] in the context of online GAs.

Moreover the integration between online GA and other non evolutionary techniques such as fuzzy and probabilistic analysis are worth to be investigated.

References

[1] A. Kobsa and W. Wahlster, editors. User Models in Dialog Systems. Springer Verlag, London, 1989.

[2] L. A. Zadeh: Fuzzy Sets Information and Control 8(3): 338-353 (1965)

[3] M.A.S.; Monfared, S.J. Steiner Fuzzy adaptive scheduling and control systems in Fuzzy Sets and Systems Vol.115, n. 2 pp. 231-246, 2000.

[4] J.Binder, D.Koller, S.Russell, K.Kanazawa, Adaptive probabilistic networks with hidden variables. Machine Learning, 1997.

[5] J. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, 1975.

[6] Whitley, D. An overview of evolutionary algorithms. Information and Software Technology, (2001).

[7] J. A. Anderson.An Introduction to Neural Networks,MIT Press Boston 1995

[8] M. T Hagen, H. B. Demuth, M. Beale, Neural Network Design, PWS Publishing Co. Boston 1996.

[9] T. Masui. Graphic object layout with interactive genetic algorithms. Proc. IEEE Visual Languages '92, 1992.

[10] J. G. Peñalver and J. J. Merelo. Optimizing web page layout using an annealed genetic algorithm as client-side script. In Proceedings PPSN, Parallel Problem Solving from Nature V, Lecture Notes in Computer Science. Springer-Verlag, 1998.

[12] H. Takagi, Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation in Proceedings of the IEEE01, IEEE Press, 2001.

[13] M. Dorigo and U. Schnepf, Genetics-based Machine Learning and Behaviour Based Robotics: A New Synthesis, IEEE Transactions on Systems, Man and Cybernetics, vol.23,n.1,pp. 141-154,1993.

[14] L. A. Becker, M. Seshadri, GP-evolved Technical Trading Rules Can Outperform Buy and Hold, in 3rd International Workshop on Computational Intelligence in Economics and Finance. Sept 2003.

[15] J. Kay, B. Kummerfeld, P. Lauder, Managing private user models and shared personas in Proceedings of Workshop on User Modelling for Ubiquitous Computing, User Modeling 2003.

[16] M. K. Reiter and A. D. Rubin, Crowds: anonymity for Web transactions, ACM Transactions on Information and System Security,vol.1, n.1,pp.66-92, 1998.

[17] N. Kushmerick, J. McKee, F. Toolan, Towards zero-input personalization: Referrer-based page prediction, Lecture Notes in Computer Science, vol.1892, Springer-Verlag, 2000.

[18] M. Koutri, S. Daskalaki, N. Avouris, Adaptive Interaction with Web Sites: an Overview of Methods and Techniques, in Proc. of the 4th Int. Workshop on Computer Science and Information technologies CSIT02, Patras Greece,(2002).

[19] A. Oliver, N. Monmarché, G. Venturini, Interactive design of web sites with a genetic algorithm. Proceedings of the IADIS International Conference WWW/Internet, pages 355-362, Lisbon, 2002.

[20] J.González; J.J.Merelo; P.A.Castillo; V.Rivas; G.Romero; A.Prieto. Optimized web newspaper layout using simulated annealing. In Sánchez-Andrés, Mira, editor. IWANN99, LNCS. Springer-Verlag, June 1999.

EMPIRICAL METHODS FOR DEVELOPMENT AND EXPANDING OF THE BULGARIAN WORDNET

Pavlina Ivanova, George Totkov, and Tatiana Kalcheva

Plovdiv University, 4 Tzar Asen str., 4000 Plovdiv, Bulgaria
[totkov, pavlina]@pu.acad.bg, selinashery@abv.bg

Abstract: Some basic points from the automated creation of a Bulgarian WordNet – an analogue of the Princeton WordNet, are treated. The used computer tools, the received results and their estimation are discussed. A side effect from the proposed approach is the receiving of patterns for the Bulgarian syntactic analyzer.

Keywords: Empirical Methods in NLP, WordNet

1. Introduction

WordNet is developed in the Princeton University [2,4] as a lexical database of English. The first multilingual database to realize such approach is *EuroWordNet* (EWN) ([11], [12]) consisting of eight European languages. The monolingual databases are related to the *Princeton WordNet* (PWN) (and in this way to each other) via an interlingual index (ILI).

The *Bulgarian WN* (BWN) has been developed as a cooperative task involving the Plovdiv University and the Department for Computer Modelling of Bulgarian Language at the Bulgarian Academy of Sciences (DCMB). The work is part of an EC funded project (IST-2000-29388) *BalkaNet* [7] for the creation of a multilingual lexical database (like EWN) for 6 Balkan languages (Bulgarian, Greek, Romanian, Serbian and Turkish, Czech).

2. Forming of a BWN

The main stages in the automatic creation of a BWN (A_BWN) are presented in [8]. We discuss further the tools and the results received in this process – namely the extraction of synsets from an English-Bulgarian dictionary (EBD) and the receiving of A_BWN.

Our starting point is the transformed EBD [6] with more than 160,000 entries. Each different meaning of an English word is placed on a different row. Each row contains the English word (entry) and its translation

equivalents (TE) in Bulgarian. A link is added (where it was possible) between the EBD rows and the PWN synsets (via the ILI) [9].

Each TE row may contain Bulgarian words and phrases separated with the following signs: comma, colon, semi-colon, full stop, slash and brackets. In order to receive the different synonyms from a TE row we had to differentiate the punctuation marks used as 'separators' from the ones marking some orthographical rule. E.g. in the translation of "*anticipant*" – *"човек, който чака, чакащ"*, the first comma is not a separator while the second one is.

A special tool *BWN Extractor* (BWNE) is designed for the solving of the problem. The BWNE was created to extract almost automatically meaningful rules for forming Bulgarian synsets corresponding to PWN. In the first place, the Bulgarian words in TE rows were processed by Bulgarian Morphological Analyzer BulMorph 2.0 [10] in order to get a list of their morphological characteristics (MC). As a result we received a string-pattern in which every Bulgarian word from the TE row was replaced with a special symbol(s) coding its MCs (e. g. N denotes a noun, A – adjective, V – verb, D – adverb, Vm – a verb in indicative mood, Va – the verb 'be', Vp – participle, Nc – common noun, Q – particle, etc.) The morphological alternatives (ambiguities) are separated with '|' and the results from the robust morphological analysis [10] are marked with the sign '^'.

Table 1 presents syntactic patterns (SynP), obtained with BWNE and ordered according to their frequency in the processed TE.

SynP	Noun	Verb	Adjective	Adverb	Total
Nc	10134	11	45	2	10192
Nc , Nc	4123	5	8	0	4136
A	59	2	3749	25	3835
Vm	20	2463	24	1	2508
A , A	13	0	2460	11	2484
Vm , Vm	11	2215	8	2	2236
A Nc	2070	2	36	1	2109
Nc , Nc , Nc	1009	0	2	0	1011
Nc|Vn	913	0	5	0	918

Table 1. The first 9 syntactic patterns received by BWNE

What this statistics shows is that, for example, when the TE row of an English word consists of two nouns separated by comma (Nc, Nc) in 4123 of 4136 cases (more than 99.6%) the English word is also a noun and the corresponding two Bulgarian words (nouns) are two synonyms. Only the cases when the part of speech (POS) does not match are questionable and need to be marked by expert using BWNE.

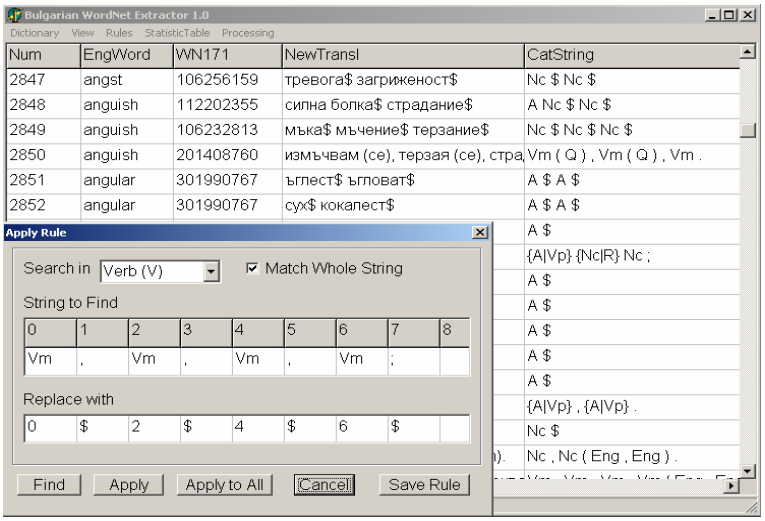


Figure 1. The Rule Editor window of the BWNE

The rules for the separation of the synonyms are based on the automatically received SynP. Moreover, BWNE provides a special *Rule Editor*. Figure 1 shows the creation of a rule to be applied on all rows corresponding to an English entry defined as 'verb'.

The functional capabilities of the *Rule Editor* are: a) automatically synthesizes rules, starting with the most likely ones; b) allows additional editing of the automatically synthesised rules; c) represents all the rows in EBD corresponding to the processed SynPs in *View* mode; d) allows changes in the respective rows in EBD in *Edit* mode; e) gives possibility for successive processing of rows from EBD (one by one or in group) in *Apply* mode; f) provides *Save Rule* mode, etc.

Experiments show that approximately 45,000 rows (TE) from the initial EBD can be automatically processed with the first 100 synthesized rules. The next 3,000 rules process additional 20,000 rows. In this way about 65,000 rows of EBD are almost automatically processed with 3,300 rules.

The extracted synonyms form A_BWN, containing about 42,000 Bulgarian synsets linked to the corresponding English synsets in PWN.

Table 2 presents 15 of the 3,300 automatically synthesized rules. Each rule consists of three parts: *POS* of the entry for whose TE a given rule is applied; *Left side* containing the (searched) string-pattern and *Right side* defining the replace string – a sequence of numbers (position of the Left side components) separated by the sign '$'. E.g. rule 15 means that 4 synonyms will be extracted in all the TE rows (for which the ILI corresponds to a 'verb') matching the pattern *verb1/verb2 noun1/noun2.* The four extracted synonyms (separated by '$') are as follows: *verb1 noun1 $ verb1 noun2 $ verb2 noun1 $ verb2 noun2$.*

Note that in rules 9-12 the comma is not (always) a separator. Its role depends from the POS of the entry – a comma followed by a relative pronoun (Pr) is a separator when the corresponding POS is A (rule 1) but it isn't when the POS is N (rule 10).

№	POS	Left Side	Right Side
1.	A	A , A , Pr Pp Vm	1 $ 3 $ 5 6 7 $
2.	A	D {A|Vp} , A	1 2 $ 4 $
3.	A	A ; R A Nc	1 $ 3 4 5 $
4.	A	Vp , A , A , R A Nc	1 $ 3 $ 5 $ 7 8 9 $
5.	D	D , D , R Pd {A|Nc}	1 $ 3 $ 5 6 7 $
6.	D	R A Nc / Nc	1 2 3 $ 1 2 5 $
7.	N	A / Vp Nc	1 4 $ 3 4 $
8.	N	A Nc , {An|D} Nc , {An|D|Nc}^	1 2 $ 4 5 $ 7 $
9.	N	An Nc , Vp R A Nc	1 2 3 4 5 6 7 $
10.	N	Nc , Pr Vm / Vm	1 2 3 4 $ 1 2 3 6 $
11.	N	Nc , R Pr Q Vm Nc	1 2 3 4 5 6 7 $
12.	V	Vm (Nc , Nc , {Nc|Np}) ; Vm	1 2 3 4 5 6 7 8 $ 10 $
13.	V	Vm (Q) , Vm (D)	1 3 $ 1 $ 6 7 8 9 $
14.	V	Vm , Nc Va R	1 $ 3 4 5 $
15.	V	Vm / Vm Nc / Nc	1 4 $ 1 6 $ 3 4 $ 3 6 $

Table 2. Rules for the extraction of Bulgarian synonyms

3. Evaluation of the A_BWN

In order to validate the A_BWN we used BWN prototype[1]. The presented result is for an A_BWN consisting of 39,109 Bulgarian synsets and containing 9,936 (common) ILI with the BWN prototype.

Let denote the number of the common literals (different words and phrases in a synset) with E, the number of the A_BWN literals –with F and the number of the A_BWN literals in the intersection – with P[2]. In order to estimate the A_BWN we use two measures:

$$\text{Precision} = \frac{P}{F} \text{ and } \text{Recall} = \frac{P}{E}.$$

The number of literals in the BWN prototype is 18,520 and in the A_BWN – 21,302. The average number of literals in a synset is 1.864 and 2.144 respectively. The number of literals common to A_BWN and the BWN prototype is 9,449. The number of synsets common to A_BWN and the BWN is 9,936. The average number of common literals in a synset is 0.951. The *Recall* is 51.02% and the *Precision* is 44.36%.

The new synsets in A_BWN (more than 33,000 additional ILI) give opportunity for further expanding of the BWN prototype.

4. Receiving of syntactic patterns

A side effect of the proposed approach is the receiving of syntactic patterns for 4 phrase types in Bulgarian: NP (noun phrase), VP (verb phrase), AP (adjective phrase) and AdvP (adverbial phrase). For example Table 3 presents the first 4 (applied) rules for A (see Table 2).

№	POS	Right Side
1.	A	A $ A $ Pr Pp Vm $
2.	A	D {A|Vp} $ A $
3.	A	A $ R A Nc$
4.	A	Vp $ A $ A $ R A Nc$

Table 3. The (applied) rules 1-4 from Table 2

In fact the received SP for the structure of AP in Bulgarian:

[1] The prototype, containing 15,007 Bulgarian synsets, is created (manually) by the DCMB experts.

[2] The literals that don't match the literals in the BWN prototype are not necessarily "incorrect".

AP := A | Pr Pp Vm | D {A|Vp} | Pr Q Vm | R A Nc | Vp

has to be checked by expert.

The first 10 SP (with greatest frequency) are presented in Table 4.

The experiments show that in this way we define some meaningful rules for the structure of NP, VP, AP and AdvP. The most frequent patterns are most likely to produce correct rules. Using the proposed approach we received 1762 syntactic patterns for the Bulgarian phrases: 1470 for NP, 175 – AP, 169 – VP and 79 – AdvP.

№	SyntacticPattern	NP	VP	AP	AdvP	Total
1.	Nc	**10744**	3	26	2	10775
2.	A	57	0	**3786**	14	3857
3.	A Nc	**3553**	1	0	3	3557
4.	Vm	10	3435	34	1	3480
5.	{Nc|Vn}	**1328**	4	3	0	1335
6.	Vm Q	0	**958**	2	0	960
7.	Vp	39	0	**887**	7	933
8.	Nc^	**871**	1	1	0	873
9.	Vm R Nc	1	**782**	0	0	783
10.	Vm Nc	2	**725**	0	0	727
Total		26028	8187	7336	902	42453

Table 4. The first 10 syntactic patterns

5. Perspectives

A method for improvement of Bulgarian Synonym Dictionary (BDS) and removing logical discrepancies in synonym rows is described in [3, 9]. The next step to be done is the expanding and correction of the synsets in A_BWN using the improved synsets from regular BDS [5].

A tool analogous to the *Split/Merge* program [9] is under development. The main features of the tool are: a) displaying all the synsets from A_BWN and BDS, in which a chosen word (or phrase) takes part; b) choice of an A_BWN synset to be processed; c) finding the BDS rows which are closest to the chosen synset [9].

The method for extracting syntactic patterns can be applied to *other lexical resources*, for example to Bulgarian Thesaurus [1]. Additional MCs (number, gender, definiteness, etc.) can be used for synthesis of more precise syntactic rules.

The receiving of precise syntactic patterns can be used for the almost

automatic creation of a *Bulgarian computer grammar* (including thousands of syntactic rules). The creation of the computer grammar is a crucial step towards the development of a syntactic analyzer of Bulgarian texts.

References

1. Andrejchin L. (ed.), Bulgarian Explanatory Dictionary. Sofia, Nauka i Izkustvo, 1999 (in Bulgarian).
2. Fellbaum C. (ed.), WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, London, England, 1998.
3. Ivanova P., Totkov G., Automated Improving and Forming Synsets on Conventional (non computer based) Synonym Dictionaries, Proceedings of the International Conf. Automation and informatics'2002, Sofia, 33-36.
4. Miller G., R.Beckwith, C. Fellbaum, D. Gross and K.Miller, Introduction to WordNet: an on-line lexical database. In: International Journal of Lexicography 3(4), 1993, accessible at ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps.
5. Nanov L, A. Nanova, Bulgarian Synonym Dictionary. Sofia, Hejzal, 2000 (in Bulgarian).
6. Rankova M., T. Atanasova, I. Harlakova. English-Bulgarian Dictionary. Izd. Nauka I izkustvo, Sofia, 1990.
7. Stamou S., K. Oflazer, K. Pala, D. Christoudoulakis, D. Cristea, D. Tufiş, S. Koeva, G. Totkov, D. Dutoit, M. Grigoriadou, BALKANET: A Multilingual Semantic Network for the Balkan Languages, Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, 12-14.
8. Totkov G., Towards Building Bulgarian WordNet: Language Resources and Tools, Proc. of the ICT&P'03, Sofia, 2003, 31-40.
9. Totkov G., P. Ivanova, Iv. Riskov, Automated Improving and Forming WordNet Synsets on Conventional (non computer based) Synonym and Bilingual Dictionaries. in A. Narin'iyani (ed.), Comp. Ling. and its Applications, Proc. of the Int. Workshop DIALOGUE'2003, Protvino, June 2003.
10. Totkov G., R. Doneva, Bipartite Finite State Transducers as Morphology Analyser, Synthesizer, Lemmatizer and Unknown-Word Guesser, Proc. of 2nd Intern. Seminar „Computer Treatment of Slavonic Languages" SLOVKO'2003, Oct. 24-25, 2003, Bratislava (in print).
11. Vossen P. (ed.), EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Final Document, 1998, 108p.
12. Vossen P. Building a multilingual database with wordnets for several European languages. http://www.hum.uva.nl/~ewn, 1999.

AN OPTIMAL DISTRIBUTED ALGORITHM FOR ALL-PAIRS SHORTEST-PATH

Saroja Kanchi and David Vineyard

Department of Science and Mathematics, Kettering University,
1700 West Third Avenue, Flint, Michigan 48504-4898
skanchi@kettering.edu; dvineyar@kettering.edu

Abstract: In this paper the network problem of determining all-pairs shortest-path is examined. A distributed algorithm which runs in O(n) time on a network of n nodes is presented. The number of messages of the algorithm is O(e+n log n) where e is the number of communication links of the network. We prove that this algorithm is time optimal.

Keywords: distributed algorithm, all-pairs shortest-path, computer network.

1. Introduction

In this paper we examine the distributed all-pairs shortest-path problem. The all-pairs shortest-path problem is the problem in which the shortest path between every pair of nodes in a network is determined. In the distributed version of the problem, a distributed algorithm is sought such that at the termination of the algorithm, every node knows the shortest path between any two nodes of the network. Floyd published a centralized algorithm [Floyd, 1962], which has been converted into a distributed algorithm by Toueg [Toueg, 1980]. The time complexity of this algorithm is $O(n^2)$ [Toueg, 1980].

The distributed shortest path problem and its variations have been studied because of its many applications. A decentralized algorithm for finding shortest paths in a network was presented by Abraham and Rhodes [Abram, 1978]. A distributed algorithm for finding shortest distances in an undirected graph was presented by Ravichandran, et. al. [Ravichandran, 1986] in which at the termination of the algorithm, each node contains the shortest path between itself and all other nodes. The algorithm described by Chandry and Misra [Chandry, 1982] finds shortest path from a node *i* to node *j* in a directed graph.

Determining topological properties of a network by distributed

computation have received considerable attention. A number of papers have covered the topic of finding a minimum weight spanning tree [Awerbuch, 1987], [Korach, 1984], [Garay, 1998]. The problems of leader election, counting, and related problems [Awerbuch, 1987], [Singh, 1995], [Korach, 1984], [Kutten, 1998], [Kanchi, 1993] have also been studied.

In this paper we use the solution for finding a minimum weight spanning tree for finding a time optimal algorithm for the distributed all-pairs shortest-path problem.

There has no previous distributed algorithm to find the all-pairs shortest-path in a general graph, other than the distributed version of a centralized algorithm given by Floyd [Floyd, 1962]. Therefore the idea of using a spanning tree and the center of the tree to find all-pairs shortest paths is a new element in this algorithm.

2. Model

The distributed network is considered to be an undirected weighted communications graph $G=(V,E)$, with processors forming the nodes, V, and bidirectional weighted communication links between processors forming the edges, E of the graph. No processor knows the topology of the network. No common memory is shared between processors and there is no global clock. All processors have unique identities from a totally ordered set. No processor knows the identity of any other processor. Each processor knows the links incident to it. For the duration of the algorithm, the network is assumed to be reliable, i.e., there will be no failures of the nodes or links for the duration of the algorithm.

The local computation at any node is assumed to take negligible time compared to the time required to transmit a message along a link. The asynchronous nature of the network permits undetermined communication delays in the delivery of a message. However, for the purpose of determining the time complexity, we assume that each message is delivered in $O(1)$ time along a link, irrespective of the size of the message. The correctness of the algorithm does not depend on this assumption.

The algorithm we present does not depend on any initiator node(s). At any time, one or more nodes may wake up and begin the execution of the algorithm. At the end of the algorithm, all nodes know the shortest path between any two nodes of the network. This data is stored in a square matrix, D, where entry (i,j) contains the shortest path between the nodes i and j.

Given any spanning tree *T* of a graph *G*, the edges that are not in *T* are called the *co-tree edges* with respect to *T*.

The size of the set *V* is denoted by *n*. The size of the set *E* is denoted by *e*.

3. Informal Description of the Algorithm

In this section we describe the algorithm at a high level. The algorithm consists of four steps as described below.

Step I: Finding a spanning tree, *T*, of the weighted graph, *G*:

Initially, all nodes are *inactive*. The first major part of the algorithm is to find a spanning tree, *T*, of the underlying unweighted graph. This can be accomplished by any one of the spanning tree finding algorithms, and we use the algorithm given by Awerbuch [Awerbuch, 1987], which takes time *O(n)*.

The spanning tree algorithm ensures every node can identify the links incident on it as either an edge in the tree *T* or a co-tree edge with respect to *T*.

Step II: Each node determines the identities of its neighbours in the graph *G*:

Each node must determine the identities of its neighbours in graph G. This can be accomplished by each node sending its identity along each link incident to it. The time complexity of this step is *O(1)*. Since each link carries exactly two messages, one from each of the incident nodes to the link, the number of messages is *2e*.

Step III: Determination of the All-Pairs Shortest-Distance matrix *D*:

This step of the algorithm deals with the transmission of distance information in *G* along the tree edges of *T*. Initially, each vertex constructs a local distance matrix that has row and column labels corresponding to the vertex and its neighbours.

Starting at each leaf node, partial distance information is transmitted along the tree edges of *T*. Whenever the partial distance matrix of a neighbour is received at a non-leaf node, new columns and rows are added to the partial distance matrix of that node and existing distance data is updated. When a non-leaf node receives partial distance matrix information from all but one of its neighbours, it becomes a transmitting node and sends its partial distance matrix to the neighbour from which it did not receive a

partial distance matrix message.

At the end of this step, exactly one or two nodes, called the saturated node(s) of the tree, would contain Shortest-Distance matrix, *D*, of the entire graph *G*. We will show that the time complexity of this step is *O(n)*.

Step IV: Communicating the All-Pairs Shortest-Distance matrix *D* to every node:

This communication originates at the one or two nodes that are described in Step 3, and messages travel using only tree edges of *T*. This step has complexities of *O(n)* time and *O(n)* number of messages.

4. Notation Used in the Algorithm

Messages transmitted in this algorithm are of the following three types:

IDENTIFICATION: This type of message is used in Step II, where each node transmits its unique identity to each of its neighbours in the graph *G*.

PARTIAL DISTANCE MATRIX: This type of message in used in Step III, where the partial distance matrix calculated locally at a given node is sent along a single tree edge.

FINAL DISTANCE MATRIX: This type of message is used in Step IV, where the final distance matrix is sent to all the tree neighbours.

The nodes are in one of four states throughout the execution of the algorithm.

INACTIVE: Nodes are in Inactive state prior to the start of the algorithm. Initially all nodes are Inactive.

RECEIVING: Any non-leaf node that is receiving and processing partial distance matrices from other nodes is said to be in Receiving state. A node in Receiving state has not yet transmitted its partial distance matrix.

TRANSMITTING: A node is in Transmitting state if it has received partial distance matrices from all but one of its neighbours (this is trivially true for a leaf node). A node in Transmitting state sends its updated partial distance matrix to one other node from which it did not receive a partial distance matrix.

SATURATED: A node is in Saturated state if has received partial distance matrices from all its neighbours in the tree *T*.

5. Algorithm

In this section we describe the distributed algorithm for finding the all-pairs shortest-distance matrix.

ALGORITHM (ALL-PAIRS SHORTEST-PATH ALGORITHM)

1. Every node sets its state to Inactive.
2. Construct a spanning tree, *T* of the underlying unweighted graph. Any good asynchronous spanning tree algorithm can be used. The only modification to the spanning tree algorithm, which is required for our algorithm, is that any node with a single neighbour in the tree (a leaf node) must change its state to Transmitting at the end of the spanning tree algorithm. Similarly, any node with more than one neighbour in the tree (an interior node) must change its state to Receiving.
3. Each node *i* determines the identities of its neighbours in *G* and stores identity and distance data in a matrix PD_i. For instance, a node *i* that is adjacent to nodes *j* and *k*, creates entries *(i,j)*, *(j,k)* and *(i,k)* in PD_i. The value of $PD_i[i,j]$, $PD_i[i,k]$ would be the weights of the edges *(i,j)* and *(i,k)* respectively, and the value of $PD_i[j,k]$ would be the sum of weights of the edges *(i,j)* and *(i,k)*. See INITIALIZE_PARTIAL_DISTANCE_MATRIX subroutine below.
4. Determine All-Pairs Shortest Distance Matrix *D* of the graph *G*. Each node's behaviour is determined by its state.

   For each node $i \in V$
   If the state of *i* is Receiving
   Run the subroutine RECEIVING_NODE_PROCESSING(*i*);
   If the state of *i* is Transmitting
   Run the subroutine TRANSMITTING_NODE_PROCESSING(*i*)

   As a result, at most 2 transmitting nodes will receive a message from all neighbours and are marked

   Saturated.
5. Transmit the final All-Pairs Shortest-Distance matrix to every node from a Saturated node. Any Saturated node contains the final all pairs shortest distance matrix *D*. The Saturated node(s) will create a final message consisting of *D* and send this message to all its neighbours in the spanning tree *T*. Any node in the spanning tree that receives *D* will store *D* locally and send *D* to all its tree neighbours except the tree neighbour from which it received *D*.

SUBROUTINE (INITIALIZE_PARTIAL_DISTANCE_MATRIX)

1. For each node $i \in V$
2. i transmits an Identification message containing its identity along each edge incident at i in G
3. i, upon receiving the identities of its m neighbours, creates a distance matrix, PD_i, of size $(m+1) \times (m+1)$ and assigns the values to , $PD_i[j,k]$ as given below.
   3.1. For each j, $k \in$ indexes of PD_i
   3.2. If $j == k$ then $PD_i[j,k] \leftarrow 0$.
   3.3. If $j == i$ or $k == i$ then $PD_i[j,k] \leftarrow$ weight of the edge between j and k.
   3.4. Else $PD_i[j,k] \leftarrow PD_i[j,i] + PD_i[i,k]$.
   3.5. EndFor

SUBROUTINE (RECEIVING_NODE_PROCESSING(i))

1. Let $Tnbr_i$ be the set of neighbours of node i in Tree T created in Step 2 of the all-pairs shortest-path algorithm.
2. Let *count* be the number of the partial distance matrices that i has received since it changed state to Receiving. Initially count is set to 0.
3. Let Links_Used be a vector of size $|Tnbr_i|$ of type Boolean in which all entries are initialized to False.
4. While count < $|Tnbr_i| - 1$
5. Receive message PD_j from neighbour j
6. count++
7. Link_Used[j] ←True
8. Call ProcessMessage(PD_j)
9. EndWhile
10. Set the state of node i to Transmitting.

SUBROUTINE (PROCESSMESSAGE(PD_j))

1. For each index k in PD_j
2. if k is not an index of PD_i
3. extend PD_i by one row and one column corresponding to k
4. For all indexes m in PD_i
5. Set $PD_i[k, m] \leftarrow PD_i[m, k] \leftarrow \infty$
6. EndFor
7. Set $PD_i[k, k] \leftarrow 0$

8. EndIf
9. EndFor
10. For each k, $m \in$ indexes of PD_j
11. if $PD_i[k,m] > PD_j[k,m]$
12. $PD_i[k,m] \leftarrow PD_j[k,m]$
13. EndFor
14. For each k, m, $n \in$ indexes of PD_i
15. if $PD_i[k,m] > PD_i[k,n] + PD_i[n,m]$
16. $PD_i[k,m] \leftarrow PD_i[k,n] + PD_i[n,m]$
17. EndFor

SUBROUTINE (TRANSMITTING_NODE_PROCESSING(i))

1. Node i transmits PD_i to its only neighbour in T from which it has not received a partial distance matrix.
2. If i receives another partial distance message, say from j, then i calls ProcessMessage(PD_j) and marks itself as Saturated.

6. Correctness

In this section we show that the All-Pairs Shortest-Path algorithm produces the correct result.

Lemma 1 There are at most two Saturated nodes.

PROOF: The algorithm starts at leaf nodes of the tree, and matrices are transmitted to internal nodes. Each internal node, in turn chooses the one node from which it has not received any partial distance matrix as its parent and transmits the partial distance matrix to that node. In this manner eventually the matrices reach the one or two centers of the tree. These centers become the Saturated nodes.

Lemma 2 The shortest distance between any two nodes is known to a Saturated node.

PROOF: We will prove this using induction on the number of edges in the shortest path. Any shortest path consisting of a single edge is known to the Saturated node(s), since every node, by Step 3, creates a partial distance matrix and all these matrices are transmitted eventually to the Saturated node(s).

Assume that if there are fewer than k edges in the shortest path between

two nodes, then that path is known to the Saturated node(s). Consider two nodes *x* and *y* such that the shortest path *P* between *x* and *y* has *k* edges. Let $P = (x = v_0, v_1, v_2, ...,v_{k-1}, v_k = y)$. Assume that the Saturated node(s) contains a ``path'' *P'* between *x* and *y*, but that the sum of the edge weights of *P'* is greater than the sum of the edge weights of *P*. Then the two paths must differ in at least one edge. Let (v_i, v_{i+1}) be the first edge in *P* that is not in *P'*. Note that v_i could be the same as *x* or v_{i+1} could be same as *y*. But since *P* is the shortest path from *x* to *y*, the path $(x, v_1, v_2, \ldots, v_i)$ is a shortest path from *x* to v_i. Similarly, the path $(v_{i+1}, v_{i+2}, \ldots,v_{k-1}, y)$ is a shortest path from v_{i+1} to *y*. Note that these paths must contain fewer than k edges, since *P* has *k* edges. But by the induction hypotheses the Saturated node contains the shortest path from *x* to v_i and from v_{i+1} to *y* since the number of edges in each of these shortest paths is less than *k*. Also, by Step 4 of the all-pairs shortest-path algorithm, Process_Message combines these two shortest paths to obtain the shortest path between *x* and *y*. Therefore the Saturated node must have the path *P*.

Theorem 1 The All-Pairs Shortest-Path Algorithm guarantees that all nodes in *G* know the all-pairs shortest-paths.

PROOF: By Lemma 1, there are exactly one or two Saturated nodes. By Lemma 2, a Saturated node knows the all-pairs shortest-path matrix *D*. Step 5 of the algorithm is a broadcast of this information to all nodes in the spanning tree, hence in the graph.

7. Complexity

In this section, we show that Algorithm 1 takes *O(n)* time and *O(e + n log n)* number of messages. Note that the subroutines Initialize_Partial_Distance_Matrix, Receiving_Node_Processing(i), ProcessMessage, and Transmitting_Node_Processing each perform local processing and are thus considered to take *O(1)* time.

Theorem 2 The all-pairs shortest-path algorithm terminates in *O(n)* time.

PROOF: Step 1 of the algorithm takes *O(1)* time. Step 2 of the algorithm, i.e., constructing the spanning tree, takes *O(n)* time. [Awerbuch, 1987]. Step 3 of the algorithm takes *O(1)* time, since each node sends one message on each tree link. Step 4 of the algorithm takes time proportional to the height of the tree with a Saturated node as a root. This is at most *O(n)*. Step 5 takes the same time as Step 4 since the messages travel from the root to

the leaves of the tree. The time complexity of the algorithm is dominated by Step 2, and is thus *O(n)*.

Theorem 3 The all-pairs shortest-path algorithm has *O(e + n log n)* bound on the number of messages.

PROOF: The number of messages in Step 2 of the algorithm is *O(e + n log n)* [Awerbuch, 1987]. The number of messages in Step 3 of the algorithm is 2*e* since each edge is used for exactly two IDENTIFICATION messages. The number of messages in Step 4 of the algorithm is *O(n)* because the partial distance matrices are transmitted from leaf nodes to the root of the tree (Saturated node) using only edges of *T*. The spanning tree has *n*-1 edges and exactly one message is sent along each tree edge, thus the number of messages is *O(n)*. Note that if there are two Saturated nodes, the edge between them is used twice. Similarly, the number of messages in Step 5 of the algorithm is *O(n)*. Therefore the number of messages generated by the algorithm is bounded by

O(e +n log n).

8. Optimality

We claim that our distributed algorithm is time optimal for finding all-pairs shortest-path. This follows since a solution to the leader election problem can be obtained from a solution to the all-pairs shortest-path problem with no additional communication time. For instance, each node can locally elect the node with the highest identity as the leader. Since the time optimal leader election algorithm [Awerbuch, 1987] takes *O(n)* time, our *O(n)* time algorithm for all-pairs shortest-path is also time optimal.

9. Conclusion

We have developed a distributed algorithm for the all-pairs shortest-path problem which is optimal in time and number of messages. The optimal time is *O(n)*. The optimal number of messages is *O(e + n log n)*.

Bibliography

[Abram, 1978] J.M. Abram and I.B. Rhodes, A decentralized shortest path algorithm in Proc. of the 16th Allerton Conf. on Communication, Control, and Computing (Monticello, Ill.), pp. 271-277, 1978

[Awerbuch, 1987] B. Awerbuch, Optimal distributed algorithms for minimum-weight spanning tree, counting, leader election and related problems, in Proc. 19th ACM Symp. on Theory of Computing, ACM, New York, pp. 230-240, 1987

[Chandry, 1982] K.M. Chandry and J. Misra, Distributed computation on graphs: shortest path algorithms, Comm. ACM 25, pp. 833-837, Nov. 1982

[Floyd, 1962] R. Floyd, Algorithm 97: shortest path, Comm. ACM 5, 1962

[Garay, 1998] J. Garay, S. Kutten, and D. Peleg, A sublinear time distributed algorithm for minimum-weight spanning trees, SIAM J. Comput., Vol. 27, No. 1, pp. 302-316, February 1998

[Kanchi 1993] S.P. Kanchi and J.L. Kim, Alternate algorithms for leader election on reliable and unreliable complete networks, Proc. of the sixth international conf. on parallel and distributed computing and systems p.118-121, October 1993

[Korach, 1984] E. Korach, S. Moran, and S. Zaks, Tight lower and upper bounds for some distributed algorithms for a complete network of processors, Proc. of 1985 PODC Conf., Vancouver, BC, pp. 199-207, August 1984

[Kutten, 1998] S. Kutten and D. Peleg, Fast distributed construction of small k-dominating sets and applications, Journal of Algorithms 28, pp. 40-66, 1998

[Ravichandran, 1986] A. Ravichandran, S.G. Menon, and R.K. Shyamasundar, A distributed algorithm for finding the shortest paths in an undirected graph, Technical Report CS-86-13, Department of Computer Science, Pennsylvania State University, May 1986

[Singh, 1995] G. Singh and A. Bernstein, A highly asynchronous minimum spanning tree protocol, Distrib. Comput., pp 151-161, 1995

[Toueg, 1980] S. Toueg, An all-pairs shortest-path distributed algorithm, Res. Rep. RC-8327, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., 1980

MULTI-DOMAIN INFORMATION MODEL

Krassimir Markov

Institute of Mathematics and Informatics, BAS,
P.O. Box: 775, Sofia-1090, Bulgaria;
e-mail: foi@nlcv.net

Abstract: The Multi-Domain Information Model for organisation of the information bases is presented.

Keywords: Multi Domain Information Model, Information Bases, Knowledge Representation.

1. Introduction

The "Multi-Domain Information Model" (MDIM) has been established twenty years ago. For a long period it has been used as a basis for organisation of various information bases. The first publication containing some details of MDIM is [Markov, 1984] but the model has not been fully presented till now. In addition, over the years, the model has been extended with some new concepts like "information space", "metaindex", "polyindexation", etc. which we will introduce in this paper.

The present paper aims to present MDIM as a coherent whole.

2. Information Domain

Definition 1. *Basic information element* "e" of MDIM is an arbitrary long string of indivisible information fragments (bytes in the version for IBM PC; symbols; etc.). ■

Let E_1 is a set of basic information elements:

$$E_1 = \{e_i \mid e_i \in E_1, i=1,\ldots, m_1\}.$$

Let μ_1 is a function which defines a biunique correspondence between elements of the set E_1 and elements of a set C_1 of positive integer numbers: $C_1 = \{c_i \mid c_i \in N, i:=1,\ldots, m_1\}$, i.e.

$$\mu_1 : E_1 \leftrightarrow C_1$$

Definition 2. The elements of C_1 are said to be *co-ordinates* of the elements of E_1. ■

Definition 3. The triple $S_1 = (E_1, \mu_1, C_1)$ is said to be an *information domain* of range one *(one-dimensional information space)* . ■

Remark: In the previous publications, the information domain S_1 was

denoted by D and the co-ordinates c_i were called "codes" of the corresponded information elements.

3. Information Spaces

Definition 4. The triple $S_n = (E_n, \mu_n, C_n)$, $n \geq 2$, is said to be an (*complex* or *multi-domain*) *information space* of range n iff 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 E_n and elements of the set C_n of positive integer numbers (co-ordinates of range n):

$$C_n = \{c_k \mid c_j \in N, k:=1,\ldots,m_n\}, \text{ i.e.}$$

$$\mu_n : E_n \leftrightarrow C_n \blacksquare$$

Definition 5. Every basic information element "e" is considered as an *information space* S_0 of range 0. ■

It is clear that the information space $S_0 = (E_0, \mu_0, C_0)$, is constructed in the same manner as all others:

- the indivisible information fragments (bytes) b_i, $i=1,\ldots,m_0$ are considered as elements of E_0,
- the position p_i (integer number) of b_i in the string e is considered as co-ordinate of b_i, i.e. $C_0 = \{p_k \mid p_k \in N, k:=1,\ldots,m_0\}$,
- function μ_0 is defined by the physical order of b_i in e and we have: $\mu_0 : E_0 \leftrightarrow C_0$

When it is necessary the string S_0 may be considered as a set of *sub-elements (sub-strings)* which may contain one or more indivisible information fragments (bytes). The number and length of the sub-elements may be variable. This option is very helpful but it closely depends on the concrete realizations and is considered as none standard characteristic of MDIM.

Definition 6. The information space S_n of range n is called *information base* of range n. ■

Usually, the concept information base without indication of the range is used as generalized concept to denote all information spaces in use during given time period.

4. Indexes

Definition 7. The sequence $A = (c_n, c_{n-1}, \ldots, c_1)$ where $c_i \in C_i$, $i=1,\ldots,n$,

is called *space address* of range n of an 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 co-ordinates.

Definition 8. Every sequence of space addresses $A_1,A_2,\ldots,A_k$, where k is arbitrary positive number, is said to be an (address) *index*. ■

Definition 9. Every ordered subset I_i, $I_i \subset C_i \subset N$ of co-ordinates (i – arbitrary positive number) is said to be a (space) *index*. ■

It is clear that space index is a kind of address index.

5. Polyindexation

Every index may be considered as basic information element (i.e. as a string) and may be stored in a point of any information domain. In such case, it will have a space address which may be pointed again.

Definition 10. Every *index* which point only to indexes is said to be a *metaindex*. ■

Every metaindex may be considered as basic information element (i.e. as a string) and may be stored in a point of any information domain, too. So, it will have a space address which may be pointed again, etc. This way, we may build a hierarchy of metaindexes.

Definition 11. The approach of representing the interconnections between elements of the information domains as well as between spaces using hierarchies of metaindexes is called *polyindexation*. ■

6. Aggregates

Let $G = \{S_i \mid i=1,...,m\}$ is a set of information spaces.

Let $\tau=\{\nu_{ij} \mid \nu_{ij} : S_i \rightarrow S_j,\ i=const,\ j=1,\ldots m\}$ is a set of mappings of one "main" information space $S_i \subset G$, $i=const$, into the others $S_j \subset G$, $j=1,\ldots m$, and, in particular, into itself.

Definition 12. The couple $Đ = (G, \tau)$ is said to be an "*aggregate*". ■

It is clear we can build **m** aggregates using the set *G* because every information space $S_i \subset G$, $j=1,\ldots m$, may be chosen to be a main information space.

Remark: In the previous publications, the aggregate Đ was called

generalized domain.

7. Operations in MDIM

After defining the information structures we need to present the operations which are admissible in the model.

It is clear; the operations are closely connected to the defined structures. So, we have operations with:

- basic information elements (**BIE**)
- spaces
- indexes
- metaindexes

In MDIM, we assume that **all** information elements of **all** information spaces **exist**. If for any $S_i : E_i=\varnothing \wedge C_i=\varnothing$, than it is called *empty.* Usually, most of the information elements and spaces are empty. This is very important for practical realizations.

7.1. Operation with basic information elements

Because of the rule for existing of the all structures given above we have need of only two operations:

- updating the BIE
- getting the value of BIE

For both types of operations we need two service operations:

- getting length of BIE
- positioning in the BIE

Updating, or simply – writing the element, has several modifications with obvious meaning:

- writing of a BIE as a whole
- appending a BIE
- inserting in a BIE
- cutting a part of BIE
- replacing a part of BIE
- deleting a BIE

The operation for getting the value of BIE is only one – **Read** a portion from BIE starting from given position. We may receive the whole BIE if the starting position is the beginning of BIE and the length of the portion is equal to the BIE length.

7.2. Operation with spaces

With a single space we may do only one operation – clearing (deleting) the space, i.e. replacing the all BIE of the space with ∅. After this operation the BIE of the space will have zero length.

With two spaces we may provide two operations with two modifications every:

- copying the first space in the second
- moving the first space in the second

The modifications concern the type of processing the BIE of the recipient space. We may have:

- copy with clear
- move with clear
- copy with merge
- move with merge

The "clear" modifications first clear the recipient space and after that provide copy or move operation.

The merge modifications may have two types of processing:

- destructive
- constructive

The *destructive merging* may be "conservative" or "alternative". In the conservative approach the recipient space BIE remain in the result if it is with none zero length. In the other approach – the donor space BIE remain in the result.

In the *constructive merging* the result is any composition of the corresponded BIE of the two spaces.

Of course, the move operation deletes the donor space after the operation.

7.3. Operation with indexes and metaindexes

The indexes are the main approach for describing the interconnections between the BIE.

At the first place, we may operate with and in the indexes C_i, $i=1,2,\ldots,n$

of the spaces. We may receive the co-ordinates of the next or previous empty or none empty elements of the space starting from any given co-ordinate. The possibility to count the number of none empty elements is useful for practical realisations.

The operations with indexes and metaindexes may be classified in two main types:

- logical operations
- information operations

The first type is content independent operations based on usual logical operations between sets. The difference from usual sets is that the information spaces are build by interconnection between two main sets:

- set of co-ordinates
- set of information elements

The logical operations defined in the MDIM 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 information spaces these operations have at least two principally different realizations based on:

- co-ordinates
- information elements

The operations based on co-ordinates are determined by the existence of the corresponding space information elements. So, the values of the co-ordinates of the existing information elements determine the operations.

In the other case, the values of the BIE determine the logical operations.

In both cases the result of the logical operations is any index, respectively – metaindex.

The information operations are context depended and need special realizations for concrete purposes.

The main information operation is creating the indexes and metaindexes. This may be very complicated processes and could not be given in advance. The main purpose of the MDIM is to give up possibility for access to the practically unlimited information space and easy approach for building interconnection between its elements. The goal of the concrete applications is to build tools for creating and operating with the indexes and metaindexes and to implement these tools in the realization of user requested systems.

For instance such tools may realize the transfer from one structure to

another, information search, sorting, making reports, more complicated information processing, etc.

The information operations can be grouped into four sets corresponding to the main information structures:

- basic information elements
- information domains
- information spaces
- index or metaindex structures

8. Discussion

Usually, the submission of any new information model needs to be discussed in connection to already existing models and theories. We have no place in this paper to analyze all known models. Because of this we will point only two of them we assume as more important:

- theory of the named sets [Burgin, 1984]
- relation model of Codd [Codd, 1970]

Our proposition is that the MDIM has the same and more modeling possibilities than named sets and relation model.

8.1. Theory of the named sets

For our further discussion we need some information from [Burgin and Gladun, 1989] .

If α is a relation of X with Y i.e. $\alpha \subseteq X \times Y$, $A \subseteq X$, $B \subseteq Y$ then

$\alpha(A)=\{y|\exists x \in A\ ((x,y)\in\alpha)\}$, $\alpha^{-1}(B)=\{x|\exists y \in B\ ((x,y)\in\alpha)\}$,

$\alpha|_{(A,B)}=\{(x,y)\in\alpha \mid x\in A\ y\in B\ \}$.

The empty set is denoted by $\varnothing$.

Definition B&G-1. A named by $\boldsymbol{M}$ set (an N-set) is a triple

$$\boldsymbol{X}=(X,\alpha,I)$$

where X is a set from some fixed class of sets and is called the support set of the named set $\boldsymbol{X}$. I is a set from some (may be another) fixed class of sets and is called the set of names of the named set $\boldsymbol{X}$. $\alpha: X \to Y$ is a map or a correspondence (a relation) from X to I and belongs to a given class of relations $\boldsymbol{M}$.

A name $a \in I$ is called empty if $\alpha^{-1}(B) = \varnothing$.

Named sets as special cases include: usual sets, fuzzy sets, multisets,

enumerations, sequences (countable as well as uncountable), etc. A lot of examples of named sets we may find in linguistics studying semantic aspects that are connected with applying different elements of a language (words, phrases, texts) with their meaning. [Burgin and Gladun, 1989, p.121-122].

The Theory of named sets (TNS) has been established about 1982 [Burgin 1984]. Independently, the MDIM has been developed in the period from 1980-1982 and its first publication was [Markov 1984] .

We may find many common ideas in the two approaches. Here we will point at two main characteristic of MDIM.

Proposition 1. Every information space is a named set.

Proof: By definition, the set E is the support set, C is the set of the names and μ is a function of naming.

Proposition 2. Every named set may be represented by an aggregate.

Proof: It is simple to build the named set by an aggregate using:

- two information spaces: one for the names and one for the elements of the named set,
- aggregation mapping which is identical to the named set mapping.

■

This way all possibilities of the TNS exist in the MDIM. In other hand, the polyindexation does not exist as theoretical base in the TNS. The aggregates are more general constructs than named sets. At the end, MDIM is designed to support practical realizations whereas the TNS is a theoretic logical construction for reasoning.

The conclusion is that the MDIM has the same and more modeling possibilities than named sets.

8.2. Relation model of Codd

The Cood's Relation theory [Codd 1970] is so popular that we do not need to explain it here. For our discussion we will proof one very important proposition.

Proposition 3. The relation in the sense of the model of Codd may be represented by an aggregate.

Proof: It is easy to see that if the aggregation mappings of the generalized domain are one-one mappings it will be relation in the sense of the model of Codd. ■

In the same time many possibilities of MDIM could not be represented by the relation model or this is very expensive work. Especially, the polyindexation could not be represented by relations. The representation of the information spaces of range more than three is very expensive for the practical realizations.

So, we may say that MDIM is more universal and convenient for practical realizations than the relation model.

9. Conclusion

The Multi-Domain Information Model (MDIM) for organisation of the information bases has been presented in this paper. The information structures and operations of MDIM have been presented.

The correspondences between MDIM and named sets (Propositions 1 and 2) as well as the relation model (Proposition 3) were shown. Our conclusion is that the MDIM has the same and more modeling possibilities than named sets and relation model.

At the end, we need to discuss some more general conclusions.

We consider *the real world* as a space of *entities.* The entities are built by other entities, connected with *relationships*. The entities and relationships between them form the internal *structure* of the entity they build. To create the entity of a certain structural level of the world, it is necessary to have:

- the entities of the lower structural level;
- establishing of the forming relationship.

The entity can dialectically be considered as a relationship between its entities of all internal structural levels. [Markov et al 2003]. Every entity may be considered as relationship between "*atoms*" which are entities on the lowest structural level where there exists another relationship and so on.

This way we may distinguish three types of relationships: explicit (forming relationships), implicit (forming relationships at lower levels) and mixed (in case we distinguish the relationships from lower levels as elements of the forming relationship of given level).

In our model, the information atoms are the basic information elements. It is easy to see that they may contain more complex structures such as domains, spaces, generalized domains, indexes, metaindexes, etc.

This means: *the complexity of the real word can be reflected by the complexity of the MDIM realizations.*

This inference gives us one very fruitful idea – to use MDIM as a model for memory structuring in intelligent systems [Gladun 2003].

Finally, we need to point out that for more than twenty years the MDIM realizations have shown the power of this model. The concrete systems based on MDIM information bases now work on more than one thousand installations all over the Bulgaria.

Acknowledgments

Author is indebted to Ms. Krassimira Ivanova and Mr. Ilia Mitov for support and collaboration during the latest 15 years. Due to theirs hard work the model presented in this paper has been implemented in practice.

Special thanks to Mr. Boicho Kokinov for fruitful remarks and advices.

Author owes a dept of gratitude to Prof. S.S. Lavrov for his important support and recommendations given in the beginning of work presented in this paper.

Bibliography

[Burgin 1984] Burgin M.S. Named Sets and Information representation. VII Allunion conference on mathematical logic, Novosibirsk, 1984. (in Russian)

[Burgin and Gladun, 1989] Mathematical Foundations of Semantic Networks Theory. Lecture Notes in Computer Science, No. 364, pp. 117-135.

[Codd 1970] Codd E.F. A Relation Model of Data for Large Shared Data Banks. CACM 13, No.6 (June 1970).

[Gladun 2003] Gladun V. Intelligent Systems Memory Structuring // International Journal: Information Theories and Applications.-2003, V.10, №1.

[Markov 1984] Kr.Markov. A Multi-domain Access Method. // Proceedings of the International Conference on Computer Based Scientific Research. Plovdiv, 1984. pp. 558-563.

[Markov et al 2003] Kr.Markov, Kr.Ivanova, I.Mitov. General Information Theory. Basic Formulations. // FOI-COMMERCE, Sofia, 2003.

INFORMATION SECURITY OF ARCHIVED OBJECTS

Dimitrina Polimirova–Nickolova

National Laboratory of Computer Virology – BAS
1113 Sofia, Acad. G. Bonchev Str., Block 8, Office 104
poly@nlcv.bas.bg

Abstract: Six basic types of archiving programs are described in the paper, as well as their advantages and disadvantages with respect to the information security. Analysis and appraisal are made of the results obtained in experiments, related to the use of encrypting mechanisms before and after the archiving process.

Keywords: Web Security, Mail Security, Information Security, Archive Programs, Compressed Objects, Methods Of Encryption.

The present situation

In the development of the computer science the creation and the use of archived objects is a classical research problem, which has found different resolutions for decades past. Nowadays the availability of several dozens of methods and their varieties represent an excellent demonstration of the ambitions of the information systems' programmers and designers for a real high-speed and high-effective compression of information flows.

The following basic types of archiving programs could be defined with respect to the information security of compressed objects, obtained after examination of more than 320 archiving programs, known by now:

1) E-mail archiving programs – in this kind of archiving programs the relative homogeneity of the information flow (e-mail traffic) is used and the most suitable methods of compression are selected. There are some differences among the basic existing e-mail clients (MS Outlook, MS Outlook Express, Netscape Mail, Opera Mail, Eudora Mail, Pegasus Mail etc.), which make possible the applying of different realizations of the compressing process. The advantages consist in the multiple reduction of the saved e-mail folders' volume and in the high degree of security against unauthorized access (viruses, worms, spyware, malware etc.). The disadvantages above all are related to the consumption of computing resources to realize the right and the reverse transformation.

The basic six extensions and their corresponding applications that are

characteristic for this type of archiving program are:

Extension	Program / Information
DBX	Outlook Express Email Folder
IDX	Outlook Express Mailbox Index
PCE	Eudora Mailbox Name Map
MSG	Pegasus Mail Stored Messages to Be Sent
SNM	Netscape Mail Email Message File
BOE	Outlook Express Backup File

2) Converting archiving programs – these are archiving programs, that have the possibility to transform objects compressed by a given method in objects compressed by another method. Two variants exist with regard to this transformation: a) without a restoration of the object in its initial appearance; b) with a restoration of the object in its initial appearance. Their advantages consist in the use of a compression method, which is optimal for a given type of information (e.g.: .jpg, .gif, .doc, .xls, .ppt etc.). Their disadvantages reside in the high complexity of operating environment.

The basic six extensions and their corresponding applications that are characteristic for this type of archiving program are:

Extension	Program / Information
ACE	WinAce Compressed File
RAR	WinRAR Compressed Archive (RarLab)
ZIP	Compressed Archive File
AIN	AIN Compressed Archive
GZIP	GNU Zip Compressed Archive
UC2	Compressed File

3) Multiple archiving programs – these are programs which perform some successive kinds of archiving processing on the object for compression by using several methods of compression differing by their characteristics. In this manner, a different (fully optimized) method of compression could be applied for the different parts of the object. The advantages lay in the very high flexibility, functionality and adaptivity to the different parts of the compressed objects, which differ by their internal structure. The disadvantages are connected to the high initial expenditure needed for the creation of library of modules for similar methods of compression and the realization of a relevant environment, suitable for analysis of the separate parts of the objects.

The basic six extensions and their corresponding applications that are characteristic for this type of archiving program are:

Extension	Program / Information
ARJ	Robert Jung ARJ Compressed Archive
JAR	JAR Archive (ARJ Software, Inc.)
TAR	Tape Archive File
AI	Ai Archiver Archive
LHA	Compressed Archive File
ZOO	ZOO Compressed Archive File

4) Image archiving programs – this is an extremely live problem in the present-day real-time processing of video and image web-objects. The predominating trend in this processing is the obligatory compressing of the object immediately after its creation. The transmission and the processing of the object are fully realized in a compressed state to the last moment of its reproduction on the relevant media. The advantages consist in the significant reduction of the objects' dimension and the time needed for transmission, retransmission and processing. The disadvantages are connected to the high expenditure for the hardware components, which realize the compression partly or fully. A reasonable compromise in this respect is the combined (software-hardware) methods of compression.

The basic six extensions and their corresponding applications that are characteristic for this type of archiving program are:

Extension	Program / Information
AIS	ACDSee Image Sequence File
B&W	Image Lab
BIL	ArcView Image File (ESRI)
BIN	Micrografx Designer 7 Project Image
CPT	Corel Photo-Paint Image (Corel)
PDB	PhotoDeluxe Image (Adobe)

5) Data archiving programs – these are programs specialized in the creation, the processing and the use of compressed objects which result from information flows owning "data" characteristics. In the different platforms and operating systems the notion "data" has a different sense. In this instance we are concerned only by the fact, that the data in the different phases of their existence pass in a compressed form, exist for a fixed time

in this form and a little time before to be "processed" the compressed objects are decompressed. The advantages lay in the reasonable degree of the optimal use of the resources. The disadvantages consist in the "superfluous" operations for compression and decompression.

The basic six extensions and their corresponding applications that are characteristic for this type of archiving program are:

Extension	Program / Information
DOC	Word Document (Microsoft)
PDF	Acrobat Portable Document Format (Adobe)
TXT	Text File
XLS	Excel Worksheet (Microsoft)
XML	Extensible Markup Language File
PPT	Power Point Presentation (Microsoft)

6) Executable archiving programs – the aim of these programs is to accomplish some specificity of the compression, connected with the possibilities for running the compressed objects. These are active objects, which own the capability for algorithmic branching of events depending on the used scenario. The advantages are connected to the extremely precise use of computing resources and the very high degree of protection against "reverse engineering". The disadvantages consist in the dependence from the platform, the operating system, the applications on use and the human factor.

The basic six extensions and their corresponding applications that are characteristic for this type of archiving program are:

Extension	Program / Information
EXE	Executable File (Microsoft)
PE	Portable Executable File
PL	Linux Shell Executable Binary
FOX	FoxBase/FoxProt Executable File
FMX	Oracle Executable Form (FRM)
XXY	SPARC Executable Script File

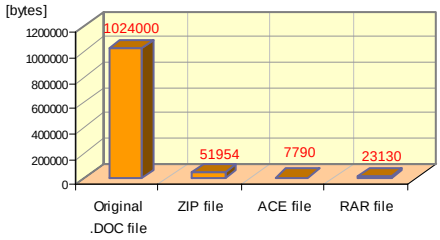
The problem

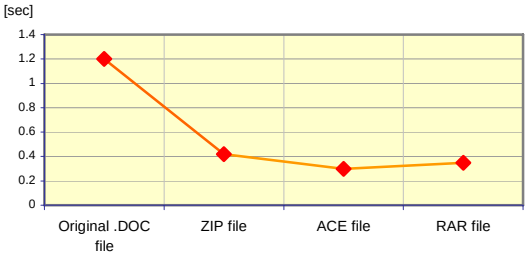
The protection of the information is accomplished by data encryption. The data encryption is a process in which the contents of a message or file are entangled to such a degree that it becomes unintelligible for nobody. To make possible the decoding of the message or the reversing of the file in its

initial state, it is necessary to own some key or access code. This conception is similar to the one for data compression. Thus, two different goals could be reached in actual fact by using the same approach:

1) The reduction of the size, which is accomplished by the data encoding during the compression.
2) In the case of encrypting, the data encoding aims to make the data unreadable.

The results of the experiments that were carried out will be shortly exposed to encourage the achievement of better effects on the enhancement of the security of objects, especially for compressed objects. Their goal was to examine and analyze the combination of data compression and data encryption [1].





The first study analyzes the SPEED of the encoding process. With regard to this examination, the following tasks were put:

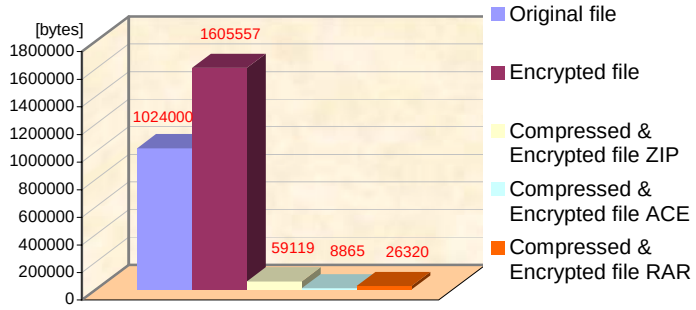
1) Valuation of the resulting files, compressed with popular archiving programs. Different file formats are used in the examination: totally 18 specific file extensions (3 for each type) regarding information security are chosen for the six types of archiving programs. Particular

experiments are made and results are obtained for all 18 extensions. Because of the lack of place here are shown only the results for the most used file extension .DOC.

2) Encryption of the objects before the compression.
3) Encryption of the objects after the compression.
4) Comparison of the time period needed for encryption of the objects before and after the compression.

The second study analyzes the SIZE of the object created after the encryption. In this connection, the following tasks were set:

1) To encrypt objects with different file formats. In this case, all the 18 specific file extensions are used.
2) To encrypt the files compressed during the first study whose extensions are among the chosen 18 specific ones.
3) To compare the size of the original and the encrypted files. Only the results from the comparison between the size of the original and the encrypted .DOC file and from the comparison between the size of the original and the encrypted after compression .DOC file are shown.



The following assessments could be made from experiments, which were carried out:

1) The speed of the encoding process is higher if the object has been compressed before the encoding. This is due to the decrease after the compression of the entire amount of information for encoding.
2) The size of the resulting file decreases, if it is first compressed and encrypted after that. In many cases if the object is encoded without

compression, its size is increased.

3) Some future investigations could be made in connection with the size of the password used in the encryption process and the effect of the passwords on the process of compression [2, 3, 4].

Conclusions

A thorough examination of the influence of some chosen parameters of the information security on the methods for compression of objects is required.

It is necessary also to create a set of criteria for appraisal of the different commercial compressing and archiving programs in connection with the information security.

Bibliography

[1] Alistair Moffat, Andrew Turpin, Compression and Coding Algorithms, Kluwer Academic Publishers, 2002

[2] Ed Skoudis, Counter Hack: A Step-by-Step Guide to Computer Attacks and Effective Defenses, Prentice-Hall PTR, 2002.

[3] K. Jallad, J. Katz, and B. Schneier, Implementation of Chosen-Ciphertext Attacks against PGP and GnuPG, Information Security Conference 2002 Proceedings.

[4] Rob Shimonski, Introduction to Password Cracking, IBM Developer Works Hacking Techniques article, July 2002.

DEVELOPMENT AND TRENDS OF PRIVATE INFORMATION RETRIEVAL

Veselina Jecheva[3]

Bourgas Free University,
8000 Bourgas, 101 Alexandrovska Str.,
e-mail: vessi@bfu.bg

Abstract: The present paper introduces private information retrieval as an approach to achieve user data privacy when a public database is accessed. It considers a detailed survey of its development and current trends.

Keywords: private information retrieval; database security; communication complexity.

The protection of the private user data in the open network communication is a critical problem in the contemporary world. Query execution to public access databases runs some risks for data security during the query execution as well as during the communication between the client and the database server. The purpose of the private information retrieval (PIR) is the user to be able to retrieve some data from a database server keeping the content of the data secret from the server.

Formally the PIR problem could be defined using the more generalized problem of private retrieving of the *i*-th bit from an *N*-bit string, stored at the server. The "privacy" feature means that the number *i* of the bit remains secret from the server, i.e. it does not receive the information which bit the user is interested in.

The private user data security and the user preferences security in particular is information of great importance in the contemporary systems. This private data could be used against the user, so it confidentiality has to be preserved.

During many years information security relied on the assumption that the user data should be kept confidential for everyone except the server. The

[3] The author is a PhD student at National Laboratory of Computer Virology, Bulgarian Academy of Sciences.

server, respectively its owner, was treated as a reliable part which would not misuse the user data received in a communication. Obviously all servers could not be treated as reliable parts, since even the biggest portals admit security breaches and user information disclosure. In many cases the companies consider their databases, containing millions user profiles and shopping preferences, to be among their valuable assets which could be a subject of a commercial deal without the users' permission. And finally, the company may be forced to sell its database because of bankruptcy.

The user preferences preservation in the contemporary systems depends on the owner's reliability and financial state, as well as on the server's security level. The solutions of the PIR problem should guarantee user data privacy from any part of the communication process, including the database server.

Examples. Let's consider some application examples for PIR problem, as well as some trivial approaches.

1. Electronic publications, music, video, pictures, etc. database. These kinds of databases are usually hosted at commercial servers, which could not be considered to be reliable enough. They should afford an opportunity to the user to hide its identity and private data using the PIR approach.

2. Medical database. Some unethical employers could be interested in how often a potential employee's medical records have been accessed, since frequent access might indicate health problems and eventually expensive treatment.

3. Pharmaceutical database. Pharmaceutical companies are specialized in drugs invention and production, as well as in gathering information about drugs basic components and its properties. The process of synthesizing of a new drug requires information about many components from different databases. The company could be forced to buy some competitions' entire databases in order to hide its confidential future plans. These expenses could be avoided by PIR protocol application in order to buy only the necessary information in a private way.

4. Cartography database. A user of such a database is the Special Operations Department of the ministry of defense in planning of operation in a given region. The administrators or the IT department managing the map database can gather some information about the region of interest, which may cause some secure information flow. When using PIR protocol the private information could remain secret for everyone except the authorized

personnel.

Oil companies are common users of a map database also, so they probably would rather their competitors not to be familiar with their latest drilling locations.

5. Patent database. These databases are usually used by scientists, inventors and potential investors. Information security violation at the server-side could cause the user a lot of problems. Let for example a scientist has discovered a great invention and would like to patent it. He has to ascertain that this invention has not been patented yet and sends a query to the patent database. The server administrator has access to its query and the positive or negative answer. In both cases the information is critical for the user and has to remain private. PIR solves this problem: the user receives the information about the patent he is interested in, but the server does not obtain any information about the data downloaded.

The following trivial approaches could be applied for PIR problem solving:

- ✓ **Entire database transfer.** From theoretical point of view, the entire database transfer solves the PIR problem: the client can process all necessary queries on its local copy of the database when all the parts, including the server, are unaware of the queries content, and consequently, stay unaware of the user private data.

  Drawback of this approach: The expenses of the entire database download depend on the database cost and the cost of the transfer from the server to the client. Usually the database cost exceeds many times the transfer cost and is too expensive for the common user to pay. So this approach cannot be practically applied in the general case.

- ✓ **Anonymous queries.** There are various practices for user transfer anonymization, as well as payment systems anonymization. These methods do not solve the PIR problem, since the server can still gather some statistical data about user queries (the most accessed record, how often a specific record has been accessed at a given time interval, etc.).

  Drawback of this approach: Most anonymization techniques depend on the third trusted party[4], which leads us back to the PIR problem of

[4] http://www.anonymizer.com, http://www.iprivacy.com, http://www.privatebuy.com

achieving user data privacy.

PIR approaches. PIR problem has been defined for the first time in 1995 in [5]. Chor et al. describe a method that enables the user access to *k* replicated copies of a database *(k≥2)* and privately retrieve information stored in the database. This means that each of the database copies gets no information about the item retrieved by the user. For a single database, achieving this type of privacy requires communication the whole database.

A set of papers has been published since the PIR problem was initially formulated in 1995. We will consider and classify some results, accordingly to the assumptions that authors rely on in these papers.

1. Theoretical PIR. This approach works on the assumption that user data privacy cannot be violated independently from the computational power of the cheater. Chor et. al prove that each theoretical solution of the PIR problem requires communication with the server with a lower bound equal to the database size [5]. Consequently the optimal solution with respect to the communication amount, i.e. communication price, is downloading the entire database. Such a solution is referred to as *trivial*. Each non-trivial PIR solution includes communication amount less than the database size.

Chor et al. consider non-trivial solution of the PIR problem in the case of availability of k replicated copies of the database, when each database copy cannot communicate to each other, as well as the case where t servers can communicate to each other and are allowed to cooperate against the user.

Based on the results described, Ambainis [1] finds the following non-trivial PIR solutions:

- ✓ A scheme with *k (k≥2)* identical database copies, non-communicating to each other – a non-trivial PIR solution with communication complexity $O(N^{\frac{1}{2k-1}})$.
- ✓ A scheme with $O(\log N)$ identical database copies, non-communicating to each other – a non-trivial PIR solution with communication complexity $O(\log^2 N \log\log N)$.

Finding a non-trivial PIR solution has been considered in a set of articles since 1997.

PIR of blocks. This problem is a PIR problem extension where records of

the database have been considered as blocks, each consisting of fixed number of bits. The purpose is a data block download instead of single bit. Theoretical PIR of blocks has been introduced in [5] for the first time and has been investigated in [4]. It is important for practical PIR applications to find solutions of the PIR of blocks.

2. Computational PIR. In this case Chor and Gilboa [3] base their work on the following assumption in order to decrease PIR traffic: the database servers are presumed to be computationally bounded, i.e. under an appropriate intractability assumption, the databases cannot gain information about *i*. For each $\varepsilon > 0$ [3] suggests a solution of the PIR problem in the case of two databases with communication complexity $O(N^{\varepsilon})$.

In [11] Ostrovsky and Shoup propose a PIR protocol, allowing the database servers to store the *i*-th record in the database in a way that they do gain information about *i*. The protocol described solves both the theoretical and computational PIR problem in the case of two or more servers.

Computational PIR in the case of single database. Chor et al. prove in [5] that theoretical PIR problem has no non-trivial solutions in the case of single database. The replacement of the information-theoretical security assumption with the intractability assumption described in [8], allows achievement of non-trivial solution, i.e. PIR problem solution in the case with single database [9]. This protocol has communication complexity $O(N^{\varepsilon})$ for each $\varepsilon > 0$. It is based on the following approach: encryption of the user query in such a way, that the server can still process it using specific algorithms and returns the result to the client. The server does not gain access neither to the original clear query, nor the result which can be decrypted only by the user.

Using another intractability assumption [2] Cachin et al. describe protocol for computational PIR including polylogarithmic communication complexity. This protocol has improved communication complexity compared to the polynomial complexity of the protocol proposed in [9]. This result is effective because the user has to send *logN* bits to the server in order to address the *i*-th bit independently from the protocol includes the feature of keeping the data privacy or not.

3. Symmetrical PIR. The purpose of the symmetrical PIR is the database privacy, as well as the user data privacy. This method has to guarantee the user will not receive more data than it has requested during the session with the database server. The symmetrical data privacy is a database security, which is very important feature of the contemporary applications from practical point of view. Symmetrical PIR solution for single database was considered for the first time in [9], and other symmetrical PIR protocols were described in [10].

4. Hardware PIR. Smith и Safford [12] consider the PIR problem in the case of single database, based on the assumption that all network traffic passes through a special device, which guarantees data privacy. This device is referred to as secure coprocessor, installed on the database server. The client encrypts its query and sends it to the coprocessor which decrypts the query, executes it, encrypts the result and returns it back to the user. The server gets no access to the query because the major property of the secure coprocessor is its built-in RAM which cannot be accessed from anyone except the coprocessor itself. Besides that the coprocessor reads all database content and the server gets no information about the result sent to the client.

5. Other PIR approaches. The main purpose of most examined works was optimization of the traffic between the client and the server, since communication is considered to be the most expensive resource. The most suggested solutions include amount of calculations executed by the database server, at least linear function in the database size, which is considerable in most practical applications. The calculations amount follows from the fact the server has to read all database records in order to execute the user query. Otherwise the user data privacy would be violated.

In order to solve this problem Gertner et al. have suggested a PIR solution where most calculations have been executed by special purpose servers instead of the database server [7]. These auxiliary servers keep arbitrary portions of the original database instead of its replicated copies, in such a way that none of the auxiliary servers could recover the initial database content. The protocol described includes computation amount at the database server reduced to *O(1)*, but the computation amount at the auxiliary servers is linear function at the database size.

Di-Crescenzo et al. have described another PIR scheme [6], which

includes special-purpose servers. Most computation and communications are moved offline, i.e. have been executed once, regardless of the number of the further user queries. Both models described in [6] and [7] do not secure user data privacy when all servers cooperate against the user.

PIR future tendencies.

- ✓ It would be useful for practical applications the amount of online calculations and network traffic to be further optimized taking full advantages of preprocessing and offline computation and communication.
- ✓ It is important in many particular cases the query definition to not be limited to the retrieving of the i-th bit or i-th record from a database. It is useful the query formulation to be similar to those used in the databases, SQL queries for example. For instance a DNA sequence similarity search in a DNA database cannot be performed using the usual PIR queries.
- ✓ There is special e-commerce platforms needed in order to apply PIR approach in the practice. Essential parts of such platforms are reliable payment algorithms.
- ✓ The present paper does not consider applications-specific PIR protocols. For example, the protocol designed for patent database could differ from those applied in the case of conventional databases.

PIR protocols allow the user to protect its data privacy, keeping the retrieved database records secret. The present paper represented a comprehensive survey of the PIR problem, focused on its potential applications, the results achieved and future development directions.

References

1. Ambainis A., Upper bound on the communication complexity of private information retrieval, In Proc. of 24th ICALP, 1997.
2. Cachin C., S. Micali, M. Stadler, Computationally private information retrieval with polylogarithmic communication, In Proc. of EUROCRYPT'99, 1999.
3. Chor B., N. Gilboa, Computationally private information retrieval, In Proc. of 29th STOC, 1997.
4. Chor B., N. Gilboa, M. Naor, Private information retrieval by keywords, Technical report, Technion: Israel Institute of Technology, 1997.

5. Chor B., O. Goldreich, E. Kushilevitz, M. Sudan, Private information retrieval, In Proc. of 36th FOCS, 1995.
6. Crescenzo G.D., Y. Ishai, R. Ostrovsky, Universal service-providers for database private information retrieval, In Proc. of 17th PODC, 1998.
7. Gertner Y., S. Goldwasser, T. Malkin, A random server model for private information retrieval, In Proc. of 2nd RANDOM, 1998.
8. Goldwasser S., S. Micali, Probabilistic encryption, Journal of Computer and System Sciences, 1984.
9. Kushilevitz E., R. Ostrovsky, Replication is NOT needed: Single-database computationally private information retrieval, In Proc. of 38th FOCS, 1997.
10. Mishra S.K., On Symmetrically Private Information Retrieval, PhD thesis, Indian Statistical Institute, Calcutta, Aug. 2000.
11. Ostrovsky R., V. Shoup, Private information storage, In Proc. of 29th STOC, 1997.
12. Smith S.W., D. Safford, Practical private information retrieval with secure coprocessors, Technical report, IBM T.J. Watson Research Center, July 2000.

CREATING TOPIC MAPS FOR E-LEARNING

Darina Dicheva and Christo Dichev

Winston-Salem State University
601 Martin Luther King Jr. Dr., Winston Salem, N.C. 27110
{dichevad,dichevc}@wssu.edu

Abstract: E-learning can be defined as learning facilitated through the use of information and communications technologies. The success of e-learning applications depends on the amount of effort involved in developing online learning materials by authors and retrieving learning resources by learners. Thus, accessibility, findability, and reusability of learning resources are critical issues in e-learning. To address these issues we propose an authoring environment for creating ontology-aware repositories of learning materials based on the new ISO standard - XML Topic Maps. Topic Maps provide a paradigm for organizing and retrieving online information and for interchanging semantic information on the Web. In this paper, we discuss briefly a general framework for developing Topic Maps-based e-learning applications and present our authoring tool - ETM-Editor - for creating Topic Maps-based e-learning materials.

Keywords: concept-based courseware, e-learning, ontologies, Topic Maps, Semantic Web

Introduction

E-learning can be defined as learning facilitated and supported through the use of information and communications technologies. It embraces learning of all types including academic instruction, professional training, and lifelong learning. Typically, it uses the Internet to bridge distances and enables people to learn no matter where they are. The successful and effective implementation of an e-learning strategy depends on more than just providing the technology, training and equipment. A crucial issue in e-learning is how to organize and classify the learning content so that learners and instructors can find what they need when they really need it. The success of e-learning applications depends on the amount of effort that is involved in developing online teaching and learning material by authors and in retrieving relevant learning resources by learners. Thus, accessibility, findability, and usability of learning resources are critical issues in e-learning. A solution to these problems involves providing a suitable model,

coupled with tools that support rapid development of e-learning applications.

To address the above critical issues we propose a (prototype) courseware development tool based on the ISO 13250 XTM standard - XML Topic Maps [1]. Topic Maps (TMs) are emerging technology, that can be used as a means to organize and retrieve information in e-learning repositories in a more efficient and meaningful way. The expressive power of Topic Maps, commonly perceived as a method for indexing of information resources, places the standard very close to artificial intelligence and knowledge modelling. Topic Maps resemble semantic networks and conceptual graphs, but offer more - a unique, standards-based way of encoding and exchanging of knowledge. Topic Maps provide an external meta-structure (a knowledge navigation layer or ontology) in form of a dynamic, semantically based hypertext. As a result, TM-based courseware can offer the following benefits:

- *For learners:* easy finding of relevant content; "browsing" in a subject field (knowledge domain) that supports exploratory learning; "help system" (orientation) through the knowledge layer (represented by interrelations or associations between domain concepts); learner-centric learning process adapted to individual learner's interests and needs.
- *For instructors:* structuring and presenting the content as a Semantic Web; distributed courseware development and ongoing further development; reuse and exchange of teaching and learning materials.

We have proposed a general framework for developing Topic Maps-based e-learning applications [3] and have used it to implement an authoring tool for creating Topic Maps-based e-learning materials - ETM-Editor (Educational Topic Maps Editor). In this paper, we discuss briefly the framework and present the ETM-Editor.

General Framework for TM-Based Environments

The proposed framework is aimed at supporting the development of *ontology-aware* repositories of learning materials, including support for learning content organization and flexible communication. It is focused on enabling authors (educational institutions) to capture, share and access knowledge.

Subject ontologies aim at capturing domain knowledge in a generic way and provide commonly agreed upon understanding of a subject domain, which may be reused and shared across people and applications. Ontology editing is an essential aspect for all ontology-aware systems. An important

issue within ontology editing is the underlying ontology model or "structure" that is to be edited. A key feature of the underlying model of our ontology-driven framework for developing repositories of learning resources (objects) is a *network of concepts*. This involves creating views of a specific domain in terms of domain concepts and relationships among them that suggest the semantics of resources relevant to that domain. Such a conceptual structure would enhance information retrieval within the repository since the set of concepts, relationships, and inference rules defined by the domain ontology constrain the possible interpretations.

Thus, the proposed general framework of ontology-aware discipline-specific repositories is based on building a conceptual structure that represents the subject domain ontology and using it for structuring and classification of the learning content [3]. The classification involves linking learning objects (content) to the relevant ontology terms (concepts), i.e. using the ontological structure to *index* the repository content. This will allow applications and users to understand the relationships between the resources and thus will insure efficient topical access to them. By providing shared agreement on the subjects meaning, ontologies can serve as a means of establishing a conceptually concise basis for communicating knowledge for many purposes, for example, in ontology-based merging of digital repositories. In e-learning applications, the ontology-based content is typically constructed and maintained in a collaborative effort, using appropriate authoring tools. It also requires 'browsing' tools for exploration of the ontology-based repositories. Therefore, the focus in the proposed framework is on the information repository, the information-authoring module, and the information retrieval module. An architecture built within this framework utilizes the advantages of concept-based and standards-based content organization, which will benefit both learners and authors. For learners it will support efficient contextual retrieval of information relevant to their needs and for authors - the reusability, shareability, and interoperability of created instructional materials.

We have proposed a layered information structure of the repository consisting of three layers, each of which captures a different aspect of the information space - conceptual, resource-related, and contextual. Therefore, the authoring tool that we developed within the proposed framework supports semantic layer (ontology) authoring, resource authoring, and context authoring.

The ETM Editor

It is very important that ontology-aware e-learning applications provide support for both ontology development and ontology usage. In the last decade, a number of environments for ontology construction and use have emerged, such as Protégé-2000 [4] (see http://protege.stanford.edu/). However, when decision has to be made about how practically to build ontology, several basic questions arise related to the tools to be used, such as:

- Is it possible to reuse existing ontologies in the specific subject domain?
- Is it possible to merge similar subject ontologies?
- How can e-learning applications interoperate ontology-based content?
- How are ontologies and related content stored (e.g. in XML, databases, etc.)?
- Is *translation* provided between different ontology implementation languages?

Taking into account such considerations, we have designed and implemented a prototype of an authoring environment, ETM-Editor, which enables the creation of ontology-aware courseware using domain concepts for structuring content and presentation. We have chosen to use the new ISO standard – XTM (XML Topic Maps) - to implement the environment. Ontologies and Topic Maps are complementary tools that aim at giving a more global vision than terminologies, thesauri and concepts systems. Whereas the Topic Maps specification ensures syntactic interoperability, ontologies provide semantic interoperability.

Topic Maps

Topic Maps provide a paradigm for organizing and retrieving online information and for interchanging semantic information on the Web [1]. Basically, Topic Maps are collections of *topics*, *associations* and *scopes* [5]. In Topic Maps the concepts are reified in topics, and they can be categorized using types. TM describe by means of topics what an information set is about. An association expresses a semantic relationship between topics, and the extent of validity of this association is called scope. TM can be viewed as a method for structuring and organizing information on the semantic and metadata level. They allow to incrementally add meaning and express new relationships among resources (such as *isExampleOf, hasSubjectOf, isAuthorOf, dependsOn* etc), making explicit the particular

contextual relationships of interest. Thus, Topic Maps fulfil a universal need to associate information to form the information networks that make up a knowledge base.

The advantage of using the Topic Maps technology for developing digital learning resources is twofold: from one side it supports convenient and intuitive presentation and manipulation of interrelated concepts embedded in information resources, and from another, the learning material is in a standard format, which makes it interchangeable i.e. it can be used in other TM-based systems. In addition, the Topic Maps standard defines the way in which two or more Topic Maps may be combined or *merged*. Topic Maps merging is highly useful as it provides a means for different users to share and combine their Topic Maps in a controlled manner, which supports TM reusability, shareability, and interoperability.

While Topic Maps offer a powerful and promising technology for intelligent organization and access of information in general, creating Topic Maps for e-learning is not a clear and simple task at present [2]. A number of commercial TM tools are available and vendors such as Empolis, Infoloom, Mondeca, and Ontopia have commercial offerings consisting, for the most part, of TM engines and browser applications (see http://www.topicmap.com). However, most of the existing TM editing tools offer only limited functionality and user support due to their generality. Therefore, specialized tools for supporting end-users have to be developed in different areas, incorporating the specifics of that area. According to our knowledge, there are no specialized TM editors for the area of education developed so far. This was our motivation for implementing the ETM-Editor.

Design Principles

The ETM Editor is an ontology editor allowing the user to build ontology driven learning repositories using Topic Maps. It provides ontology and metadata engineering capabilities coupled with basic document management facilities. The editor incorporates an interface that reflects the language, thinking, and processes in the application domain – courseware development. A generic Topic Map editor that would enable creating topic T or association A of type X in scope S was not the right solution. The driving idea was to create a Topic Map editor tailored to particular tasks and reflecting certain application logic that will be solved by applying the Topic Map paradigm. Instead of subjecting authors to all the TM complexities, e-learning applications that are developed should not require authors to know

that they are working with Topic Maps.

The ETM Editor is designed as a tool, in which authors externalize their subject knowledge in domain conceptual structures represented by Topic Maps. The basic steps in building a Topic Map assume that the user has a conceptual specification of the learning content in terms of its topic set, association set, occurrence set, and their interconnections. In our previous work, we have identified some problems that users often encounter in creating conceptual structures. We have explored further a number of research questions related to using TM for representing educational resources.

The ETM Editor is designed to offer enhanced, specialized support for creating concept-based digital learning repositories. It benefits from the TM's basic feature to support easy and effective merge of existing information resources while maintaining their meaningful structure. This allows for flexibility and expediency in re-using and extending existing repositories. The learning content created by the editor is compliant with the XML Topic Maps standard and thus interchangeable with other standard TM tools. The questions, raised above resulted in the following more specific design criteria:

- Provide intuitive interface that:
  - reduces authors' cognitive overload when creating and editing concept-based learning content,
  - is easily learnable,
  - allows merging Topic Maps from different educational repositories.
- Provide extensive support to authors in developing educational Topic Maps, including:
  - support for creating and modifying existing TM-based learning objects, including easy access and manipulation of TM constructs using a custom (e-learning domain) language,
  - easy comparison and merge of independently built learning objects (e.g. from different repositories),
  - checking for inconsistencies in the learning content.

The ETM Editor's functionality includes the following capabilities:

- Maintaining concepts: adding concepts, deleting concepts, linking concepts to other concepts.
- Creating learning objects: defining learning object types, adding learning

objects, deleting learning objects, modifying learning objects, merging learning objects.

- Creating contexts (organizing learning objects): linking learning objects conceptually, organizing learning objects hierarchically, and defining different views.
- Importing/exporting Topic Maps, i.e. 'transporting' Topic Maps from one application (repository or system) to another.

ETM Editor Implementation

The ETM Editor is an authoring environment that supports the overall, cyclic process of learning repository engineering including adding and updating repository entities and related metadata as well as browsing and codifying them. The ETM Editor is Topic Maps-based, thus the main objects that it manipulates are topics (representing domain ontology concepts), relationships between them (corresponding to the TM associations), resources, and views (implementing the TM scoping feature). The editor provides window-based, event driven graphical user interface with pop-up dialogs, menus grouped by function, and drop-down lists that minimize typing. The key goals for the editor are the same as for any software tool designed to support a library type collection of information – collocation of related information and effective navigation of the collection. A screenshot of the ETM Editor interface is shown on Fig. 1.

The Editor GUI includes four different sections: *Topic Map*, *Topics*, *Relationships*, and *Views,* with the *Topics* section shown above. On the left side, the topics ontology is represented in the form of "browser tree" that allows navigation through the topics. This topic tree can explicate the local context while also presenting an overview of the whole Topic Map. In addition, it allows determining connected components. While editing Topic Maps, it might happen that several "islands" are created that have nothing in common. Such islands can be identified fast using the tree view. Using the ETM Editor the user can create concept hierarchy for a given domain. The result is displayed in the tree view, which can be navigated by clicking on the nodes. For easier construction of relationships the editor has a pool of predefined (courseware authoring related) relationship types and role types.

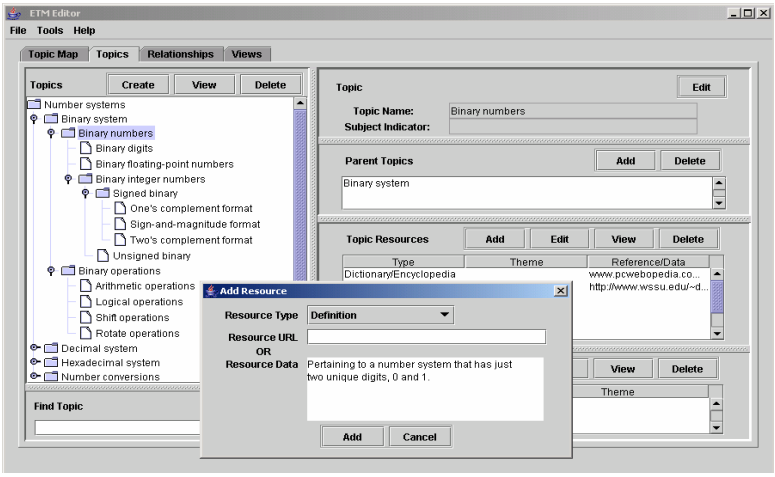


Figure 1. A screenshot from the ETM Editor interface.

The GUI provides intuitive control of all functions. The user can create, delete or change a topic, a resource, an association (between topics with specified roles and types) and a scope in the Topic Map. All changes are immediately reflected in the tree view. The author can specify what parts of the Topic Map the viewer or web application is to show to users. This context filter uses the scopes property defined in the Topic Map, and lets users decide which topic names they prefer to use, and what occurrences and associations they do not wish to see. Using the 'Find Topic' option the user can also search for topics by their names. When a topic is selected, all information related to that topic, including topic's subject indicator, type (parent topic), names and resources, can be viewed in the right-side panels: 'Parent Topics', 'Topic Resources' and 'Topic Names' and their popping-up dialogs. The ETM Editor is implemented as a client-server application developed in Java and using the TM4J Topic Map Engine [6], which is an open source providing a comprehensive API that allows creating and modifying Topic Map structures stored either in-memory or persistently in a database. It has open modular architecture that allows easy extension of its functionality.

Conclusions

The work reported here is aimed at contributing to the development and

use of efficiently searchable, reusable, and interchangeable discipline-specific repositories of learning objects on the Web. We proposed an authoring environment supporting the development of standards-based ontology-aware online learning materials. The next step in our agenda is the design and development of a browser for Topic Maps-based learning materials. It will support learners efficiently to navigate educational Topic Maps and search for useful resources. The latter is crucial in project-based and self-directed learning where the learners are actively engaged in retrieval of relevant information.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DUE-0333069 "NSDL: Towards Reusable and Shareable Courseware: Topic Maps-Based Digital Libraries." We would like to thank our students Somath Nao and Yanli Sun for their implementation efforts.

References

1. Biezunski, M., Bryan, M., & Newcomb, S., ISO/IEC 13250:2000 Topic Maps: Information Technology, www.y12.doe.gov/sgml/sc34/document/0129.pdf.
2. Dichev C., Dicheva D., & Aroyo L. Using Topic Maps for Web-based Education, J. Advanced Technology for Learning, Vol. 1, No 1, 2004, 1-9.
3. Dicheva, D., & Dichev, C. A Framework for Concept-Based Digital Course Libraries. Journal of Interactive Learning Research (accepted), 2004.
4. McGuinness, D. Ontologies Come of Age. In Fensel, D. Hendler, Lieberman, H. & Wahlster, W. (Eds) Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential. MIT Press, 2002.
5. Park, J. & Hunting, S. XML Topic Maps: Creating and Using Topic Maps for the Web, Addison-Wesley, 2002.
6. TM4J: Topic Maps for Java, http://tm4j.org/dg-d0e35.html.

REFACTORING COMPONENT BASED SOFTWARE WITH ASPECTS

Dimitar Birov

Sofia University, FMI
James Bouchier blvd. 5 Sofia, 1164
birov@fmi.uni-sofia.bg

Abstract: This paper focuses on the study of the problems of refactoring of software systems developed from components using aspect oriented paradigm and methodology. We present an aspect oriented view of refactoring of the component based software on each phase of component based development - specification, implementation, package, assembly and deployment and propose an Aspect Component Based Software Refactoring methodology as an addition to Aspect Component Based Software Engineering.

Keywords: Aspect Component Based Software Engineering, Aspect Oriented Programming, Aspects, Components, Refactoring

1. Introduction

The goal of the process of refactoring of software is to improve existing design through decoupling, decomposition, and specification of object-oriented code. *Refactoring* means to change "… a software system in such a way that it does not alter the external behaviour of the code yet improves its internal structure" [3].

The goals of refactoring are: to keep the software as simple as possible; elimination of code duplication; revisit distribution of responsibilities; and redesign code, reducing its complexity. Refactoring decomposes, decouples, and simplifies system structure into easy manageable, readable, understandable, independent units. At programming language level they can be abstract data types, object, classes. At system construction level they can be components.

Component Oriented Programming aims at producing software components for a component market and for later composition. Component-Based Development is a technology for the construction of high-quality, large software systems in timely and affordable manners, reducing developing costs and efforts. A component should be able to be developed,

acquired and incorporated into the system and composed with other components independently in time and space [8]. It improves the flexibility, reliability, and reusability of the final application due to the use/reuse of software components already tested and validated.

Aspect oriented programming (AOP) [5, 6] views a software system as a combined implementation of multiple concerns. In the simplest form, there are *core* concerns that are natural components (that provides the actual functionality) of software. Additionally there is a system-level concerns such like logging, debugging and authentication that tend to affect several other concerns. For instance, a logging implemented into software implicates set of components and classes. Each component should have its own code for logging, making the components less specialized, resulting in making it very difficult to predict effect changes to the logging code will have. This phenomenon is called *crosscutting concerns*. A typical software system may consist of several kinds of concerns including business logic, data persistence, logging and debugging, authentication, security etc.

Aspect-Oriented Software Development (AOSD) is software-developing paradigm that cope with the issues arisen by crosscutting concerns. Aspects are first class entities (such like classes) in order to encapsulate crosscutting concerns. An aspect can be implemented in a separate module, though the functionalities provided by it are spread across the application even crosscutted.

Refactoring as a topic of software development constantly discussed last decades. Aspect-orientation and refactoring are both concepts for decoupling, decomposition, and simplification of code. Jan Wloka [9] explores whether refactoring and AOP can coexist. He proposed refactoring patterns (similar to design patterns [4]) as specific guides for performing refactoring, following initial catalogue presented by Martin Fowler [3]. Wloka stated that refactoring over complex structure using just classical object-oriented methods is difficult and sometimes impossible because of the deep dependence of parts of the system. Possible decision is crosscutting concern that is spread across many types and methods. In this paper, we extend this to components.

Clemente and Hernández analyzed problem of crosscutting produced during component development and proposed extension named Aspect Component Software Engineering (ACBSE) [2] of Component Based Software Engineering (CBSE). They reported improvements and next advantages of ACBSE: improving reusability by strength of independency of

components; increasing adaptability of software; scalability and compressibility.

Current paper can be considered as an extension and addition to the ACBSE considering refactoring of components. We will discuss refactoring in all phases of component development.

The rest of the paper is organized as follows. Section 2 introduces specifics of component based software development. Section 3 defines limitations of orthogonal design requirements of component-based development. Section 4 presents an extension of Aspect Component Based Software Engineering (ACBSE) [2] to refactoring of the CBS and proposes an Aspect Component Based Software Refactoring (ACBSRE) methodology.

2. Component Based Development

The idea of building software systems from building blocks or independent modules is not new. Because dividing into specification and implementation each module can be integrated into and composed with others to build expected software system. Building an application passed through following phases in presence of components: description (specification) and implementation (CBD phases); interconnection and deployment (package assembly and deployment phases). The components are not so reusable and adaptable as we can expect and it is because of the natural phenomena of crosscutting among modules of the system.

Because business rule is a process in software systems, changes in software specification and requirements influence mainly changes and refactoring of the business rules implementation of the software. Components are composed from *functional* and *non-functional* code. Business rule is regarded as functional property of the component. Synchronization, notification, logging, exception handling, memory management are typical crosscutting concerns and they are specified as non-functional properties.

Global design improvements through refactoring are impeded by crosscutting concerns. A concept for encapsulating the crosscutting implementation of concerns emerged naturally.

3. Applying AOP to Component Architecture Design

The main characteristics of software developed with AOP are flexibility, adaptability and reusability of the elements used to compose the system [6]. AOP provides various mechanisms to express, adapt, isolate and reuse *crosscutting concerns* in the software development to obtain these main characteristics.

Aspect-oriented programming allows us to decompose software systems in different dimensions. We can use a vertical decomposition process to establish the primary decomposition model of the architecture. We then use aspect-oriented techniques to compose ”horizontally” or to ”superimpose” the implementation for orthogonal design requirements onto the primary model, without modifying the existing architecture. We refer to that decomposition process as the horizontal decomposition.

Two design requirements are orthogonal to each other if one can be implemented without coordinating with the other, such as in the case of the requirements for efficiency and the requirements for location transparency. Since each of the orthogonal requirements must have its own most appropriate decomposition model, the vertical decomposition process, which generates one decomposition model, may not be an optimized solution for both requirements. In fact, it generates a model with tangled logic, as indicated by Gregor Kiczales [5]. The presence of orthogonal design requirements is a limitation in component-based system, which has appropriate decision proposed by aspect-oriented approach.

4. Refactoring Components Development Using AOP

Aspect orientation improves reusability, adaptability and flexibility of the component software. Aspect-Oriented Component System is built from three artefacts: the component program, the aspect program and the aspect weaver. An AspectJ [7, 10] or AspectWerkz [11] languages (both are natural extensions of Java), could be used for creating aspect program. EJB or CORBA [12] can be viewed as examples of tools for creating component program.

As a first step of refactoring we need to identify the presence of the crosscutting points in the primary decomposition model. This way the tangled code is transformed to three types of groupings of classes in the aspect-oriented implementation, namely, primary classes, aspect implementation classes, and the weaving classes. The importance of such

division is that it allows all three components to be designed, tested and evolved with unprecedented independence and freedom.

4.1. Redesign and Re-specification of Components

Process of refactoring of a design and component specification could start with reviewing previously design/specification documents describing interface, specifications and business rule(s) of the current component. If such design documents are not available we need to extract information from component code or from component behavior. Of great importance is to specify and understand interfaces which component provides as well interfaces that component requires. A *use clause* specifies interfaces the component expected or requires from other components in software system. Next important step is obtaining a clear view and separation of functional and non-functional properties of the component based on business rules of the component and the system. Business rules implement functional properties and reflect the specification and interconnection between components in the design phase [1]. Then review/redesign of business rules of the component and refactoring could be performed with avoidance of introduction of fresh errors by using aspect oriented design patterns [4].

Alternative approach [2] is to delay crosscutting at the implementation phase. An advantage of this is that the components can be easily adapted to the requirements of the context. And the final implementation decision can be done when a significant information for it have already been specified or realized. We can use type of component dependency (a *strong* dependency is defined when dependent component description and use are critical for main functionality of the component and *weak* dependency is defined when dependant component functionality is closely related with the context or framework environment and removing it from use clause of the current component will not disturb main component functionality) as criteria for deciding what to specify and refactor on respecification phase and which part to be delayed till implementation phase.

4.2. Implementation Phase

Aspect oriented approach make decoupling and modularizing of the reimplementation of the components easy. Aspects naturally modularize and encapsulate different aspects of the component dependency. Potential candidates for “aspectualizing” are *weak* dependant components. Component developing process naturally divided to component developer

who will concentrate over developing (extracting from source) main functionality based on *strong* dependencies between components and "aspects" developer who will use aspect-oriented techniques for decoupling or minimize interdependencies between components. For instance interfaces of the weak dependant component are potential source for started point of aspects functionality implementation.

4.3. System Package Phase

At this phase developer can specify policies for each component and configure security, transactions and persistence [12]. As well container configuration can be changed.

XML descriptors are commonly used for describing component properties in Package phase of component development. XML is very suitable language for describing interconnections between components. These way system architecture is described on an interface level – required and provided interfaces are described for component. The XML component descriptor specification can be produced from UML representation handled in system design phase through translation the dependencies of component The XML component descriptor specifies strong and weak dependencies between components. Developer can apply aspect-oriented approach to refactoring business rules, invent and add new ones and new dependencies in changed context and environment. Then pre-processing and recompilation of the component code is needed. The dependencies are expressed as aspect implementations through a generic aspect-oriented programming language, such as AspectJ [10] or AspectWerktz [11].

4.4. System Assembly and Deployment Phase

At this phase, location of each component has to be described. It is stronger requirement if we design distributed system. The UML diagrams for assembly and deployment of component-based systems are translated to XML Assembly Descriptor, which consists from component and property descriptors. A specific properties file, read by all components (positioned locally or on net) is generated from XML descriptor. It specifies location of each component of the software system using different types – URL, IOR, NameService.

4.5. Advantages of refactoring with aspects

Refactoring using aspects improves *reusability* of components both ways: through encapsulating and modularize "weak" dependencies between components and their non-functional properties and through increasing independence from context. Last, one is a key for *adaptability* of components – adding new aspects to components and specifying them in component descriptors is a way one can change component functionality and adapt it to the context. This way a new functionality, respectively a new component (implementation and specification) is obtained. In addition, a new feature – *scalability* of the component software is reached. All information about the system and interconnection of the components is represented using XML language and UML design description schemata. This improves *compressibility* and *readability* of the software system.

5. Conclusions

In this paper we have presented refactoring methodology as an addition to Aspect Oriented Component Software Engineering. We discussed refactoring steps following the life cycle stages of a component-based system and explore full advantages of AOP and technology allows. As result a more modularized, flexible, adaptable, reusable and compressible software systems are obtained.

References

1. Chessman, J. and Daniels, J., UML Components: A Simple Process for Specifying Component-Based Software, Addison-Wesley, 2001.
2. Clemente, P., Hernández, J., Aspect Component Based Software Engineering, The Second AOSD Workshop on Aspects, Components, and Patterns for Infrastructure Software (ACP4IS) 2002
3. Fowler, M., Refactoring: Improving the Design of Existing Code, Addison-Wesley Longman, 1999
4. Hannemann, J., Kiczales, G., Design Pattern Implementation in Java and AspectJ, Proceedings of the 17th OOPSLA 2002, pp. 161 – 173
5. Kiczales G., "Aspect Oriented Programming," ACM Computing Surveys (CSUR), vol. 28, no. 4, 1996.
6. Kiczales, G. Aspect-Oriented Programming. 1997. Procedings of ECOOP, Springer Verang. LNCS 1241.
7. Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J., and Griswold, W. An Overview of AspectJ. 2001. Proceedings of ECOOP, Springer Verang. LNCS 2072.

8. Szyperski, C., Component Software: Beyond Object- Oriented Programming, Addison-Wesley, 1998.
9. Wloka, J., Refactoring in Presence of Aspects, Position paper, 13th Workshop for PhD Students in Object-Oriented Systems
10. AspectJ, www.aspectj.org
11. AspectWerkz, aspectwerkz.codehaus.org
12. Object Management Group (OMG). Specification of Corba Component Model (CCM). 1999.

FUZZY SETS AND WEB SITES CLASSIFICATION

Georgi Furnadzhiev

Institute of Mathematics and Informatics, BAS
Information Research Department
Acad. Georgi Bonchev St., Block 8, Sofia 1113, Bulgaria
furnadjieff@math.bas.bg

Abstract: In the presented paper a fuzzy sets implementation into web sites classification is considered. Web sites external features are addressed and the possibility to use them for the classification is proved.

Keywords: web mining, fuzzy sets, classification

Introduction

There are more than 4.10^9 Google indexed web pages in the World Wide Web. Finding relevant information is very difficult. Searching information is a main problem. When we find many results, it is a good idea to classify them.

Using web search engines we can choose: result file type, language, domain, etc. Often we receive a message "This web site is added to directory X in category (ies)..." in the result list. This directory contains qualitative, but very small subset of all web sites in the world, and for most results, we do not have any information about their types. This makes a big part of our result uncategorized. We can group them by region or language, for example, but not regarding their content. It will be good if we can, using a web crawler or metasearch engine, to specify a web site type from a given set at least. Other useful opportunity will be to classify uncategorized part of the result list of our search query. It is not the same using Google to find word "accommodation" in science conferences' web sites or travel agencies' web sites.

Web Sites Classification

Unknown objects classification is a main part of machine learning and data mining research. When we classify a set of objects, we need

- formal object and classes descriptions
- classification model

- training set and training mechanism
- rules adding unknown objects into a class

There are created many automatic classification approaches, based on artificial neural networks, decision trees, genetic algorithms, etc. The classification process follows the steps:

- Model choice.
- Training. We use a relatively small and labelled subset, called training set. The labels mean belonging to a class. Based on this training set, we construct the classifier.
- Unknown objects classification.

Web resources classification is an application of traditional data mining techniques in respect to the specific area. ([8], [9]) The datasets contain web sites. All web sites classification could be possible, if we have a good ontology describing the current state of the art. However, it is a very difficult activity. We have to know at least the current situation in the entire web. A rational idea is to have a specific sub-ontology and use it to decide on the particular problems.

Talking about web sites classification, we have to keep in mind two main arguments.

- The hyperlinks between web sites do not reflect on their types. The authors are not obligated to relate their web sites to any other ones.
- The most adequate web sites description approach is using quality data. We can detect features or count them only.

There are many realised approaches to determine web site type or automatic construction of web directories. In general, we can find two very popular directions – adapted for web documents text-mining techniques ([2]) and web structure mining techniques ([5], [6], [7]). In first case, authors prefer to weight different parts of the web sites or the web pages, and in the second – to use web structure in general. There are examples for domain specific classification ([3],[4]).

Here we try to prove how the type of web sites affects their external features. We try to find how the content influences the external view.

Web Sites External Features

We have to define what an external feature is firstly. Every web site can be considered from two points of view:

- Internal – this is the site structure, meta tags, technologies, formal languages used in site creation, etc
- External – this is the visible part of the site

For example, when the user clicks "Sign in", it could be a button, or (GIF or JPEG) image or text hyperlink in the different cases. It can start a script, written in some formal language, providing the same semantics.

When talking about links here, we mean external views of the same web site's links.

Fuzzy Sets

Fuzzy sets are presented in 1965 [1] and are very popular in the last decades. They are good mechanism for describing real word objects. In the other hand, the programme realisation of main operations with finite sets is easy. Here we remind some definitions.

Definition 1: Let X is a given set. A fuzzy set A over X is the set of doubles $(x, \mu_A(x))$ for every element x of X, when $\mu_A(x)$ is real number and $0 \le \mu_A(x) \le 1$

Every fuzzy set over X can be described by means of the function $\mu_A(x)$.

Definition 2: Let A is fuzzy set over X. A is empty (universal) if $\mu_A(x)=0$ ($\mu_A(x)=1$) for all x.

Definition 3: Let A and B are fuzzy sets over X. The fuzzy set C is A and B section if and only if $\mu_C(x)=min(\mu_A(x), \mu_B(x))$.

Definition 4: Let A and B are fuzzy sets over X. The fuzzy set C is A and B union if and only if $\mu_C(x)=\max(\mu_A(x), \mu_B(x))$.

Definition 5: Let A is fuzzy set over X. The set $\bar{A}$ is addition to A if and only if $\mu_{\bar{A}}(x)=1-\mu_A(x)$

Web Sites Features and Fuzzy Sets

The fuzzy sets are good mechanism for describing the features of the web sites classes. There are not any formal models for the web sites creation and the authors are not obligated to include anything. Moreover, main purpose in the web is to be distinctive. However, content and specific

area has an effect on the language, structure, representation of the data, etc. We can expect similar information to be presented in similar ways. From this point of view we cannot say a given feature is specific for a web sites class or not, but we can define a relative belonging into a set of features describing the class. That makes the fuzzy sets very relevant.

How to Prove

To define a fuzzy set describing a class, we need to discover a relatively small training set of web sites and their descriptions. For every member of this set we have to find the features contained in them at first, and compare the given results at second. With a simple comparison and counting, we find relatively belonging into a set of features for this class (and this small set). This makes our results as so accurate as our training set is representative.

We need to prove whether our fuzzy sets are relevant or not. Of course, the initial fuzzy set is not enough for the classes' description. It is possible to find one or more elements for all classes, but we have to find the specific ones. In a formal model if we have the classes $C_1 \dots C_n$, and T_i i= 1 … n are the fuzzy sets given from the first step, we actually are interested in sets

$$T_i \setminus \bigcup_{j \neq i} T_j \, (*)$$

for every i= 1, …, n. Here we can use the equation

$$A \setminus B = A \bigcap \overline{B}$$

where A and B are arbitrary sets. This representation will help us to apply definitions for the section and the union of fuzzy sets. If for every i = 1, …, n all of the sets (*) are not empty and are not the universal set, we can say we have found lists of features describing given classes.

This model is temporary because of the temporary nature of the web. It is exact for the training set only, not for all web sites in the world, belonging into the classes. Moreover, it provides correlations among the given classes, but not among all classes, which could exist in the world around. To improve the model correctness and accuracy we have two ways:

- Using carefully selected and relatively big training sets
- Frequently testing the training set for changes and actualise the features database and sets (*)

How to Classify

The next task is to find a rule for unknown web site evaluation. A natural approach is to consider every web site description like a fuzzy set too and find all of the distances between this description and the fuzzy sets, associated with the classes. An uncategorized web site belongs to a class, if and only if, the distance between the site and the class is the smallest. The distance can be defined in many different ways. Actually, this is clustering web with preliminary defined cluster centres. In our works, we compare the Hamming and the Euclidean metrics. The metrics choice can be automated. It is necessary to have program applying two or more metrics or similarity functions. In the second case, the system must prove how similarity is bigger. The system can simultaneously follow two criteria:

1. Better total correctness, and in case they are equal -
2. The web sites distribution after the test. The statistical dispersion is good measure there.

For metrics choice, the same training set can be used.

Objects and Classes Descriptions

Web sites' descriptions in this model are simple. For every one we define a vector $V_i(v_{i1},v_{i2}, \dots , v_{in})$ where v_{ij}=1 if the feature j is found in the site, and v_{ij}=0 if the feature is not found in the site.

If we have m web sites belonging into a given class, we define the vector T with components:

$$t_j = \frac{1}{m}\sum_{k=1}^{m} v_{kj}$$

It is not difficult to see that

- vector T defines a fuzzy set
- if we have two or more classes and mark them with T_i their vectors, then $D_i = T_i \setminus \bigcup_{k\neq i} T_k = T_i \bigcap \overline{(\bigcup_{k\neq i} T_k)}$ is a fuzzy class descriptor for every i.

Experiment

We made experiments with 100 web sites from five following types

T1. University web sites
T2. Newspaper web sites
T3. International unions web sites
T4. Governmental web sites
T5. Parliament web sites

We used 20 web sites by class. Their first nontrivial pages have been considered. Here by *nontrivial page* we mean the first page after simple Enter page. We used Yahoo! Directory for finding representative for all world-training sets, from different languages, countries and continents with respect of their relative distribution. When we described these web sites, we obtain 127 different features. We count the features found into the classes. We compare the classes by (*) and obtain classes descriptors. Here we give elements x with $\mu(x) \geq 0.5$ for every class.

T1: University web sites (30 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "Alumni"	0.70
Link "About university"	0.70
Link "Structure"	0.65
Link "Events"	0.65
Link "Library"	0.60
Link "Researches"	0.60
Link "Students"	0.60
One colour background	0.55

T2: Newspaper web sites (60 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "News"	0.60
Link "Sport news"	0.60
Link "Archives"	0.50
Link "Advertising"	0.50

T4: Governmental web sites (57 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "Searching"	0.50

T3: International unions web sites (52 elements with nonzero value of $\mu(x)$)

Feature	Belonging
Link "About us"	0.55

T5: Parliament web sites – 45 elements with nonzero value of $\mu(x)$ but $\mu(x) \leq 0.45$ for all. The first five are

Feature	Belonging
Language choice	0.45
Link “Contacts”	0.35
Links to institution’s documents	0.35
Link "News"	0.35
One colour background	0.35

In our tests, we compare the Hamming and the Euclidean metrics and test them with 10 random selected web sites – by two for class. The results are given in the following tables

Euclidean metrics						
	1	**2**	**3**	**4**	**5**	Correctness
T1	2					100%
T2		2				100%
T3			2			100%
T4			2			0%
T5				1	1	50%
Total	2	2	4	1	1	65%

Hamming metrics						
	1	**2**	**3**	**4**	**5**	Correctness
T1	2					100%
T2		1			1	50%
T3			1		1	50%
T4					2	0%
T5					2	100%
Total	2	1	1		6	60%

Here “Correctness” is the percent of web sites correctly added into their class’ sets. Based on the results we can say the Euclidean metrics is better.

Less correctness for some types we can explain with classes' similarity. Distance matrix between classes is as follow

T1	0,00				
T2	2,48	0,00			
T3	1,99	1,71	0,00		
T4	2,04	1,82	1,04	0,00	
T5	1,94	1,80	0,94	1,03	0,00
	T1	**T2**	**T3**	**T4**	**T5**

As we can see, the best results in metrics tests we obtain for most "isolated" classes.

Conclusions

Based on the experiment results we can say this approach have acceptable correctness for further studies and applications. The best results are observed for less similar classes. The main weak points are similar classes' areas. The approach is applicable to most general web sites categories.

It is a good idea to prove the approach in a similar web sites classification. There are many huge categories in the web directories. We can apply the approach for subcategories creation. We can expect similar classes, but the web sites are similar too.

Other result is fuzzy sets are suitable mechanism for web sites classes description and study.

The results manifest how important the web site structure is. The most of the described features are external representation of this structure. It is prove in practise the proposition web sites are independent objects for classification.

References

1. Zadeh L., Fuzzy sets, Information and control, Vol. 8, 1965 (338 – 353)
2. Pierre J., On the Automated Classification of Web Sites. Linköping Electronic Articles in Computer and Information Science, Vol. 6(2001): nr 0. http://www.ep.liu.se/ea/cis/2001/000/. February 4, 2001
3. Ardo A., T. Koch, and L. Nooden. The construction of a robot generated subject index. EU Project DESIRE II D3.6a, Working Paper 1 1999. http://www.lub.lu.se/desire/DESIRE36a-WP1.html
4. Kock T., A. Ardo. Automatic classification of full-text HTML documents from one specific subject area. EU Project DESIRE II D3.6a, Working Paper 2 2000. http://www.lub.lu.se/desire/DESIRE36a-WP2.html
5. Attardi G., A. Gulli, and F. Sebastiani. Automatic Web Page Categorization by Link and Context Analysis. In Chris Hutchison and Gaetano Lanzarone (eds.), Proceedings of THAI'99, European Symposium on Telematics, Hypermedia and Artificial Intelligence, 105-119, 1999.
6. Cho J., H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. In Computer Networks and ISDN Systems (WWW7), Vol. 30, 1998.
7. Rennie J., A. McCallum. Using Reinforcement Learning to Spider the Web Efficiently. Proceedings of the Sixteenth International Conference on Machine Learning, 1999.
8. Han, J. and Chang, K. C.-C. Data Mining for Web Intelligence, IEEE Computer, Nov. 2002
9. M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim, Data Mining and the Web: Past, Present and Future, Proceedings of WIDM99, Kansas City, U.S.A., 1999.

APPROACH FOR DYNAMIC EVALUATION IN DISTANCE LEARNING ENVIRONMENT

Daniela Orozova and Maria Monova-Zheleva

Burgas Free University
101 Alexandrovska St., Burgas 8000, Bulgaria
orozova@bfu.bg, mariaj@abv.bg

Abstract: The use of information and communications technology (ICT) in education is leading to fundamental changes in traditional learning and teaching practices. Distance learning is one of the most widespread forms of open learning. The distance learning systems use different approaches in regard to the organization and management of the learning. The development of methods and tools for dynamic assessment and evaluation of the level of knowledge and skills obtained by learners is a real challenge. The different forms and strategies for assessment and evaluation are described in this paper. Concrete approach for dynamic evaluation of computer tests is presented also.

Keywords: Assessment in Education, Evaluation, Computer Assisted assessment (CAA), Computer-based assessment (CBA), Computer-based tests.

Introduction

The distance learning systems have to be adaptable to the users' needs and preferences. They have to allow actualization, adding of new functions, and using of databases for the different learning domains as well. Another important task concerning the distance learning systems implementation is the development of effective methods and tools for systematic gathering of information about the level of learners' knowledge and skills, and for reporting the learners' achievements and growth.

Evaluation is a judgement about the quality of a response, product or performance, based on established criteria and standards. In regard to learning the assessment, evaluation and reporting student achievement and growth are integral parts. This means that the methods and tools for assessment and evaluation have to be based on common educational standards and criteria.

Assessment in education

Assessment is the systematic gathering of information about what students know, are able to do, and are working toward. Assessment should be continuous, collaborative, consultative and based on an agreed set of criteria. Assessment focuses on the critical or significant aspects of the learning to be demonstrated by the student. Students benefit when they clearly understand the learning goals and learning expectations. [1] Assessment in education is a complicated issue today. There is a large debate about the tools that we use to measure student learning. Assessment methods and tools include as follows: observation; student self-assessments; daily practice assignments; quizzes; samples of student work; pencil-and-paper tests; holistic rating scales; projects; oral and written reports; reviews of performance; portfolio assessments. In addition to "auditing" student performance, assessment is expected to improve student understanding through the use of "authentic" tasks and relevant, ongoing feedback. Advances in computer technology are beginning to catch up with these challenges. Computer Assisted assessment (CAA) is a common term for the use of computers in the assessment of student learning. Various other forms exist, such as Computer-Aided Assessment, computerized assessment, Computer Based assessment (CBA) and computer-based testing [2].

Evaluation

Evaluation of student performance is the process of making decisions based on the interpretation of evidence gathered through the assessment for the purpose of goal setting and/or reporting. Teachers/experts use their insight, knowledge about learning, and experience with students, along with the specific criteria they establish, and to make judgments about student performance in relation to expected learning outcomes. This information is used in order to make decisions about effective instruction for a learner or groups of learners, redirect efforts, and establish future learning goals.

When evaluation is seen as an opportunity to promote learning rather than as a final judgment, it shows learners their strengths and suggests how they can develop further. Students can use this information to redirect efforts, make plans to practice the learning, and establish personal learning goals. Teachers are better able to assist in goal setting, provide support, and enhance the student's learning.

Using the forms of written assessment reduces the influence of some

deforming effects such as the effect of the gender, the effect of the voice, the effect of the halo, etc.

Assessment Techniques

Assessment techniques are usually either norm referenced or criterion referenced. Norm referenced assesses an individual's performance in relation to the norms established by a peer group. Criterion referenced occurs when a student is assessed on his or her ability to meet a required level of skill or competence. Computer Assisted Assessment is usually criterion referenced.

- Norm-referenced evaluation is used for large-scale system assessments; it is not to be used for classroom assessment. Norm-referenced evaluation compares student achievement to that of others rather than comparing how well a student meets the criteria of a specified set of learning outcomes. A classroom does not provide a large enough reference group for a norm-referenced evaluation system.
- Criterion-referenced evaluation compares student performance to established criteria rather than to the performance of other students. Criterion-referenced evaluation is most appropriate for evaluating student performance in the classroom. Criterion-referenced evaluation requires that teachers establish criteria based on the expected learning outcomes. Criteria can be used to evaluate student performance in relation to learning outcomes. The criteria are used to guide, monitor, and evaluate learning.

Computer assisted testing

Standardized testing remains the dominant approach to student evaluation today. The term test refers to set of questions/problems in definite subject domain and the corresponding system for evaluation of the answers/solutions. The test results report on the level of obtained knowledge and skills in the subject domain at a given stage of pedagogical process.

The main feature of the computer test as a form of assessment is the lack of direct with the learner. Well-written computer assisted testing is more likely to be objective testing: testing that can be marked objectively and thus offer high reliability. The benefit is that the tests can be marked quickly and easily, and adapted to meet a wide range of learning outcomes. Tests have

the following potential [2]:

- to incorporate a wide range of media;
- to incorporate hints into test questions;
- to link online assessment to feedback;
- to assign other learning activities based on the test results;
- to make randomized selection can be made from large question banks;
- to be administrated easily.

The test is a tool for assessment and evaluation of abilities and knowledge. The development of computer tests and approaches for evaluation is a very important task about distance learning systems implementation. Construction of good objective tests requires skill and practice and so is initially time consuming. Implementation of CAA system can be costly. Hardware and software must be carefully monitored to avoid failure during examinations.

In the context of the approach for dynamical evaluation presented below, the tests have to include only questions/problems from one of the following categories:

- Unique answer: These questions have one, unique answer (string) which is well defined. Their evaluation is made automatically from the system comparing the learner's answer with the correct answer.
- True/False (T/F): The answer can take the value true or false. Similarly, the evaluation is made from the system. Some research has also shown this test format to perform poorly with regard to reliability. Nevertheless, in some situations, the "multiple true-false" (or Type-X format) can offer adequate results. [3]
- Multiple-choice (M-C) questions: These questions have several possible answer choices (correct answer as well as three to four distracters). The correct answer is only one. Because M-C questions can be scored automatically and their scores offer high reliability, they are the assessment of choice for many distance learning instructors. But, M-C test results can often be weak with respect to validity, especially when testing students for higher-order thinking or competence in procedural skills. [3]

The approach for dynamic evaluation of computer-based tests

This is an approach for evaluation of different skills and defined in advance different types of errors and knowledge units. The common score is based on the assessment of these parameters.

The scores reflecting the learner's level of assimilation of given knowledge are in fact the scores of questions, problems or tests. The scores are real numbers in the interval [-1,1]. In the beginning when there is no information about the learner, the value of the scores is (0). Another approach is in the beginning of the testing each learner to starts with score (1), i.e. "A". During the testing, the learner strives for keeping this score to the end. Similar approach is the starting score of learners to be (-1), i.e. "F", and during testing all the correct answers given by learner improve his/her score.

For each type of error are supported two counters [5,6]:

- for the number of cases when it would be possible an error to be made;
- for the number of wrong attempts.

After the end of the event for each error types are formed scores by the following formula:

$$Score\_medial=1-Number\_wrong\_attempts/Number\_all\_attempts$$

Score_medial is a number in the interval [0,1] and represents the probability the attempts of that user to be without errors. The final score for the level of assimilation of given knowledge is in the interval [-1,1] and is calculated by following formula:

$$Score\_final=2*Score\_medial-1$$

Bearing in mind the formulas given above the common score is:

$$Score=1-2*\ Number\_wrong\_attempts/Number\_all\_attempts$$

When the number of attempts is 0 the score is 0 too. In the end of the event the score of the type of error is formed. The influence of this score on the model [4] of user is given by the formula:

$$Score\_new=(1-K)*Score\_current+K*Score$$

K is a number in the interval [0,1] representing the influence of a score of the error type over the score in the model of the user. K is calculated as follows:

$$K=0.5*(Number\_all\_attempts-1)/\ Number\_all\_attempts$$

Thus, the new score for the given question decreases as the number of unsuccessful attempts to find the correct answer for this question increases.

Apart from that the score is multiplied by 0.5 so that the influence of the old score over the new one to be limited. The score of each skill (the level of assimilation of given knowledge) consists of information about the scores of previous questions and about the last question.

The tests are generated on the base of randomized selection from large question banks. In the test generation algorithm have to be included a rule for the sequencing of questions so that a difficult question to be given only to the user who on the previous question with average difficulty have answered correctly.

The average of scores for the different skills and for the different error types forms the final result.

The assessment of the level of assimilation for the given knowledge unit is according the formula:

$$Score\_new=(1-K)*Score\_current+K'*Sorce\_test$$

Score_test is a score, which is based on the established educational criteria. This score is given by expert/teacher or by expert system. K' represents the influence of the Score_test over the knowledge unit (KU). K' is calculated as follows:

$$K'=Test\_difficulty/KU\_difficulty$$

The scores from the interval [-1,1] are interpreted in the following manner. If the score is near 1 then the system is more sure that the learner have assimilated correctly the corresponding knowledge unit. If the score is near –1 this means that the knowledge unit has been wrongly assimilated.

In other words, if the score is a positive number then it represents the level of learner's knowledge compared with the knowledge of the expert. In case that the score is negative then it represents the wrongly assimilated knowledge [5].

In respect to the equality of the learners, the tests of all of them have to consist of equal number of questions.

Conclusions

During the testing the student are given some hints and additional instructions. Apart from the test score the students receive the correct answers for the test questions and references to the learning materials and additional learning resources. The described approach is appropriate for student self-assessments in the process of learning.

References

[1] Wiggins, Grant (1998) Educative Assessment, San Francisco: Jossey-Bass Inc., pp.7

[2] Knowles, J. O'Keefe (1999) Computer assisted assessment. Computer assisted assessment Center, http://www.le.ac.uk/TALENT/book/c3p2.htm

[3] Haladyna, Thomas M. (1999). Developing and Validating Multiple-Choice Test Items. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

[4] Barr A., Feigenbaum E., The Handbook of Artificial Intelligence, vol.2, 1984.

[5] Langova D., K. Paskalev, Cognitive modeling in intelligent tutoring systems, Contemporary trends in the development of fundamental and applied sciences - VI national conference of the Bulgarian Scientists union, Stara Zagora, 1995, pp. 279-284.

[6] D. Orozova, H. Aladjov, K. Atanasov, Generalized Net For Machine Learning With Current Estimations, Advanced Studies in Contemporary Mathematics, Ku-Duk Press, 2001, Vol.3, No.2, pp.61-76.

DESIGN OF E-LEARNING CONTENT – APPROACHES AND METHODS

Maria Monova-Zheleva

Burgas Free University
101 Alexandrovska St, Burgas 8000, Bulgaria
e-mail: mariaj@abv.bg

Abstract: Learning Technologies have been evolving over the last decades, and have gone through many phases and approaches: early mainframe based programmed systems, microcomputer software packages written in native programming languages for specific machines, CBT systems, authoring systems, and more recently after the Internet explosion, web-based systems, and e-Learning Systems. Objective of this paper is approach for e-Learning content design by combining concepts and methods of didactical design, information, and interaction. In this paper a complex method is presented, which helps to structure the learning content design process, to reduce the complexity of the design tasks. Some problems concerning the design and development of Web-based learning objects are also described.

Keywords: Contextual Analysis, Didactical Design, Learning Object, Learning Object Reusability, and Interoperability e-Learning Standards.

Introduction

There is an agreement among teachers, educators, psychologists, designers, and content developers that the learning content has to be structured in an instructionally sound manner to facilitate the learning process. This means that didactical and software-usability approaches have to be combined.

The modern pedagogical and psychological theories such as “open learning”, “constructivism” and “activity theory”[1,2] have to be the didactical foundation for the next generation of e-Learning software. Another important requirement is the collaborative learning strategies to be supported.

There is growing consensus around the object-based approach for constructing e-Learning content. Learning objects can be re/used in multiple learning contexts, which increases the flexibility of training, and makes

updating courses much easier to manage [3-5]. Apart from the object-oriented approach a method for supporting the design of a space and time dependant didactical information continuum consisting of hypertext, audios, videos, 3D-animations, simulations etc. is necessary. Such a method should be based on standard usability engineering methods like “contextual design”[6], “usability engineering lifecycle”[7], and the software ergonomics design method - MUSE [8].

The described method aims at reducing the complexity of the design decisions by a systematic context-oriented analysis [9] and by distinguishing two levels of abstraction – macro design level and micro design level.

Context-oriented analysis

The first important step of the context-oriented analysis is the identification of appropriate learning contents. The appropriate content is the content, which is difficult to teach with traditional teaching methods. Such content is especially appropriate for being realized within the context of an e-Learning environment.

The problems and difficulties of teaching the identified content have to be considered in detail. The next step is a detailed analysis of the didactical conditions. The main aspects of this analysis should be “educational organization”, “subject matter”, “learning environment” and “learner”. Another important aspect of the context-oriented analysis is the interoperability issue. The result of the context-oriented analysis is the vision on realizing the identified content in a virtual learning environment.

Macro design level

The main objective at this level is to combine global design decisions concerning the didactical concept and the results of the context-oriented analysis. The future users of the learning content have to be determined. At this level it is not possible to identify the individual characteristics of every single user. The users are categorized into groups by their roles during the work with the content. Some typical roles are learner (advanced and beginner), trainer/instructor, author, administrator.

The next step is the learning tasks, which the users shall perform with the content (in their certain roles), to be linked to learning objectives, which describe the desired behaviour change of the learners. Every task processing is subdivided by visible or invisible work schedule so that the

learners can control their learning processes. The learning content can be developed with the help of the methodical instrument “guide questions”[8]. In didactical concepts with an accent on learner activities the focus is on the active work with learning objects [4,5], which can be used and reused in changing educational context. In usual web-based learning materials these possibilities are normally not taken into account.

An example for a learning content conceptual architecture is given below.

The conceptual architecture is based on a four-level hierarchical conceptual scheme of the multimedia training materials [16]. According to this scheme, the courses in the Web-based learning application contain the following courseware elements:

- Course, consisting of one or more modules.
- Module, consisting of one or more lectures.
- Lecture, consisting of one or more units.
- Unit – a sequence of one or more sub-units. Each sub-unit may be of one of following semantic types:
  - Definition element, introducing new concepts, object attributes or actions. It is presented often by text files, but it may contain also graphic and multimedia files.
  - Example element, implementing e-demonstration/s, which is presented internally by graphic and multimedia files of the accepted formats.
  - Test element, implementing classical assessment forms, which is presented internally mainly by text files.
  - Exercise element, introducing task/s to be executed by the learner. It may be presented internally by all the accepted file types.

Technically these four types of semantic sub-units may be represented through any one (but only one) of the following information types: text, picture, sound, animation, and movie. The types and numbers of the integrated files in a unit are not limited.

One lecture can either be an independent HTML document or consist of a series of *objects* (units), which appear in a specific order. These objects can also be references to other documents in the World Wide Web. These references can point to some material related to the current lecture page material. A lecture page can be associated with a video or an audio file, or

both of them, which describe the lecture page as a whole. But this does not mean that there is a restriction of the number of multimedia files that can be used in the content of a lecture page also.

The designers have to model the logical relations. This means they have to define which transitions shall exist between every two pair of hyper-nodes. The meaningful sequences of nodes have to be determined so that the learners can process their learning tasks. In this context, the support of versatility and multiple versions of the learning content is a very important requirement.

After the conceptual model the "opportunities for action" model has to be created. This model has to describe the stages of the didactical process and to represent user's navigation by specifying the logical and temporal transitions between hyper-nodes.

For each of the described objects (learning tasks, learning objectives, guide questions, learning content, users, hypermedia-elements) the relevant attributes have to be documented according to the e-Learning interoperability standards. Common standards for metadata, learning objects, and learning architecture are critical for the success of the knowledge economy. In this context interoperability e-Learning standards promote reusability (inter- and intrareusability) [5], customization, and personalization of learning material as well as providing valuable services to all the users of the e-Learning environment including learners and trainers. The Dublin Core Metadata [11], IEEE LOM [12], Educom's IMS [13], ARIADNE [14], AICC [15], and ADL/SCORM [16] are the most important initiatives dealing with interoperability standards and specification promoting.

In regard to presentation view at this level, the global decisions concerning the user interface presentation have to be made. The necessary input and output hardware must be determined also.

Micro design level

At this design level, the micro design of each hyper-node has to be created on the basis of the results on the macro design level. The user roles are inherited from the macro to the micro level. For each user role, specialized learning tasks within the task of the macro level are determined. Specialized learning objectives and specialized learning objects are connected with the specialized learning tasks.

When the components of hyper-nodes are developed in detail at the micro level it is possible the designers to find out that some additional transitions have to be modeled. In this case they have to go back at the macro design level in order to add the missing transitions in the macro-framework.

As on the macro level for these specialized tasks, objects and objectives, the relevant metadata have to be defined. The possibilities of actions are realized through software tool functions. Gorny [8] differentiates four function types:

- application functions - the learner's actions on the learning objects and additional software functions, which are used,
- adapting functions - allow the teacher to adapt the software to specific groups of learners and specific methodological and organizational conditions, and to adapt the software to individual learner requirements,
- control function - control the computer system,
- meta-functions - support the usage by „help information“, warning and error messages etc.

The support of the group learning and teamwork is a big advantage of the web-based learning environments. The problem is that the tool functions have to be available simultaneously for a group of learners. In the same time, when such a function is activated several users will be influenced by this action. During the teamwork, the realization of standard “undo” function is a real challenge because some members of the team might have changed the state of their work with the learning content in the meantime.

Concerning the interactions there are two main tasks at the micro level. The first is specifying the dialogue types for the different tool functions. The second task is describing the structure of the dialogues. The following dialogue types can be distinguished: command dialogue, data input dialogue, multiple-choice dialogue and object manipulation dialogue [5]. The combination of these types forms complex dialogue structures such as dialog panels, forms etc.

The second task is related to the definition of the permitted order for the call of tool functions. To fulfil this task it is necessary the preconditions of activating of a given tool function to be specified and the situation after execution of the function (the postcondition) to be described. In case of teamwork the co-ordination, co-operation, access conditions and access rights have to be specified also. Thus, the work of multiple users with

common objects will be without access collisions.

At this level all elements of the content are completely described and the final presentation layout is developed.

Conclusions

The described method is based on iterative procedure. The design process is presented with two levels of abstraction (macro and micro level).

Each of these abstraction levels is described from conceptual, interaction and presentation point of view.

At each level the designers have to take into account the interoperability e-Learning standards and specifications as well as the relevant didactical requirements and criteria.

References

[1] Dochev, D., R. Yoshinov, R. Pavlov, Open Classrooms in the Digital Age, IV EDEN Conference, Conference proceeding, Barcelona, Nov. 19-21, 247-252 p., 2000.

[2] Evans, T., New research challenges for technology supported learning, Report on the Open Consultation Workshop and Process, Wavecrest Systems Ltd., 2001

[3] Longmire, W. A Primer on Learning Objects. Learning Circuits, 2000 http://www.learningcirquits.org/mar2000/primer.html

[4] Wiley, D. A. Connecting learning objects to instructional design theory. A definition a metaphor, and a taxonomy, D. A. Wiley (Ed.), The Instructional Use of Learning Objects. Bloomington, IN: Association for Educational Communications and Technology, 2001. http://works.opencontent.org

[5] Wiley, D. A. The instructional Use of Learning Objects, 2002, http://reusability.org/read/

[6] Beyer H., Holtzblatt K. Contextual Design, Morgan Kaufmann, San Francisco, 1998.

[7] Mayhew D. The Usability Engineering Lifecycle, Morgan Kaufmann, San Francisco, 1999.

[8] Gorny P. EXPOSE – HCI-counseling for user interface design, in Nordby, K et al (Eds.)"Human - Computer Interaction, Proc. Interact'95", Chapman&Hall, London, pp.297-304, 1995.

[9] Donker H. Usability Engineering of eLearning Software, Work with Display Units - WWDU 2002, Berlin, pp. 272 – 274, 2002.

[10] Dochev D., Pavlov R., Monova-Zheleva M. Principles, Quality Requirements and Solutions for On-The-Job E-Training in SME. EDEN Annual Conference

"The Quality Dialogue — Integrating Quality Cultures in Flexible, Distance and eLearning", Conference Proceedings, pp.518-523, 2003.

[11] Dublin Core Metadata Initiative http://purl.org/dc/

[12] IEEE LOM: Institute for Electrical and Electronic Engineers Learning Technology Standard Committee learning Object Metadata http://tsc.ieee.org/doc/wg12/LOM3.6.htm

[13] IMS: Instructional Management System Global Learning Consortium, Inc. IMS Learning Resource Meta-data Best Practices and Implementation Guide. http://www.imsproject.org/metadata/mdbest01.html

[14] ARIADNE: Alliance of Remote Instructional Authoring and Distribution Network for Europe, Educational metadata. http://ariadne.unil.ch/Metadata/

[15] AICC: Aviation Industry CBT Committee, http://www.aicc.org/pages/primer.html

[16] ADL/SCORM: Advanced Distributed Learning Network, Sharable Content Object Reference Model, http://www.adlnet.org/Scorm/

INFORMATION LOGISTICS SIMULATION METHODOLOGY AND METHODS OF INFORMATION TECHNOLOGY AND THEIR LOGISTICS

Svetlozar Kabaivanov

VTU"St.st. Cyril and Methodius", Veliko Turnovo
sv.kabaivanov@abv.bg

Abstract: Information Logistics organizes the flow of data which accompanies and which attends to the material flows of commodities and utilities. The importance of information logistics finds expression in management of basic subsystems of different types of organizations.

The purpose of the current paper is to express opinion on the place and application of information logistics into the new information and communication media and the environment of new forming theory and practice of selection, training, teaching and preparation of organization sections in structures of different levels.

Keywords: information logistics, information system (IS), simulation application, certification

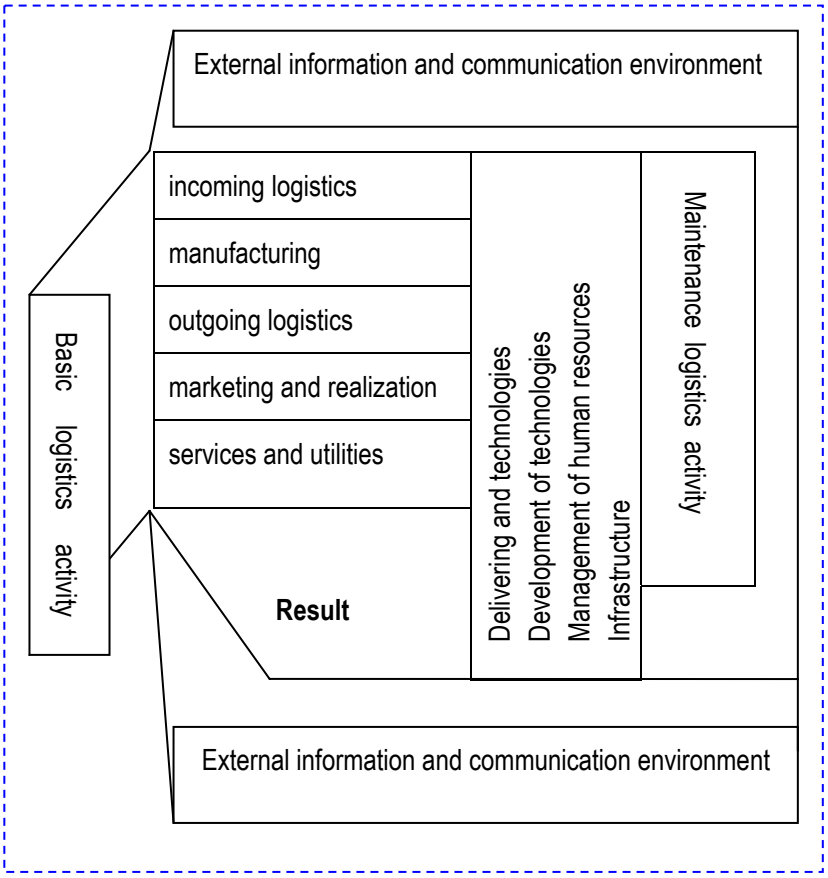
The rapid changes in information system (IS) technology are presenting firms with significant challenges and dramatic opportunities. Revolutionary advances in hardware capability coupled with nearly free-fall in prices have shifted numerous applications across the threshold from infeasible to feasible. Concurrently, important advances in software development methodologies and tools have encouraged the construction of many previously infeasible systems. In addition, the changing structure of organizations, specifically the trend toward flatter structures and the accompanying shifts in information requirements, plus the move to increased user involvement has significantly altered the types of IS being implemented.

The term "logistics" originates from the Greek word "logistike", which means the art of reasoning and calculation. One definition of logistic sounds like that: logistic – a science of planning, organization, management and control of movements and implements of material and information flows in tract and time from their primary sources to the ultimate users. According to

the US council of logistics management recently occurred the opinion that "logistics" is synonymous of management of materials and distribution. [1.]

The analysis of the results of using the logistics as a management conception indicates that it brings to new competition advantages, to form incomes of resources and means of production intended for economic subjects and also for small and big regions and states.

An attempt at showing the generalized nature of management of information resources and processes is presented in the table:



An Indefeasible part of logistics is represented in logistic information flow which includes input and output data. This subsystem of logistics is called information logistics. To create of information logistics system through

different levels into different activities it is necessary to establish a model of such system.

Computing modeling and simulation offers methods and composition of instruments for set up, probation, reproduction and application of such models.

Certification of modeling and simulation (M&S) applications poses significant technical challenges to M&S program managers, engineers and practitioners. Manage ring such measurements and evaluations requires a unifying methodology and should not be performed in an ad hoc manner. The methodology consists of the following body of methods, rules and postulates:

- Employment of subject matter experts.
- Relative criticality weight of indicators using the analytic hierarchy process.
- Hypertext certification report and interpretation of the results and others.

Subject matter experts are commonly employed for M&S evaluation for certification. Under the methodology, the knowledge and experience of them are utilized for:

- constructing a hierarchy of indicators,
- relative criticality weight of indicators,
- building a rule-based expert knowledge base
- assigning scores for the indicators.

Such indicators are tools used in evaluation and concern simulation methods.

A model is a representation or abstraction of something such as an entity, a system or an idea. Simulation is the act of experimenting with or exercising a model or a number of models under diverse objectives including acquisition, analysis and training. For example, if the analysis objective is to predict the performance of a complex system design, we experiment with a model or a distributed set of models representing the system design. If the predicted performance is used in making an acquisition decision, the process is called simulation-based acquisition.

On algorithm of processes basis can propound the following three known methods:

Naïve method: Generate the full sample and order it to find the required order statistic. Thus a generator for the original distribution and sorting, or a

faster algorithm for finding the same order statistic is needed. Finding the maximum or minimum reduces to other symbol comparisons.

Inversion: Generate the corresponding uniform order statistic and transform it with the inverse cumulative distribution function (CDF). A generator for the beta distribution and an algorithm to invert the CDF of the original distribution is needed. We used transformed density rejection (TDR) to generate from the beta distribution and numerical inversion using Newton's method and a table of size 1000.

Quick elimination (QE): The method works only for maximal or minimal. An algorithm to sample from the given distribution, restricted to a half-open interval, is necessary. (We used transformed density rejection algorithms to accomplish this task).

The implement of these instruments is based at strategic information system (IS) planning. Strategy identification emerged in the early 1900s as a formal business concept. However, it was not until the 1970s that strategic planning emerged as a discipline. Today, determining the future direction of an organization is often called by different names, depending upon the organizational level at which the planning exercise takes place. The new discipline Information logistics and its simulative methods also take place at this area of knowledge and practice. Strategy of business and other different kind of enterprises lay on planning, which is a product of superstructure based on previous experience and imaginary prognoses. For example, the typical business firm normally engages in three levels of strategy formulation, which in effect forms a strategic management hierarchy. They are:

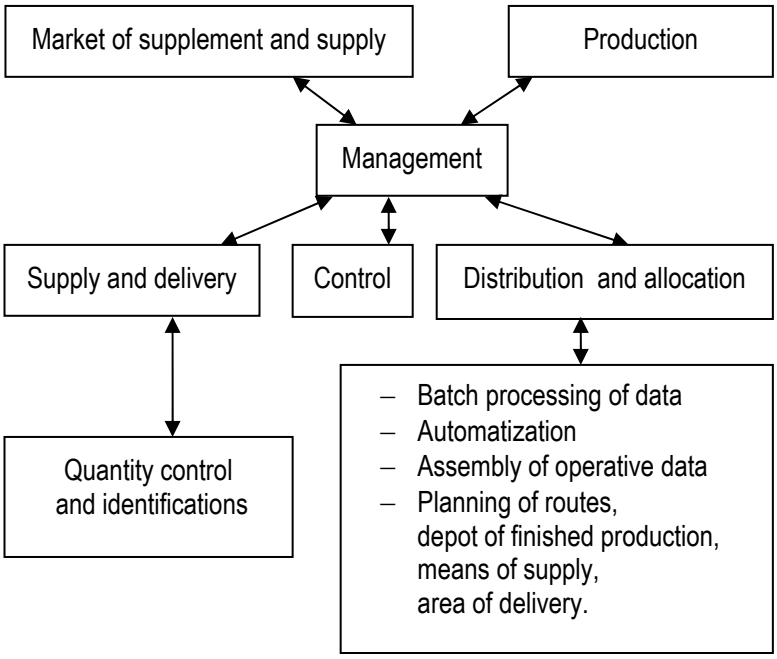
- corporate strategy at the headquarters level;
- business strategy at the level of the business unit;
- functional area support strategy at the level of the functional departments such as finance, marketing and information systems.

During the 80s took place a classification of IS planning within two planning contexts: information resource and information function. The information resource context is an approach to IS planning that addresses the management of a firm's information system using an organization-wide perspective. It is concerned with the deployment of information services in support of `organization function approach and takes a more traditional technology approach to planning, dealing mainly with the technical aspects of establishing and managing the firm's information systems. It is 'concerned with processes by which IS products are made available, i.e.,

the activates associated with identifying, selecting, and implementing IS products.

Set forth below is represented an example of organization of information logistics net of commodity production, founded on suggestion based on simulation of real organization.

Organizations are now faced with an opportunity to improve their utilization of information systems technology by improving the effectiveness of the planning efforts. The evolution of information systems planning has coincided with the progressive assimilation of information technology in organizations. The IS planning concept has assumed different context within this evolution, aligning itself with organizational needs and assuring an evolution of purpose.



There are more aspects of implementation of simulation methods, not only in economics and business structures but in military affairs, too.. Simulation also gives military and political leaders insight into potential conflicts. Commanders can now recreate on computer the complex choreography of thousands of soldiers, weapons, vehicles, and aircraft moving across a battlefield that extends over thousands of square

kilometers. In this way, military decision-makers can test strategic options before launching a campaign in earnest. They can also test the performance of new weapons systems under consideration. Even as computer simulations achieve greater realism, military operations themselves have become more computer-driven and more synthetic.

"The push toward training simulation has also spawned a huge industry. According to the trade publication Military Training and Simulation News, the U.S. Department of Defense spends about US $4 billion each year on simulation and training equipment. No other country has invested nearly as much."[5] Live training is still the most common way of readying troops. The U.S. Navy pioneered this approach after combat data revealed that a pilot's first few encounters with the enemy tended to be the deadliest; those who survived that early phase were more art to survive in general. All the U.S. armed forces now have similar live-training sites. In recent years effectiveness has been boosted by advanced electronics and instrumentation. At the urban warfare training village at Fort Polk, La., for instance, video cameras record the action as Army Rangers and Marines fight building to building and room to room. A form to laser tag called Miles- for Multiple Integrated Laser Engagement System – identifies hits and misses.

A powerful simulation-based decision tool should provide capabilities such as monitoring and steering. Monitoring provides the ability to query or sample simulation variables. These variables may be sampled continuously, displayed when a pre-defined condition is satisfied or presented on demand.

We define a simulation as being composed of a simulation application (provided by the user) that interacts with a simulation executive. The simulation executive provides primitives that allow simulation programmers to define their own applications and it implements the necessary underlying synchronization protocols. This is a layered software system, with the operating system at the bottom, the simulator executive in the middle and the simulation application.

Simulation is becoming an important tool for decision makers in time-constrained environments. High-fidelity simulations demand parallel implementations to support the decision process at real-time rates. Typical decision making applications in which parallel simulations can or do play a role include air traffic control, gaming strategy and battle management among others. The goal of this research is to provide a simulation-based decision aid for managers faced with complex planning tasks. To realize

effective decision making we propose a technology called simulation cloning that enables more efficient exploration of different possible future outcomes based on policy decisions made at well defined decision points.

References

1. Bowersox,D. D., Closs, Logistical Management, McGraw-Hill International edition, 1996. [quote4]
2. Logistics, textbook, Anikin B.A and others, Moscow, Infra-M, 2000
3. Osman Balci, A Methodology for Certification of Modeling and Simulation Application, ACM Transaction on Modeling and Computer Simulation, October 2001, volume 11, number 4
4. Philip Todorov, Distribution Politics, “Trakia-M”, Sofia, 1999.
5. Michael Macedonia, U.S. Army Simulation Training and Instrumentation Command, IEEE SPECTRUM, march 2002, p. 33.

Second Workshop on Multimedia Semantics

WEB PAGE RETRIEVAL BY STRUCTURE

William Grosky[1] and Gargee Deshpande[2]

1 - University of Michigan-Dearborn,
Computer and Information Science Department,
4901 Evergreen Road, Dearborn, Michigan 48128, USA;
email: wgrosky@umich.edu

2 - Wayne State University,
Computer Science Department,
Detroit, Michigan 48202, USA

Abstract: Our research explores the possibility of categorizing webpages and webpage genre by structure or layout. Based on our results, we believe that webpage structure could play an important role, along with textual and visual keywords, in webpage categorization and searching.

Keywords: content-based retrieval, genre detection, layout ontologies.

1. Introduction

The amount of data available electronically on the web has increased dramatically in recent years. Users generally retrieve data by browsing and searching by keywords. This is an example of *content-based search*. In this approach, search is based on the words in the heading of the page or the contents of the images displayed on the web pages, or words occurring as meta-data in pages. The overwhelming amount of information on the web requires a powerful search service to render that information accessible and useful. Without such a search strategy, finding a specific web site can be as difficult as finding a book in a library that has no card catalogue and a completely random method of storing its books.

In recent years much research has been done on querying the web. In this research, the web is viewed as a collection of multimedia documents in the form of pages connected through hyperlinks. Unlike most web search engines, the aim here is to provide more database-like query functionality. Also, application of data mining techniques to the World Wide Web, referred to as *web mining*, has been the focus of several research projects and papers. Web mining has been categorized into *web content mining* and *web usage mining*. Web content mining is the process of finding information from

the web, whereas web usage mining is the process of mining user browsing histories for access patterns [1].

We believe that it would also be desirable to see the layout of web pages when querying these pages and grouping them according to these layouts. The term *layout* connotes the spatial relationships between the page contents rendered by particular tags. Thus, web pages can be categorized according to their *layout ontology*. The term *ontology* means a specification of a conceptualization, a set of concept definitions. Broadly speaking, an ontology is a description (like a formal specification of a program) of concepts. Each web page has a structured hierarchy of tags that defines the layout ontology for that particular page. It is possible that two different web pages have a similar structure of their tag hierarchy. Then, the layout ontology of these two web pages is said to be the same. Our aim is to categorize web pages according to these structures. Our belief is that pages with similar layout ontologies have somewhat similar semantics, or at least can be categorized as belonging to the same environment. For example, we will present some preliminary experiments that show that pages from different newspapers are more similar in layout ontology to each other than to the layout ontology of commercial sites selling books. This concept can also be used for extracting data from the web depending upon content as well as the structure.

In this paper, we assume that each web page consists of HTML tags. HTML tags can be broadly categorized as *container tags* and *standalone tags*. Container tags can contain other tags inside them but standalone tags are atomic. Because of the containing capacity of some tags, each web page can be represented by a tree structure of its tags. Thus, for each web page, a tree structure of tags can be determined. Our hope is that we get a somewhat different tree structure of tags for each page from different environments.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related literature. Section 3 covers various conceptual details, while Section 4 discusses implementation details and the various technologies used in our experiments. In Section 5, we give the results of some preliminary experiments, while Section 6 gives some concluding remarks.

2. Literature Review

Our hope is that by automatically characterizing the environment of a particular web page, content-based information can more easily be extracted from it. In [2], information from unstructured and semistructured web documents is retrieved from web pages in chunks called *records*. A record is a group of information relevant to some entity. The final goal is to extract information from these records to populate a relational database. The paper describes a heuristic approach to discovering the record boundaries in web documents. It captures the structure of a document as a tree of nested HTML tags and locates the sub-tree containing the records of interest, identifying candidate separator tags within the sub-tree using five independent heuristics, finally selecting a consensus separator tag based on a combined heuristic.

The five heuristics are OM (ontology matching), SD (standard deviation), IT (identifiable separator tags), HT (highest-count tags), and RP (repeating-tag pattern). Each of these heuristics returns one or more candidate separator tags with a measure of certainty attached to each candidate. Finally, they provide a way to combine these individual heuristics to determine a consensus separator tag and hence discover record boundaries.

The technique we exploit in this paper is based on the work of [3]. In this paper, a computational geometry-based spatial color indexing methodology is examined for efficient and effective image retrieval. In this scheme, an image is evenly divided into a number of M*N non-overlapping blocks, and each individual block is abstracted as a unique feature point labelled with its spatial location, its dominant hue and its dominant saturation. For each set of feature points labelled with the same hue or saturation, a Delaunay triangulation is constructed and then a feature point histogram is computed by discretizing and counting the angles produced by this triangulation. The concatenation of these feature-point histograms serves as the image index. This research work has been the motivation for our research.

Related research field to our approach is the research being done on semistructured data. For retrieving web pages by structure, structures of web pages have to be stored and retrieved effectively. For storing semistructured data, paper [4] argues that languages supporting deduction and object-orientation are particularly well-suited, as object-orientation provides a flexible common data model for handling semistructured data.

Paper [5] presents the Lorel language designed for querying semistructured data. The main novelties of Lorel are that it makes extensive use of coercion to relieve the user from the strict typing of a query language, which is inappropriate for semistructured data, and that it provides powerful path expressions, which permit flexibility for declarative navigational access.

As against the data model that is underlying [5], [6] argues that semistructured data can be stored in relational format by exploiting the regularities inherent in existing semistructured data instances. The claim is that most of the data will be stored in relational format and future insertions can occur in a self-describing way. In [7], an approach of creating wrappers for storing semistructured data is discussed.

3. Web Page Retrieval by Structure

Motivated by the ultimate goal of automatically computing efficient and effective descriptors which symbolize web page structure, this research has been directed towards the management of information such as the levels of tags comprising a web page, the tag hierarchy, and the area covered by the tags on the web page. As nesting of tags plays important role in defining structure of the web page, dominance of tags is considered for each level.

Hope is that within broad domain of web pages this technique can be used to find the structure of web pages and categorize web pages according to the structure. Further to such categorized web pages, semistructured techniques can be applied for effective content retrieval.

The paper [2] has been the motivation behind this research. This paper examines the use of a computational geometry-based spatial color indexing methodology for efficient and effective image retrieval. In this scheme, an image is evenly divided into number of M*N non-overlapping blocks, and each individual block is abstracted as unique feature point labelled with its spatial location, dominant hue, and dominant saturation. For each set of feature points labelled with the same hue or saturation, a Delaunay triangulation is constructed, followed by computing a feature point histogram realized by discretizing and counting the angles produced by this triangulation. The concatenation of all these feature point histograms serves as the image index.

Following the same concept, we examine the use of a computational

geometry-based web page structure analysis for effective web page structure matching. In this scheme, a web page is evenly divided into number of M*N non-overlapping blocks, and each individual block is abstracted as a unique tag that covers the maximum area in that block at its level. For each feature tag selected we get a set of feature points. For each set of feature points labelled with the same tag, we construct a Delaunay triangulation and then compute the feature point histogram as mentioned above. The concatenation of these feature-point histograms serves as our web page descriptor. Web page descriptors are further used to categorize different web pages.

As mentioned previously, we assume in this paper that each web page consists of HTML tags. HTML tags can be broadly categorized as *container tags* and *standalone tags*. Container tags can contain other tags inside them but standalone tags are by themselves. Examples of container tag are the TABLE tag and the PARAGRAPH (P) tag, while examples of standalone tags are the BASE tag and the AREA tag. Because of the containing capacity of the tags, a web page corresponds to a tag tree structure, called a *tag tree*. Not all web pages have similar tag trees. In this paper, we study page layouts to try to categorize web pages semantically.

For our analysis, the level of a tag plays an important role when finding tags covering the maximum area in a block. An example of a web page tag hierarchy is as follows:

```
<HTML>
    <HEAD>
        <TITLE>
        </TITLE>
    </HEAD>
    <BODY>
        <P>
            <TABLE>
                <TR>
                    <TD>
                    </TD>
                </TR>
            </TABLE>
            <B>
            <B>
        </P>
```

```
    </BODY>
</HTML>
```

In the web page example given above, the <HTML> tag is at the highest level. Nested in the <HTML> tag are tags <HEAD> and <BODY>. Inside the <BODY> tag is a <P> tag and inside the <P> tag is a <TABLE> tag and so on. When we consider the concept of area covered by a tag on a web page, the concept of level plays an important role. In the example given above, the level of tag <HTML> is 1, the level of the <BODY> tag is 2, the level of the tags <TABLE> and <B> are 3, the level of tag <TR> is 4, and so on. When calculating the dominant tag at level 3, both <TABLE> and <B> tags are analyzed to check which tag is covering the maximum area in which block on the page. As tag <TR> is inside the <TABLE> tag, the area covered by the <TABLE> tag on the web page contains the area covered by the <TR> tag. So, for the blocks in which the <TABLE> tag is dominant at level 3, it is possible that in this same block, tag <TR> is dominant at level 4.

Now, each web page consists of tag hierarchy. We consider a few tags as characterizing features $F = \{f_1,\ldots,f_k\}$. We believe that the spatial placement and dominance of these various feature tags can be used to characterize the web pages.

The web page is divided into N*M non-overlapping blocks. For each block, at each level, the tag covering the maximum area is found. Then we find for each of the predefined feature tags, which blocks that tag was marked as the predominant tag. We mark the center co-ordinate of all such blocks. The spatial arrangement of these points is an important aspect of our work. As mentioned earlier, we construct a Delaunay triangulation and then compute the feature point histogram by discretizing and counting the angles produced by this triangulation. The concatenation of these entire feature-point histograms serves as our web page descriptor.

It has been shown that histogram intersection is especially suited for comparing histograms for content-based retrieval. Additionally, histogram intersection is an efficient way of matching histograms. The intersection of the histograms W_{query} and $M_{database}$, each of n bins, is defined as follows:

$$D(W_{query}, M_{database}) = (\sum \min(W_j, M_j)) / \sum W_j$$

The histogram of a web page characterizes the web page depending upon the placement of tags forming the web page. Thus, the above-mentioned formula can be used to check the similarity between the two web pages. If two web pages are similar in structure then the histogram of those two pages are bound to be similar. For such web pages, the above formula

returns a value close to 1. Similarly, if two pages are very different in structure then the above formula returns a value close to 0.

4. Implementation

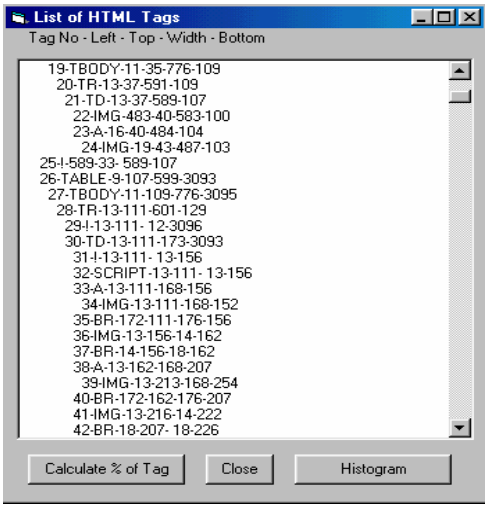
The input to our system is a web page. Our feature representation is extracted from this web page and matched against those extracted from other web page of known semantics. In more detail, we do the following:

1. Our system accepts a URL as input and displays the given web page using the Internet Explorer engine.
2. The web page displayed is analyzed to get all tags on the page with left, top, right, bottom (X1, Y1, X2, Y2) co-ordinates of area covered by each tag on the page. For each tag, the level of nesting is also saved while gathering this data.
3. The page is normalized to size 512 * 512. The original calculated co-ordinates (X1, Y1, X2, Y2) are re-calculated to map to this normalized size.
4. The page is divided into N*M disjoint blocks. The relevant coordinates of each block is calculated.
5. For each block, it is found out that which tag covers how much area.
6. Depending upon the data gathered in step 5, it is found out for each level, for each block, which tag is covers the maximum area.
7. For each of the selected feature tags, the blocks are found in which the tag covers the maximum area. Center X and center Y coordinates of these blocks are written to a file.
8. Histogram program is run on the file and histogram points calculated by the program are read back into the system. The histogram program used to calculate these points is implemented for the two largest angles of each Delauney triangle using 36 bins. Thus, each bin corresponds to 5 degrees.
9. For each web page, descriptor of (36 * Number of feature tags) bins is calculated.
10. When the descriptors of all the web pages of interest are calculated using steps 1 through 9, the distances between these pages and the database pages are calculated.

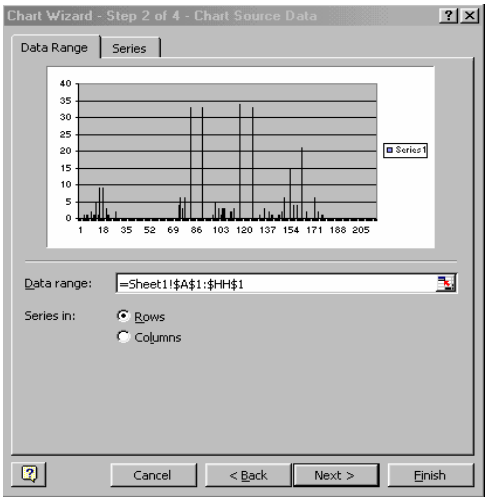
11. For each query page, the nearest pages are chosen, based on the distances calculated in step 10.
12. The web pages selected in step 11 are analyzed for category information. The most occurring category is chosen. The query page is categorized using this category.

As an example, consider the following web page:

Here is a snapshot of part of the tag tree, along with the coordinates of the rectangular area covered by the rendering of each tag:



And here is the resulting histogram for the web page:



5. Experimental Results

Our proof-of-concept experiments are carried out on newspaper web pages and e-commerce web pages. Four newspapers and two e-commerce web sites are selected as categories. The categories are: Detroit News, Times of India, Tribune India, Esakal, Amazon.com, and Buy.com.

For each of the newspaper categories, six days of newspaper front pages were analyzed, while from the e-commerce web sites, six web pages were used. Thus, a total of 36 web pages were analyzed.

Initially, we defined a large set of feature tags to ensure a powerful set of independent features for the discrimination of our two classes. This initial set of 52 feature tags were: <A>, <APPLET>, <B>, <BIG>, <BR>, <CAPTION>, <CENTER>, <CITE>, <CODE>, <COL>, <COLGROUP>, <DD>, <DIR>, <DL>, <DT>, <EM>, <FONT>, <FORM>, <H1>, <H2>, <H3>, <H4>, <H5>, <H6>, <HR>, <INPUT>, <LI>, <MENU>, <OBJECT>, <OL>, <OPTION>, <P>, <PRE>, <SELECT>, <SMALL>, <STRONG>, <SUB>, <SUP>, <TABLE>, <TBODY>, <TD>, <TEXTAREA>,<TH>, <TITLE>, <TR>, <U>, <UL>, <FRAME>, <FRAMESET>, <IMG>, <MAP>, <AREA>.

We also conducted an experiment using a reduced set of tags. For each tag, we calculated a mean descriptor, by computing bin averages over all 36 web pages. We then calculated the deviation of each descriptor from its mean. We only kept those tags with high deviations, as these tags more easily discriminate among the various pages. The tags we kept for this experiment were <FONT>, <STRONG>, and <IMG>.

In all our experiments, we compared each individual web page, using the nearest neighbour approach, to the 35 remaining pages, using both sets of tags. We tried to determine both individualized categories as well as genre categories. The former takes a match as successful only if the two pages came from the same site, while the latter takes a match as successful only if the two pages came from the same genre: newspaper versus e-commerce. Here is the table of our results.

	Individualized Categories		Genre Categories	
	Matches	Failures	Matches	Failures
52 tags	26	10	33	3
3 tags	27	9	33	3

Based on these initial results, it seems that our technique has promise for genre detection.

6. Conclusions

The aim of this research was to analyze the possibility of categorizing webpages and webpage genre by structure or layout. The original insight comes from the fact that many newspaper sites, say, have the same look and feel. Based on our results, we believe that structure could play an important role, along with textual and visual keywords, in webpage categorization and searching.

Bibliography

[1] R. Cooley, B. Mobasher, and J. Srivastava, 'Web Mining: Information and Pattern Discovery on the World Wide Web,' Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

[2] D.W. Embley, Y. Jiang, and Y.K. Ng, 'Record – Boundary Discovery in Web Documents,' Proceedings of the ACM SIGMOD Conference, 1999, pp. 467-478.

[3] Y. Tao and W.I. Grosky, 'Spatial Color Indexing Using Rotation, Translation, and Scale Invariant Anglograms,' Multimedia Tools and Applications, Volume 15, Number 3 (December 2001), pp. 247-268.

[4] B. Ludäscher, R. Himmeröder, G. Lausen, W. May, and C. Schlepphorst, 'Managing Semistructured Data with FLORID : A Deductive Object-Oriented Perspective,' Information Systems, Volume 23, Number 8 (1998), pp. 589-613.

[5] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J.L. Wiener, 'The Lorel Query Language for Semistructured Data,' International Journal on Digital Libraries, Volume 1, Number 1 (April 1997), pp. 68-88.

[6] A. Deutsch, M. Fernandez, and D. Suciu, 'Storing Semistructured data in Relations,' Proceedings of the Workshop on Query processing for Semistructured data and Non-standard Data Formats, Jerusalem, Israel, January, 1999.

[7] N. Ashish and C. Knobolk, 'Wrapper Generation for Semi-Structured Internet Sources,' SIGMOD Record, Volume 26, Number 4 (December 1997), pp. 8-15.

MPEG-7: THE MULTIMEDIA CONTENT DESCRIPTION INTERFACE

Peter Stanchev, David Green Jr., and Boyan Dimitrov

Kettering University, Flint, MI 48504, USA
pstanche@kettering.edu; dgreen@kettering.edu; bdimitro@kettering.edu

Abstract: In this paper a review of the most used MPEG-7 descriptors are presented. Some considerations for choosing the most proper descriptor for a particular image or video data set are outlined.

Keywords: MPEG-7, Multimedia, Content based retrieval

1. Introduction

More and more digital images and video are being captured and stored. In order to use this information, an efficient retrieval technique is required. One major development in this area is the content based image and video retrieval techniques which use image features for indexing and retrieval [Rabitti, 1989]. The most used features are color, texture, and shape. Several semantic image and video models are suggested [Stanchev, 1999], [Grosky, 2001]. In MPEG-7 standard different descriptors for this purpose are proposed [Manjunath, 2002]. What descriptor is the best for a particular data set? Some preferable answers of this question are given.

2. MPEG-7 Descriptors

The MPEG-7 descriptors can be classified as general visual descriptors and domain specific descriptors. The former include color, texture, shape and motion features. The latter includes face recognition descriptor. Although distance functions are not part of the standard we will present the most used distance functions. Only color, texture and shape descriptors are covered, since they are mostly used.

2.1. Color descriptors

Color is one of the most widely used image and video retrieval feature [Schettini, 2001]. The MPEG-7 standard includes five color descriptors which represents different aspects of the color and includes color distribution, spatial layout, and spatial structure of the color. The histogram

descriptors capture the global distribution of colors. The dominant color descriptor represents the dominant colors used. The color layout descriptor captures the spatial distribution or layout of the colors in a compact representation. While MPEG-7 standards accommodate different color spaces, most of the color descriptors are constrained to one or a limited number of color spaces for ensuring inter-operability.

2.1.1. Dominant color descriptor

This descriptor specifies a set of dominant colors in an image [Cieplinski, 2000]. It is good to represent color features where a small number of colors are enough to characterize the color information. The extraction algorithm quantizes the pixel color values into a set of dominant colors. The matching is done by calculating the distances between dominant color sets based on the difference between corresponding colors in any two sets of dominants.

The result of the method is a vector with integer numbers, presented as $F = \{(c_i, p_i, v_i), s\}$, (*i=1,2, …, N*), where *N* is the number of dominant colors. The vector components are: the dominant color value c_i (RGB color space vector); p_i - normalized fraction of pixels corresponding to color c_i; optimal color variance v_i, (describes the variance of the color values of the pixels in a cluster around the corresponding color); and the coherency s representing the overall spatial homogeneity of the dominant colors.

The distance algorithm uses an estimate of the mean square error, based on the assumption that the sub-distributions described by dominant colors and variances are Gaussian. Consider 2 descriptors:

$$F_1 = \left\{\left(c_1^{'}, p_1^{'}, v_1^{'}\right), s_1^{'}\right\} \qquad (i = 1,2,...,N_1) \qquad \text{and}$$

$$F_2 = \left\{\left(c_2^{'}, p_2^{'}, v_2^{'}\right), s_2^{'}\right\} \qquad (i = 1,2,...,N_2)$$

where

$$p \in [0,31], \quad c_i = rgb2luv(c_i^{'}), \quad v_i = \begin{cases} 600 & v_i^{'} = 0 \\ 900 & v_i^{'} = 1 \end{cases}, \quad p_i = \frac{\left(p_i^{'} + 0.5\right) / 31.9999}{\sum_i p_i},$$

and if

$$f_{x_i y_j} = \frac{1}{2\pi\sqrt{2\pi \times (v_{x_i}^{(l)} + v_{y_j}^{(l)}) \times (v_{x_i}^{(u)} + v_{y_j}^{(u)}) \times (v_{x_i}^{(v)} + v_{y_j}^{(v)})}} \times \exp\left\{-\frac{1}{2}\left[\frac{(c_{x_i}^{(l)} - c_{y_j}^{(l)})^2}{v_{x_i}^{(l)} + v_{y_j}^{(l)}} + \frac{(c_{x_i}^{(u)} - c_{y_j}^{(u)})^2}{v_{x_i}^{(u)} + v_{y_j}^{(u)}} + \frac{(c_{x_i}^{(v)} - c_{y_j}^{(v)})^2}{v_{x_i}^{(v)} + v_{y_j}^{(v)}}\right]\right\}$$

and if

$$D_v = \sqrt{\sum_{i=1}^{N_1}\sum_{j=1}^{N_1} p_{1_i} p_{1_j} f_{1_i 1_j} + \sum_{i=1}^{N_2}\sum_{j=1}^{N_2} p_{2_i} p_{2_j} f_{2_i 2_j} - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} p_{1_i} p_{2_j} f_{1_i 2_j}}$$

then the distance is calculated as: $D = [0.3 \times abs(s_1 - s_2) + 0.7] \times D_v$.

2.1.2. Scalable color descriptor

This descriptor performs color histogram in HSV color space encoded by a Haar transform [MPEG, 2002]. The extraction is done by quantizing the image into a 256 bin HSV color space histogram and then using the Haar transform to reduce the number of bins.

The output of the method is *a* vector from integer numbers, presented by a histogram with 64, 32 or 16 bins.

The distance matching can be done either in the Haar coefficient domain or in the histogram domain. In the case where only the coefficient signs are retained, the matching can be done efficiently in the Haar coefficient domain by calculating the Hamming distance as the number of bit positions at which the binary bits are different using an XOR operation on the two descriptors to be compared. This induces only a marginal loss in similarity matching precision compared to reconstructing the color histogram and performing histogram matching, while the computational cost is considerably lower.

2.1.3. Color layout descriptor

This descriptor performs spatial distribution of colors [Kasutani, 2001]. The extraction is being done as follows: the image is divided into 8x8 blocks. For each block, a single dominant color is selected. The resulting 8x8 image is then transformed into a series of coefficients using dominant color descriptors transformation. These are finally quantized to fit an assigned number of bits.

The method output is a vector from integer numbers, describing {*DY, DCr, DCb}* coefficients, where *Y* is the coefficient value for luminance, *Cr, Cb* coefficient values for chrominance.

For matching two descriptions *{DY, DCr, DCb}* and *{DY', DCr', DCb'}* the following formula:

$$D = \sqrt{\sum_i w_{yi}(DY_i - DY_i^{'})^2}$$

$$+ \sqrt{\sum_i w_{bi}(DCb_i - DCb_i^{'})^2} + \sqrt{\sum_i w_{ri}(DCr_i - DCr_i^{'})^2}$$

is used, where *i* represents the zigzag- scanning order of the coefficients.

2.1.4. Color structure descriptor

This descriptor is a generalization of the color histogram that encodes information about the spatial structure of the colors in an image as well as their frequency of occurrence [Messing, 2001]. The histogram is extracted in HMMD color space and non-uniformly quantizing is performed over the histogram values. This descriptor specifies spatial distribution of colors. It is calculated by letting a structuring element with image samples to visit each position in the image and then summarize the frequency of color occurrences in each structuring element location in a histogram. The structuring element always has dimensions 8x8, but the distance between the samples in the original image differs with the resolution.

The output of the method is a vector from integer numbers, presented by a 256 bin histogram.

The matching is done by minimizing the distance calculated as the sum of the differences between the corresponding bins in any two color-structure histograms.

2.1.5. Group-of-frame or Group-of-picture descriptor

This descriptor is a compound descriptor that expresses the color features of a collection of images or video frames by means of the scalable color descriptor [Ferman, 2000]. During the extraction the average, median or intersection scalable color histogram of the frame/picture group is calculated from scalable color histograms of each group/picture. The intersection histogram is a histogram with the minimum value for each bin over all histograms in the group.

The output of the descriptor is a vector from integer numbers, as in the case of scalable color descriptor.

The matching is done in the same way as for the scalable color descriptor.

2.2. Texture descriptors

The image texture is one of the most important image characteristic in both human and computer image analysis and object recognition [Manjunath, 2001]. Visual texture is a property of a region in an image. There are two texture descriptors in MPEG-7: a homogeneous texture descriptor, and edge histogram descriptor. Both these descriptors support search and retrieval based on content descriptions.

2.2.1. Homogeneous Texture

This descriptor is aimed at texture-based image-to-image matching [Ro, 2001]. During the extraction, the mean and standard deviation of the image pixel intensities is computed. Energy and energy deviation feature values are computed by applying 30 Gabor filters in the frequency domain. The polar form used in the frequency domain in this approach is more suited for rotation invariant analysis than the Cartesian form.

The output of the method is: the average value (an integer number in the interval [0,255]); standard deviation (an integer number in the interval [0,255]); energy (30 integer numbers in the interval [0,255]); energy deviation (30 integer numbers in the interval [0,255]).

The matching is done by summing the normalized weighted absolute difference between two sets of feature vectors not using rotation or scale invariant algorithms.

2.2.2. Edge histogram descriptor

This descriptor is a texture descriptor and describes the spatial distribution of four directional edges and one nondirectional edge in three different levels of localization in an image [Park, 2000]. The localization levels are the global, the semi-global and the local level. During the extraction, the image is partitioned into 16 non-overlapping sub-images with sizes depending on the original image size. It is also divided into a preferred number of image-blocks. For each image-block, a horizontal, a vertical, a 45 degree diagonal, a 135 degree diagonal and a nondirected edge value is calculated using edge extraction filters applied on the average brightness values in four sub-blocks. If the maximum edge value is greater than a threshold value, the image-block is considered to contain the corresponding edge. Otherwise, the image-block is considered to contain no edge. The image-block edge composition in the sub-images forms a local edge histogram with a total of 80 bins (5 types of edges, for each of the16 sub-

images). The global edge histogram summarizes the distribution of the different edges in the whole image by adding the corresponding local edge histogram bins into five global histogram bins one for each type of edge. The semi-global edge histogram is generated by accumulating the edge compositions in the sub-image clusters.

The output is a vector of 80 integer numbers between [0, 7].

Distance is calculated as added weighted difference between the local, global, and semi-global edge histograms respectively. Significance is measure by is the sum of absolute difference of 150 coefficients extracted from the 80 bins.

2.3. Shape descriptors

MPEG-7 supports region and contour shape descriptors. Object shape features are very powerful when used in similarity retrieval.

2.3.1. Region Shape

In the region shape descriptor, the shape of an object can be a single or multiple regions with or without holes [Kim, 1999]. The feature extraction is based on a set of Angular Radial Transform (ART) coefficients. ART is a complex 2-D transform defined on a unit disc with polar coordinates. In practice, the needed values of the basic functions are pre-calculated and put into a lookup table during the first step of the extraction. The ART transformation is then done by summing up the multiplication for each image pixel with each corresponding pixel in the lookup table, calculating the magnitudes.

The output is a vector of 35 integer numbers in the interval [0, 15].

The matching is done by calculating the minimum distance between the feature vectors for any shapes of two images. The distance for two vectors is the sum of absolute difference of coefficients.

2.3.2. Contour Shape

The contour shape descriptor presents a closed 2-D object or region contour in an image or video sequence [Mokhtarian, 1992]. During the extraction, N equidistant points are selected on the contour, starting from an arbitrary point on the contour and following the contour clockwise. The contour is then smoothed by repetitive low-pass filtering of the x and y coordinates of the selected contour points. The smoothing flattens out the concave parts of the contour. Points separating concave and convex parts

of the contour and peaks in between are then identified and the normalized values are saved in the descriptor.

2.4. An example of MPEG- 7 descriptors representation

An example of using some of the MPEG-7 descriptors in XML form on the sample image taken from TREC2002-FeatureDevelopment-mpeg1VideoSet [Smeaton, 2002], shown in Figure 1. is given after the figure.

Figure 1. Sample image (AtThisMo1954_2_KeyFrame_0_495.jpg)

```
<VideoSegment>
<MediaTime>
    <MediaTimePoint>T00:00:08:7247F30000</MediaTimePoint>
</MediaTime>

<VisualDescriptor xsi:type="ScalableColorType" numOfBitplanesDiscarded="0" numOfCoeff="64">
    <Coeff> -5 -22 -127 47 2 -6 16 29 -11 19 16 32 -42 -12 8 12 -6 7 -3 2 -15 5 0 0 -15 8 -5 0 -8 5 1 -4 3 -2 1 6 3 0 1 2 -1 -7 1 3 1 2 4 5 -5 0 4 0 -1 5 8 5 -5 0 0 -2 1 0 -3 -3</Coeff>
</VisualDescriptor>

<VisualDescriptor xsi:type="DominantColorType">
    <SpatialCoherency>20</SpatialCoherency>
    <Value>
        <Percentage>7</Percentage>
        <Index>50 39 20</Index>
        <ColorVariance>0 0 0</ColorVariance>
    </Value>
    <Value>
        <Percentage>2</Percentage>
        <Index>152 164 162</Index>
```

```
            <ColorVariance>0 0 0</ColorVariance>
        </Value>
        <Value>
            <Percentage>2</Percentage>
            <Index>143 131 115</Index>
            <ColorVariance>1 0 0</ColorVariance>
        </Value>
        <Value>
            <Percentage>3</Percentage>
            <Index>102 69 50</Index>
            <ColorVariance>1 0 0</ColorVariance>
        </Value>
        <Value>
            <Percentage>4</Percentage>
            <Index>82 92 90</Index>
            <ColorVariance>0 0 0</ColorVariance>
        </Value>
        <Value>
            <Percentage>7</Percentage>
            <Index>177 146 117</Index>
            <ColorVariance>0 0 0</ColorVariance>
        </Value>
        <Value>
            <Percentage>4</Percentage>
            <Index>227 223 217</Index>
            <ColorVariance>0 0 0</ColorVariance>
        </Value>
    </VisualDescriptor>

    <VisualDescriptor xsi:type="ColorLayoutType">
        <YDCCoeff>26</YDCCoeff>
        <CbDCCoeff>23</CbDCCoeff>
        <CrDCCoeff>37</CrDCCoeff>
        <YACCoeff5> 23 23 9 14 13</YACCoeff5>
        <CbACCoeff2> 14 12</CbACCoeff2>
        <CrACCoeff2> 19 20</CrACCoeff2>
    </VisualDescriptor>

    <VisualDescriptor xsi:type="ColorStructureType" colorQuant="4">
        <Values> 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 16 39 34 9 15 46 66 14 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 46 53 40
35 119 99 118 65 55 13 13 9 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 7 1 2 17 25 3 0 2 2 1 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 43 58 48 49 49 53 53 54 4 40 58 47 36 37 33 21 0 19
49 49 37 30 17 8 0 1 0 0 1 0 1 11 0 0 1 1 10 19 22 22 24 23 24 20 16 17 14 14 14 11 13 13
15 15 17 20 18 23 29 33 38 38 39 27</Values>
    </VisualDescriptor>
```

```
    <VisualDescriptor xsi:type="EdgeHistogramType">
        <BinCounts> 5 3 2 5 2 5 5 3 3 4 3 6 2 3 2 1 5 5 4 2 3 3 4 3 3 2 4 6 5 6 1 3 4 5 4 1 4
3 6 3 3 3 5 6 6 3 2 7 4 5 4 1 6 4 6 1 3 2 7 5 2 2 5 5 7 1 6 2 2 4 2 1 7 5 3 0 6 2 5
5</BinCounts>
    </VisualDescriptor>

    </VideoSegment>
```


3. The use of MPEG-7 descriptors

There are several problems, which have to be solved before evaluating the quality of different descriptors. The first problem is: how to choose the benchmark database? There is no common database used for content based benchmarking. Many researchers use the Corel image database (http://www.corel.com/). Another possibility is the collection used in MPEG-7 [MPEG98], but it is also copyrighted as Corel database. Other possibilities are the databases on: http://elib.cs.berkeley.edu/photos/tarlist.txt, http://www.cs.washington.edu/research/imagedatabase /groundtruth, or on http://www.white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html.
The second problem is how to measure the performance of the different descriptors. This mean to find a set of features which adequately encodes the characteristics that we intend to measure and a suitable metric. Which is the best similarity function? In 1977 Amos Tversky proposed his famous feature contrast model [Tversky, 1977]. He uses a set of binary features. In [Eidenberger, 2003] mean and standard deviation, distribution analysis and cluster analysis are used. Some of the results are: Color Layout performs badly on monochrome data. Like Color Layout, Color Structure performs inferior on monochrome data. The Dominant Color identifier performs equally well on any type of media. Scalable Color performs exactly like Color Layout and Color Structure. All color descriptors works excellent on photos but three of four perform badly on artificial media objects with few color gradations and very badly on monochrome content. An exception is the Dominant Color descriptor. This descriptor works well on each type of content. Edge Histogram performs excellent on any type of media. The Homogeneous Texture descriptor works acceptably on the Brodatz dataset. A combination of different descriptors is needed. The best descriptors for using combinations are Color Layout, Dominant Color, Edge Histogram and Texture Browsing. The others are highly dependent on these. The color histograms (Color Structure and Scalable Color) perform badly on monochrome input. Therefore, Dominant Color should be used for GoF/GoP color instead of Scalable Color. Generally, all descriptors are highly

redundant and applying complexity reduction transformations could save up to 80% of storage and transmission capacity.

In [Stanchev, 2004] we generalize this result. We propose a technique for evaluating the effectiveness of MPEG-7 image features on specific image data sets, based on well defined statistical characteristics of the data set. The aim is to improve the effectiveness of the image retrieval process based on the computed similarity on these features. We also validate this method with extensive experiments with real users.

Finally, some aspects of images are captured by none of the descriptors and existing descriptors should be either refined or new visual descriptors should be added to the standard.

4. Conclusion

Several visual descriptors exist for representing the physical content of images, for instance color histograms, textures, shapes, regions, etc. Depending on the specific characteristics of a data set, some features can be more effective than others when performing similarity search. For instance, descriptors based on color representation might be effective with a data set containing mainly black and white images. Techniques, based on statistical analysis of the data set and queries are useful.

It seems that the most intelligent descriptors are the one based on color layout. Not only does it compare the colors, but also where in the image they occur. This can be of great use if you are looking for a sunset, a face, a certain kind of landscape view etc, where similar colors usually occur in the same regions of the images. The texture and shape based search methods can also be very good, but the search results that are not among the used ground truth set can often be perceived as looking completely different compared to the query image so the use in general image databases can be questioned. On the other hand, the texture and shape based methods can recognize features such as contours and appearance that cannot be detected by the color based methods.

Even if it not possible, in general, to overcome the semantic gap in image retrieval by feature similarity, it is still possible to increase the retrieval effectiveness by a proper choice of the image features, among those in the MPEG-7 standard, depending on the characteristics of the various image data sets (obviously, the more homogeneous the data set is, better results can be obtained).

Bibliography

[Cieplinski, 2000] Leszek Cieplinski, Results of Core Experiment CT4: Extension of Dominant Colour Descriptor, MPEG-7 TR #13-06, January 2000

[Eidenberger, 2003] H. Eidenberger, "How good are the visual MPEG-7 features?", SPIE & IEEE Visual Communications and Image Processing Conference, Lugano, Switzerland, 2003

[Ferman, 2000] A. Ferman et al., "Group-of-frame/picture color histogram descriptors for multimedia applications", Proceedings of the Storage and Retrieval of the IEEE International Conference on Image Processing", Vol. 1, Vancouver, Canada, 2000, 65-68

[Grosky, 2001] Grosky W., Stanchev P., "Object-Oriented Image Database Model", 16th International Conference on Computers and Their Applications (CATA-2001), March 28-30, 2001, Seattle, Washington (94-97).

[Kasutani, 2001] E. Kasutani, A. Yamada, "The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video retrieval", Proceeding of International Conference on Image Processing 2001, Oct. 2001, Thessaloniki, Greece 2001

[Kim, 1999] W. Kim, Y. Kim, "A new Region-Based Shape Descriptor", MPEG-7 TR#15-01, December 1999

[Manjunath, 2001] Manjunath B., Ohm J., Vasudevan V., Yamada A., Color and Texture Descriptors, IEEE Transactions on circuits and systems for video technology, V. 11, No. 6, June 2001, 703-715

[Manjunath, 2002] B.S. Manjunath, P. Salembier, T. Sikora, "Introduction to MPEG-7", Wiley, 2002

[Messing, 2001] Dean S. Messing, Peter van Beek, James H. Errico, Using Colour and Local Spatial Information to Describe Images, MPEG-7 TR #13-07, January 2001

[Mokhtarian, 1992] F. Mokhtarian, A. Mackworth, "A theory of multiscale, curvature-based shape representation for planer curves", IEEEE Transaction on Pattern analysis and machine intelligence, 14 (8), 1992, 789-805

[MPEG, 2002] "MPEG-7 Overview (version 9)", ISO/IEC JTC1/SC29/WG11N5525

[MPEG98] MPEG Requirements Group, "MPEG-7: Context and objectives (version 10 Atlantic City)," Doc. ISO/IECJTC1/SC29/WG11, International Organisation for Standardisation, 1998.

[Park, 2000] D. Park, Y. Jeon, C. Won, S. Park, "Efficient use of local edge histogram descriptor", Processing of ACM International workshop on Standards, Interoperability and Practices, Marina del Rey, CA, USA, 2000, 52-54

[Rabitti, 1989] Rabitti F., Stanchev P., "GRIM_DBMS - a GRaphical IMage DataBase System", in "*Visual Database Systems*", T. Kunii (edt.) North-Holland 1989 (415-430).

[Ro, 2001] Y. Ro, M. Kim, H. Kang, B. Maniunath, J. Kim, “MPEG-7 Homogeneous Texture descriptor”, ETRI Journal 23 (2), 2001, 41-51.

[Schettini, 2001] Schettini R., Ciocca G., Zuffi S., A survey of methods for color image indexing and retrieval in image databases, in Luo R., MacDonal L., (editors) Corol Imaging Science: Exploiting Digital Media, J. Willey, 2001

[Smeaton, 2002] A. Smeaton, P. Over, The TREC-2002 Video Track report, http://www-nlpir.nist.gov/projects/t2002v/results/notebook.papers/ VIDEO.OVERVIEW.pdf

[Stanchev, 1999] Stanchev P., “General Image Database Model”, in Visual Information and Information systems, Huijsmans, D. Smeulders A., (etd.) Lecture Notes in Computer Science 1614, 1999 (29-36).

[Stanchev, 2004] P.Stanchev, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, Fausto Rabitti, Pasquale Savino, “Selection of MPEG-7 Image Features for Improving Image Similarity Search on Specific Data Sets”, The sevent IASTED International Conference “Computer graphics and imaging”, Kauai, Hawaii, 2004

[Tversky, 1977] A. Tversky, "Features of Similarity", Philosophical review, 84/4, 327-352, 1977

MPEG-7 BASED IMAGE RETRIEVAL ON THE WORLD WIDE WEB

Rajeev Agrawal[1], Farshad Fotouhi[2], Peter Stanchev[1], and Ming Dong[2]

1 - Kettering University, Flint, MI 48504, USA; ragrawal@kettering.edu; pstanchev@kettering.edu

2 - Wayne State University, Detroit, MI 48202, USA; fotouhi@cs.wayne.edu; mdong@cs.wayne.edu

Abstract: Due to the rapid growth of the number of digital media elements like image, video, audio, graphics on Internet, there is an increasing demand for effective search and retrieval techniques. Recently, many search engines have made image search as an option like Google, AlltheWeb, AltaVista, Freenet. In addition to this, Ditto, Picsearch, can search only the images on Internet. There are also other domain specific search engines available for graphics and clip art, audio, video, educational images, artwork, stock photos, science and nature [www.faganfinder.com/img]. These entire search engines are directory based. They crawls the entire Internet and index all the images in certain categories. They do not display the images in any particular order with respect to the time and context. With the availability of MPEG-7, a standard for describing multimedia content, it is now possible to store the images with its metadata in a structured format. This helps in searching and retrieving the images. The MPEG-7 standard uses XML to describe the content of multimedia information objects. Increasingly, these objects will have metadata information in the form of MPEG-7 or any other similar format associated with them. It can be used in different ways to search the object. In this paper we propose a system, which can do content based image retrieval on the World Wide Web. It displays the result in user-defined order.

Keywords: XML, MPEG-7, Metadata, Multimedia, Content Based Image Retrieval (CBIR)

1. Introduction

The CBIR has been a very active research area in the last decade. Conventional content-based image retrieval systems [1, 2, 3] use low-level features such as color, texture, shape, automatically extracted from the images. Another focus of this research is on improving the low level

features. The modifying the similarity measures make the retrieval as better as possible. It is argued in [4] that unconstrained object recognition is still beyond of current technology. The content based systems can at best capture only pre-attentive similarity, not semantic similarity. So far there has not been a single system, which can perform this task automatically without human intervention due to the nature of this problem.

The expansion of the World Wide Web (WWW) is making the problem of effective retrieval of images very important for all its users. The complexity of Web documents is rapidly increasing with the wide use of multimedia components, such as images, audio and video, associated to the traditional textual content. This requires extended capabilities of the Web query search engines in order to access images according to their multimedia content. A large number of search engines (e.g. Altavista, Yahoo, HotBot, etc.) support indexing and content-based retrieval of Web documents. Only the textual information is taken into account. Initial experimental systems providing support to the retrieval of Web documents based on their multimedia content (Webseek [5], and Amore [6]) are limited to the use of pure physical features extracted from multimedia data, such as color, shape, texture. These systems do not go beyond the use of pure physical visual properties of the images. They suffer the same severe limitations of today as the general-purpose image retrieval systems [7], such as Virage [8] and QBIC [9]. These systems consider images as independent objects, without any semantic organization in the database or any semantic inter-relationships between database objects. Many image searches also use an approach that filters out less relevant results. They analyze and index the text on the page adjacent to the image, the image link text, text in the HTML alt tag, filename or file path name. Similarly, this approach can also be used with other media files such as audio and video. Even though these search engines do not “look inside” the media files, they can give quite relevant results.

Another approach can be to look into the media file contents itself and trying to mine for textual information in the file for better multimedia indexing. For example, a Portable Network Graphics (PNG) image file can contain textual information such as title, author, description, copyright, creation time, software used, disclaimer, warning, source and comment [10]. Not all file formats contain metadata, and even if they do, an indexing engine should know how to handle all the different file formats and where to find that information in a file. It would be better if we had a data model, which could be used with different media formats and utilized a rich set of

metadata. There have been many metadata models developed. Some of them are RLG Preservation Metadata Elements, NISO Draft Standard, DIG35 Specification, Data Dictionary for Audio/Video Metadata, Metadata for Long-Term Preservation, Metadata Encoding and Transmission Standard [11]. MPEG-7 is another multimedia metadata standard. The Moving Picture Experts Group (MPEG) was established in 1988 to develop audiovisual compression standards. MPEG-1, MPEG-2, and MPEG-4 all represent the content itself, while MPEG-7 represents information about the content [12]. While the first produces the contents, the latter describes the content. There are number of tools provided in MPEG-7 - descriptors (the elements), description schemes (the structures), a Description Definition Language (DDL) (for extending the predefined set of tools) and a number of system tools. MPEG-7 can support all natural languages. DDL provides the foundation for the standard. It provides the language for defining the structure and content of MPEG-7 documents. The DDL is not a modelling language such as Unified Modelling Language (UML) but a schema language to represent the results of modelling audiovisual data (i.e. descriptors and description schemes) as a set of syntactic, structural and value constraints to which valid MPEG-7 descriptors, description schemes, and descriptions must conform. The purpose of a schema is to define a class of XML documents. The purpose of and MPEG-7 schema is to define a class of MPEG-7 documents. MPEG-7 instances are XML documents that conform to a particular MPEG-7 schema (expressed in the DDL) and that describe audiovisual content. MPEG7 has been developed after many rounds of careful discussion. It is expected that this standard would be used in searching and retrieving for all types of media objects. If we have images stored with MPEG-7 metadata, it would be easier to do semantic retrieval. MPEG-7 files contain a reference to the location of the corresponding image file. It is also possible to exploit other tools and technologies developed for XML like Xquery, XPath, etc. There has been a lot of work on XML schema integration. This plays a central role in numerous applications, such as web-oriented data integration, electronic commerce, schema evolution and migration, application evolution, data warehousing etc. In schema integration, the main objective is to find a suitable technique to match the elements in different schemas. We propose to combine XML schema integration techniques and image retrieval techniques using low-level features with or without semantic annotations.

Rest of the paper is organized as follows: Section 2 describes the

motivating examples. Section 3 relates a list of previous work and other literature survey. Section 4 describes our proposed system. Finally, we give concluding remarks in section 5.

2. Motivating Examples

The commercial image search engines available today basically search the images based on keywords. The keywords are extracted from the web page, where image appears. But the keyword-based search has its own limitation, which will be clear from the following examples.

1. If we want to search and retrieve the pictures of a person in the different stages of his/her life with respect to the time, available on different websites, that is not possible through keyword search. The keyword search would definitely retrieve the images but not integrate in the order we want. Assumption here is that different websites has the pictures of the person at different stages of his/her life and also incorporate some semantic information, which can be in MPEG-7 or in any other metadata format. The reason is keyword search just looks for the name in the surrounding text, but no in other information. E.g.: When we search the pictures of a great person like Mahatma Gandhi. Images are retrieved, but not in any order. The main reason is that no semantic information is incorporated with the images.

2. Some security agency is interested in getting more information about a person, who has perpetrated some crime and they have a photograph of this person. There is no technique available, which can return the information about this person from the Internet, if the agency uses this photograph as input (query by example method). The basic idea of this kind of search is that low level feature of the query image should be compared with all the images available on the Internet and a set of images, which are closer up to certain threshold are returned.

3. We want to search images of two cities, which belong to same country. The keyword search can include some false results. E.g. when we search for the cities Detroit and Flint together, we see some graphs, which are not the images of the cities, but refer their names in the graphics.

4. There is also no method available, which can return the result of following types of query. E.g.: Search the pictures about American history between the year 1900 and 1950.

There is no method of defining the queries between certain time range

and/or any other metric. One of the problems of not getting the desired results is that there in no or little metadata available with the images available on the Internet. Second reason is that the algorithms employed by the search engines, does not have the capability to do search based on a specific criteria like these. As we can see in the above examples, that there is still a long way to be able to apply complex queries to search the images from the World Wide Web. In addition to above examples, we may encounter large number of other kinds of queries, which are not possible through existing search engines.

3. Literature Survey

The CBIR on World Wide Web involves two research areas: images classification and, images search and retrieval techniques.

3.1. Image Classification

In the literature, a wide variety of content-based retrieval methods and systems may be found. In [13], authors have reviewed about 200 references in CBIR up to the year 2000. There are three broad classes of applications user aims when using the system: search by association, search at a specific image, and category search. [14] identifies other patterns of use: searches for one specific image, general browsing to make an interactive choice, searches for a picture to go with a broad story, searches to illustrate a document etc. An attempt to formulate a general categorization of user requests for still and moving images are found in [15]. This and similar studies reveal that the range of queries is wider than just retrieving images based on the presence or absence of objects of simple visual characteristics. To describe the image, we have to extract certain low-level features from it. There are a number of image processing operations that translate the image data into some other spatial data array. These operations may use local color, local texture, or local geometry. The main purpose of image processing in image retrieval must be to enhance aspects in the image data relevant to the query and to reduce the remaining aspects. There are several color representations like RGB, HSV, YUV and their variations.

Local shape characteristics derived from directional color derivatives have been used in [16] to derive perceptually conspicuous details in highly textured patches. In [17], a series of Gabor filters of different directions and scale have been used to enhance image properties [18]. Combining shape

and color both in invariant fashion is a powerful combination as described by [19]. The texture is defined as all what is left after color and local shape have been considered or it is defined by such terms as structure and randomness. Basic texture properties include the Markovian analysis and other generalized versions [20, 21].Other texture analysis methods are MRSAR-models [22], Wavelets [23], fractals [24] etc. A comparative study on texture classification from mostly transform-based properties can be found in [25].

In CBIR, the image is often divided in parts before features are computed from each part. There are four types of partitioning identified in [13]: string segmentation, weak segmentation, sign detection, data independent image partitioning. In [26] knowledge-based type abstraction hierarchies are used to access image data based on context and a user profile, generated automatically from cluster analysis of the database. Also in [27], the aim is to create a very large concept-space inspired by the thesaurus-based search from the information retrieval community. In [28], a variety of techniques is discussed treating retrieval as a classification problem. One approach is principal component analysis over a stack of images taken from the same class of objects. This can be done in feature space [29] or at the level of the entire image [30]. In [31], binary Bayesian classifiers are used to capture high-level concepts from low-level image features under the constraint that the test image belongs to one of the classes. Specifically, the hierarchical classification of vacation images is considered; at the highest level, images are classified as indoor or outdoor; outdoor images are further classified as city or landscape; finally, a subset of landscape images is classified into sunset, forest, and mountain classes. A large number of systems have ignored two distinct characteristics of CBIR systems: the gap between high-level concepts and low-level features, subjectivity of human perception of visual content. A relevance feedback based approach has been suggested in [32]. Other interactive approaches have been suggested in [33, 34, 35]. Example include interactive region segmentation [36]; interactive database annotation [34, 37]; usage of supervised learning before the retrieval [38, 39]; and interactive integration of keywords and high-level concepts to enhance image retrieval performance [40, 41]. In [42], an image retrieval system called SIMPLIcity (Semantics-sensitive Integrated Matching for Picture Libraries), which uses semantics classification methods, a wavelet-based approach for feature extraction, an integrated region matching based upon image segmentation,

has been proposed. There are several domain-dependent ontology based systems [43, 44]. In [45], system uses a neural network to identify objects present in the images.

3.2. Image Search and Retrieval Techniques

There are a large number of papers published in the area of image search and retrieval. We are restricting our discussion here related to image search on World Wide Web. A system is implemented in [46], by which visual information on the web is (1) collected by agents, (2) processed in both text and visual feature domains, (3) catalogued and (4) indexed for fast search and retrieval. A typical web image search engine will first traverse the Web by following the hyperlinks between documents using several autonomous Web agents or spiders. These agents detect images and download and process them and add the new information about the image to the catalog.

A *perception-based search component,* which can learn users' subjective query concepts quickly through an intelligent sampling process, is proposed in [47]. A multi-resolution feature extractor extracts perceptual features from images and a high-dimensional indexer performs non-supervised clustering using Tree-structured Vector Quantization (TSVQ) [48] to group similar objects together. iFind is a web-based image retrieval system developed at Microsoft Research, China [49]. It provides the functionalities of text based image search, query by example, and their combination. Images in the database are indexed by their low-level (visual) features, high level (semantic) features (collected from image's environment), and optionally, annotations if they are available. In [50], MISE (The MediaSys Image Search Engine) is described. This system enables the users to search, to browse, to process, and to store images according to the combination of visual and textual features with meta-data related to the images. The MediaSys servers store the meta-data, visual and textual features, and the images themselves over a large scale distributed and heterogeneous system. The article [51] investigates what MPEG-7 means to Multimedia Database systems (MMDBSs) and vice-versa. It is argued that MPEG-7 has to be considered complementary to, rather than competing with, data models employed in it. [52] describes the use of stylesheets in the search and retrieval process of multimedia information, especially for audiovisual information. MPEG-7 has been used to describe the contents of the information. The use of stylesheets over the MPEG-7 data gives flexibility during both query formulation and the presentation of search

results, and it allows a personalized way of querying and presenting.

4. The proposed system

We discussed some of the example queries in section 2, which can not be answered by any of the existing systems to the best of our knowledge. We propose a system, which will exploit the XML technology and new MPEG-7 media metadata standard. In this section, we briefly describe the Image Integration Architecture:

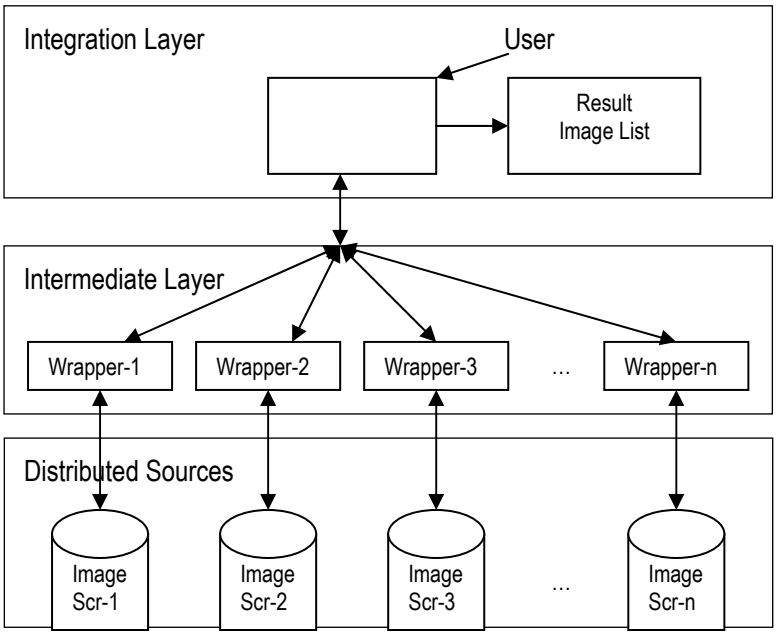


Figure 1: Image Integration architecture

Figure 1 shows three-layered image integration architecture. At the lowest level, we have different Image sources. These sources have images and have not been designed on certain agreed schema. In other words, images in these sources may be in raw JPG, BMP, GIF or any other format without any semantic information. They may contain images clustered in certain groups. They may contain metadata in the form of MPEG-7 with

partial annotation or they may contain MPEG-7 metadata with structured annotation. There may be other possibilities also.

Intermediate layer focuses on extracting image information by extracting low-level features, metadata or any other semantic information available. If there is no semantic information available, we have to rely solely on low-level features. We are considering images, which are embedded in a webpage or stored in the image database. Each image source has to be treated in a different way:

Image source with raw image formats. At intermediate level, we extract low-level features and store as MPEG -7 metadata. Since this procedure has to be automatic, we can not do annotations at this level. There is no automatic annotation technique available so far.

Image source with raw image formats but clustered in groups. We extract low-level features from the image and also store cluster information in MPEG - 7 metadata. Some intelligent technique has to be used to make cluster information useful in retrieval. We can also use traditional image search retrieval methods and look for important keywords stored in and around the image.

Image source with raw image format and with some metadata but not in MPEG – 7. We extract low-level features of the image and use the metadata while creating MPEG-7 metadata.

Image source with MPEG-7 partial or full annotation. We do not need to extract the low-level features, since they are already available in MPEG-7 metadata.

Our emphasis here is to get information about all the images in MPEG-7 format, which is essentially XML data. Then we can use XML tools to query the images in Integration layer. There has been a lot of work on XML schema Integration [53, 54, 55, 56, 57, 58]. Due to the space constraint, we are not discussing about XML schema integration here.

The user will make a query at the top level using any of the methods using keyword, query by example, range queries etc discussed in section 2. This architecture may use agent based method or the popular directory based indexing method to search the image data sources. Integration process consists of querying the results returned by the intermediate layer, refine them according to user demand and return the results back to him/her. The relevance feedback and/or other long term learning technique can be used at the highest level to improve the results. The queries similar to the examples mentioned in section 2 can be successful if we combine low level features and semantic information together to produce the results. This architecture does not merely return the search results based on the keywords associated with the image, but also takes into account the low level features of the image.

5. Conclusion and Future Research

In this paper we suggest three - layered image integration architecture at a conceptual level. This approach takes care of images stored on the websites/image databases with or without semantic information. There are many challenges we have to face in this approach like selecting appropriate schema integration technique. MPEG-7, though already declared standard, will still take some time before images have their metadata stored in this format. Therefore, it would be a grave mistake to rely on the assumption that metadata would be easily available in MPEG-7. Similarly, there are a large number of low-level features suggested by different researchers, but MPEG-7 has included only some of them. There are possibilities that better features may be released in future and we have to consider these new features in any content-based image retrieval system. We are trying to set up an experimental environment based on the approach suggested in this paper. We are in the process of collecting the images with the properties described in section 4. We believe that the proposed system would greatly enhance the quality of content-based image retrieval.

Bibliography

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by Image and video content: The QBIC system. IEEE Computer, vol. 28, no. 9, pp. 23-32, 1995.

[2] W. Y. Ma and B. S. Manjumath. Netra: A toolbox for navigating large image databases. ACM Multimedia System, vol. 7, pp. 184-198, 1999.

[3] Y. Rui, T. S. Huang. S. Mehrotra, and M. Ortega. A relevance feedback architecture for content-based multimedia information retrieval systems. IEEE Workshop on Content-based Access of Image and Video Libraries, pp. 82-89, 1997.

[4] S. Santini and R.Jain. Visual navigation in perceptual databases. International Conference on visual Information systems, San Diago, CA, Dec. 1997.

[5] J. R. Smith and S. Chang, Visually searching the Web for content, IEEE Multimedia, July-September 1997.

[6] S. Mukherjea, K. Hirata and Y. Hara. Towards a multimedia World Wide Web information retrieval engine. 6th WWW International Conference, S. Clara, CA, 6–11 May 1997.

[7] C. Meghini, f. Sebastiani and U. Straccia. Modelling the retrieval of structured documents contaning texts and images. 1st ECDL, Pis, Italy, Sep. 1997

[8] J.R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain and C.F. Shu. The Virage image search engine: An open framework for image management. SPIE 96, 1996.

[9] M. Flickner et al., Query by image and video content: the QBIC system, IEEE Computer, 28(9), September 1995.

[10] PNG (portable Network Graphics) Specification, Version 1.2., http://www.libpng.org/pub/png/spec/

[11] Metadata Standards. http://www.chin.gc.ca/English/Standards/metadata_multimedia.html

[12] B S Manjunath et. El. Introduction to MPEG-7. John Wiley, 2002.

[13] Arnold W. M. Smeulders et. el. Content-Based Image Retrieval at the End of Early Years. IEEE transactions on Pattern Analysis and Machine Intelligence, vol. 22, No. 12, Dec 2000.

[14] S. Ornager. Image Retrieval: Theoretical and Empirical User Studies on Accessing Information in Images. 60th Am. Soc. Information Science Ann. Meeting, vol. 34, pp. 202-211, 1997.

[15] L. Armitage and P. Enser. Analysis of User Need in Image Archives. J. Information Science, vol. 23, no. 4, pp. 287-299, 1997.

[16] A. Mojsilovic, J. Kovacevic, J. Hu, R.J. Safranek, and S.K. Ganapathy. Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns. IEEE Trans. Image Processing, vol. 9, no. 1, pp. 38-54, 2000.

[17] B.S. Manjunath and W.Y. Ma. Texture Features for Browsing and Retrieval of Image Data. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 837-842, Aug. 1996.

[18] R. Rodriguez-Sanchez, J.A. Garcia, J. Fdez-Valdivia, and X.R. Fdez-Vidal. The RGFF Representational Model: A System for the Automatically Learned

Partitioning of `Visual Pattern' in Digital Images. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 10, pp. 1,044-1,073, Oct. 1999.

[19] T. Gevers and A.W.M. Smeulders. Content-Based Image Retrieval by Viewpoint-Invariant Image Indexing. Image and Vision Computing, vol. 17, no. 7, pp. 475-488, 1999.

[20] S. Krishnamachari and R. Chellappa. Multiresolution Gauss-Markov Random Field Models for Texture Segmentation. IEEE Trans. Image Processing, vol. 6, no. 2, 1997.

[21] G.L. Gimel'farb and A.K. Jain. On Retrieving Textured Images from an Image Database. Pattern Recognition, vol. 29, no. 9, pp. 1,461-1,483, 1996.

[22] J. Tatemura. Browsing Images Based on Social and Content Similarity. Proc. Int'l Conf. Multimedia and Expo, 2000.

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

[24] L.M. Kaplan et al. Fast Texture Database Retrieval Using Extended Fractal Features. Storage and Retrieval for Image and Video Databases, VI, vol. 3,312, pp. 162-173, SPIE Press, 1998.

[25] T. Randen and J.H. Husoy. Filtering for Texture Classification: A Comparative Study. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 4, pp. 291-310, Apr. 1999.

[26] C.C. Hsu, W.W. Chu, and R.K. Taira. A Knowledge-Based Approach for Retrieving Images by Content. IEEE Trans. Knowledge and Data Eng., vol. 8, no. 4, pp. 522-532, 1996.

[27] H. Chen, B. Schatz, T. Ng, J. Martinez, A. Kirchhoff, and C. Lim. A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 18, no. 8, pp. 771-782, Aug. 1996.

[28] N. Vasconcelos and A. Lippman. A Probabilistic Architecture for Content-Based Image Retrieval. Proc. Computer Vision and Pattern Recognition, pp. 216-221, 2000.

[29] H. Murase and S.K. Nayar. Visual Learning and Recognition of 3D Objects from Appearance. Int'l J. Computer Vision, vol. 14, no. 1, pp. 5-24, 1995.

[30] R.W. Picard and T.P. Minka. Vision Texture for Annotation. Multimedia Systems, vol. 3, pp. 3-14, 1995.

[31] Aditya Vailaya et. el. Image Classification for content-Based Indexing. IEEE reanscations on Image Processing, Vol. 10, No. 1, Jan. 2001.

[32] Yong rui et. el. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. IEEE transactions on circuits and Video Technology. Volume 8, Number 5, 1998, pp. 644-655.

[33] A. D. Narasimhalu, Multimedia Syst. (Special Section on Content-Based Retrieval), 1995.

[34] W. Niblack, R. Barber et al. The QBIC project: Querying images by content using color, texture and shape. In Proc. SPIE Storage and Retrieval for Image and Video Databases, Feb. 1994.

[35] P. P. Ohanian and R. C. Dubes. Performance evaluation for four classes of texture features. Pattern Recognition, vol. 25, no. 8, pp. 819–833, 1992.

[36] M. Ortega, Y. Rui, and K. Chakrabarti, S. Mehrotra, and T. S. Huang. Supporting similarity queries in MARS. In Proc. ACM Conf. Multi-media, 1997.

[37] A. Pentland and R. Picard. IEEE Trans. Pattern Anal. Machine Intell. (Special Issue on Digital Libraries), 1996.

[38] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. Int. J. Comput. Vision, 1996.

[39] R. W. Picard and T. P. Minka. Vision texture for annotation. Multi-media Syst. (Special Issue on Content-Based Retrieval).

[40] J. Dowe. Content-based retrieval in multimedia imaging. In Proc. SPIE Storage and Retrieval for Image and Video Databases, 1993.

[41] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in MARS. In Proc. IEEE Int. Conf. ImageProcessing, 1997.

[42] J. Z. Wang, G. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries. In IEEE Trans. on pattern Analysis and Machine Intelligence, volume 23, pages 947--963, 2001.

[43] C. Breen, L. Khan, and A. Kumar. Image Classification Using Neural Networks and Ontologies. IEEE DEXA, International Workshop on Web Semantics, France, Sept 2002.

[44] L. Khan and D. McLeod. Audio Structuring and Personalized Retrieval Using Ontologies. IEEE Advances in Digital Libraries, Library of Congress, Washington, DC, May 2000.

[45] Casey Breen, Latifur Khan, Arun Kumar, and Lei Wang. Ontology-based image classification using neural networks", SPIE 2002

[46] J. R. Smith, S.-F. Chang, “Searching for Images and Videos on the World-Wide Web,” Columbia University, No. CU/CTR/TR 459-96-25, Aug. 1996.

[47] Wei-Cheng Lai, Edward Chang, and Kwang-Ting (Tim) Cheng. An Anatomy of a Large-scale Image Search Engine. WWW 2002, 7-11 May 2002, Honolulu, Hawaii.

[48] A. Gersho and R. Gray. Vector Quantization and Signal Compression. Kluwer Academic, 1991

[49] Hong-Jiang Zhang, Zheng Chen, Wen-Yin Liu and Mingjing Li. Relevance Feedback in Content-Based Image Search. Invited Keynote, 12th Int. Conf. on New Information Technology (NIT), May 29-31, Beijing.

[50] Panrit Tosukhowong, Frederic Andres, Kinji Ono, Jose Martinez, Noureddine Mouaddib, Nicolas Dessaigne and Douglas C. Schmidt A Flexible Image Search Engine. ACM, MM99, Oct 30 - Nov 5, 1999, Orlando, Florida

[51] Harold Kosch. MPEG-7 and Multimedia Database Systems. SIGMOD Record, Vol. 31, No.2, June 2002.

[52] Mark van Setten, Erik Oltmas, Mettina Veenstra. Personalized Video Search and Retrieval using MPEG-7 and Stylesheets. https://doc.telin.nl/dscgi/ds.py/Get/File-8842/

[53] V. S. Subrahmanian, Sibel Adali, Anne Brink, Ross Emery, James J. Lu, Adil Rajput, Timothy J. Rogers, Robert Ross, and Charles Ward: HERMES: A heterogeneous Reasoning and Mediator System. Technical Report, University of Maryland, Maryland, 1995.

[54] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom: The TSIMMIS Project: Integration of Heterogeneous Information Sources. IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.

[55] Rajeev Agrawal, Mukesh Mohania, Yahiko Kambayashi, S S Bhowmick, S Madria: An Architecture for XML Schema Integration. ICDL: Research and Practices, Kyoto, Japan, 2000.

[56] Ralf Behrens: A Grammar Based Model for XML Schema Integration. BNCD, 2000

[57] Ronaldo dos Santos Mello, Carlos Alberto Heuer: A Bottom-Up Approach for Integration of XML Sources. WIIW 2001

[58] Dongwon Lee, Murali Mani, Wesley W. Chu: Effective Schema Conversions between XML and Relational Models. In European Conf. on Artificial Intelligence (ECAI), Knowledge Transformation Workshop (ECAI-OT), Lyon, France, July 2002.

ASPICO: ADVANCED SCIENTIFIC PORTAL FOR INTERNATIONAL COOPERATION ON DIGITAL CULTURAL CONTENT

Frédéric Andrès, Jérôme Godard[*]*, and Kinji Ono*

National Institute of Informatics, University of Advanced Studies
Hitotsubashi 2-1-2 Chiyoda-ku, Tokyo 101-8430, Japan
andres@nii.ac.jp; jerome@grad.nii.ac.jp; ono@nii.ac.jp

Abstract: In this paper, we present the development of an advanced open source multi-lingual cooperative portal system dedicated to archive and semantic management, and to cooperative exchange for research and education purpose about the historical silk roads.

Keywords: Digital Silk Roads, Semantic Understanding, Metadata Annotation

1. Introduction

Following the evolution of cultural heritage archives, new requirements for semantic understanding in a multi-lingual and multi-disciplinary cultural field such as the historical silk roads have been pointed out in major symposiums [Ono 2001, Ono 2003] related this field. The *Advanced Scientific Portal for International COoperation* (ASPICO) aims at providing a web portal service to enable international and multi-disciplinary researchers and fellows to cooperate on research about the historical Silk Roads. The platform is open source and available for every one for research and education. It makes it possible to provide multilingual semantic extraction service in order to process digital cultural artefacts from a cross disciplinary point of view, based on cooperative annotation support, metadata extraction and classification. It is based on powerful and industry-leading software such as Linux OS, Apache, MySQL and Java. It is an independent platform, allowing internal usage (e.g. intranet), external usage (e.g. extranet) and access for general public via the internet. It provides flexible features, allowing easy customization for individual needs. Also it provides in a

[*] This research is partially supported by a grant (*bourse Lavoisier*) from the French Ministry of Foreign Affairs (Ministère des *Affaires Etrangères*).

transparent way a fully multilingual support. So searches on the data can be performed on all the information in any language. The platform is standards compliant based on the usage of XML which allows complex data interaction and analysis to take place both within the server and on the client side. The platform has a distributed architecture so autonomous systems can be located in different institutions in order to be consulted simultaneously and to aggregate final results.

In Section 2, we introduce the platform architecture. Then the section 3 describes the Image Content Recognition aspects and it reviews the state of the art in the field of active contour. Finally, Section 4 concludes and gives the direction of the future work.

2. The Platform Architecture

The platform architecture (Fig 1) includes database servers, the ontologies, the query interface and resource entry service, and high resolution image viewing.

2.1. Ontology Management

A key feature of the system is the multilingual ontology support. It is organized as a set of multi-layer ontology. Each ontology is related to one field such as history, geography, architecture, art… This approach enables search by contextual content as it relies on annotated documents and features extraction processes. Each set of ontology is based on an object-identifier bridge and mono-lingual Unicode (UTF-8) encoding ontologies.

2.2. Query Interface

Queries are performed via a web browser based interface. Screens for simple or advanced queries can be easily created and the fields to be viewed customized by the system administrator. In addition, date or numeric size fields can be searched by specifying a range of dates or sizes between which searches are performed. Users are able to select the working language and the domain of interest as well as the number of results returned and whether resource results are shown.

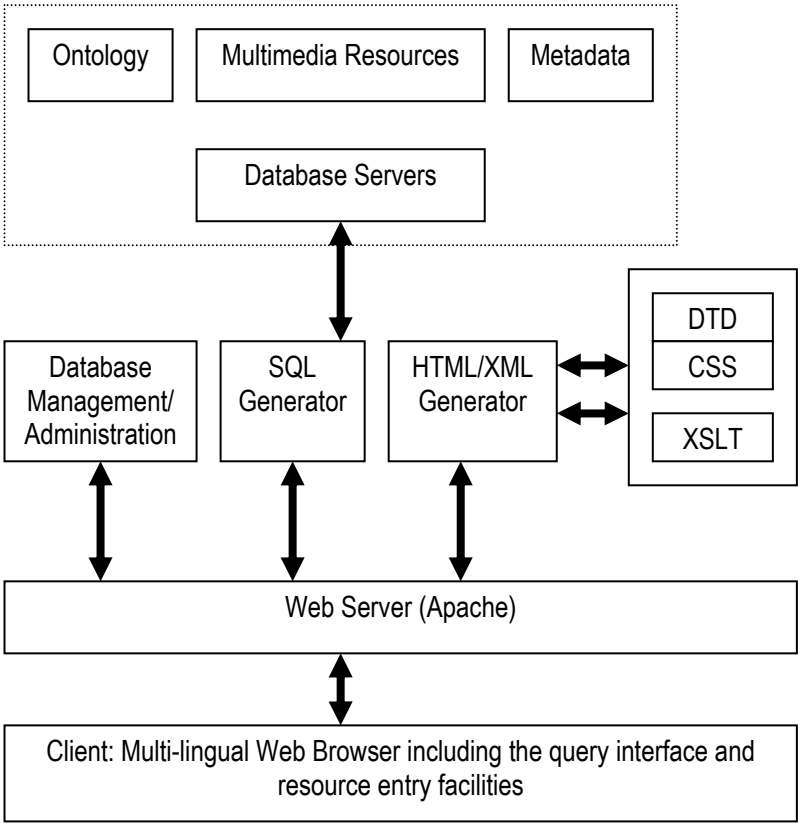


Fig 1: Architecture of the ASPICO platform

The interface is divided in three parts:

1- Historical and material resources related to artifacts;
2- Technical or management information related to photographic resources;
3- Technical or management information related to document resources;

Where applicable, the user can choose technical terms from a list of relevant terms classified alphabetically, or can type something directly in. Ontologies in 21 languages will be able to be consulted on line. Full text searches can be made within each field.

The display or the output format of the results (e.g. HTML, XML, plain

text, formatted tabular, list of images, graphical, statistical analyses etc) is independent of the storage structure in order to optimize the delivery process. It typically follows a methodology based on context-dependent cultural resource accesses [Godard 2003].

2.3. Resource Entry

Resource entry is also performed via a web browser interface similar to that used for querying. Users who enter Resource need to log on. Write, modification and suppression rights can be assigned and controlled by the system administrator for each user; some predefined types of user provide group management abilities. Information such as name of the user and date of the entry are automatically filled in by the system. To maintain the integrity of the resource being entered into the system, the controlled lists of relevant vocabulary within the thesaurus are used for each translatable field. When uploading resource via the web interface, users are required to enter some preliminary metadata related to the resource. When a resource is saved into the main database, the metadata is translated into a language independent code representation.

2.4. High Resolution Image Viewing

Another key component of the resource management system is the capability to remotely view high resolution images of both 2D paintings and 3D objects. Each image resource is stored as both a JPEG thumbnail for rapid previewing and in tiled pyramidal TIFF format for high-resolution viewing. A java applet permits multi-resolution viewing in conjunction with a tile server. This viewing system is based on the Internet Imaging Protocol. The viewer works by requesting only the tiles at the appropriate resolution required for viewing a particular part of the image. The requested tiles are then dynamically JPEG encoded by the server and sent to the applet. In this way, images of any size can be viewed quickly across the internet.

2.5. Web Site Mirroring

The resource archive will be mirrored to permit link-ups between distributed research databases. The system limits the access to data according to users rights to indoor users (INTRANET), outdoor users (EXTRANET) and to the Web users.

2.6. Multilingual Access

2.6.1. Multilingual Ontology Integration

The multilingual Ontology support based on protégé 2000 [protégé 2000] has been set up for the indexing of the scientific resource contents in the frame of the DSR project, cooperation between UNESCO and NII.

2.6.2. Ontology Terms Translation

The controlled lists of technical terms from each ontology as well as the free text information fields (such as the titles) have been translated with the support of domain experts. Unicode module has been integrated for the Asiatic language support.

3. Image Content Recognition

Content-based image retrieval is a challenging and active research area [Jain 1998] with the potential to provide powerful tools for image searching and semantic understanding. Although many techniques have been described in the research literature, the capabilities of current content matching systems are still basic general purpose approaches. According to the field of the image content, specific methods can be developed to specialize the image content recognition processing accordingly. General techniques based on such features as color distribution, texture, outline shape and spatial color distribution have been popular in the research literature and in content based retrieval systems.

The digitization of the photo resource archives at NII allows us to sample and group together images with similar characteristics thereby providing the reference material for testing such image content recognition software.

3.1. Semi-Automatic Content Recognition of Images

Regarding to this aspect, requirements have been determined according to the resource type as part of the resource metadata. If the resource type is document, the recognition is done for each image included inside the document. Iconography characterization is one the most complex issue in this field. Some shape recognition works well such as portrait, landscape, buildings, or themes such as crucifixion, or virgin and child. Cooperation with UNESCO and DSR experts has been established to provide a multi-lingual support on these categories for the DSR project.

3.2. Automatic Shape Identification

One user-requirement for the DSR project is the identification of shapes for painting and buildings in order to provide richer statistics for searches. This is useful for restricting areas of interest and avoiding backgrounds. This is carried out using recognition of deformable models. Research activities concerning deformable models can be partitioned in two types in [Jain 1998]:

- Free-form model, also called active contours, which allows representing any shape by using a minimizing energy algorithm.
- Parametric model allows defining and encoding specific geometric properties of the shape (moments, angles…)

Let us review the two classes.

Free-form model

An initial contour, or snake, C is defined by the coordinates $\{x(s), y(s)\}$, 0<s<1. The method was initially introduced by [Kass 1988] and involves the energy-minimization contour C by controlling the three forces:

- The internal forces E_{int}, which define the constraints concerning the shape of the model (more or less smooth).
- The images forces E_{image} which distort the contour according to the variations of pixels values (grey level or colour values).
- The external forces E_{con}.

The willed contour is thus obtained by minimizing the energy given by:

$$E_{snake} = \int_0^1 \{E_{\text{int}}(x(s), y(s)) + E_{image}(x(s), y(s)) + E_{con}(x(s), y(s))\} ds$$

The external forces E_{con} will be not used in what follows.

The internal forces are mainly defined by the coordinates of the snake C

$$E_{\text{int}}(s) = \tfrac{1}{2}\{\alpha(s)\left[x_s(s)^2 + y_s(s)^2\right] + \beta(s)\left[x_{ss}(s)^2 + y_{ss}(s)^2\right]\}$$

Where the subscripts on x and y define derivative form. The coefficients α and β indicate the strength of the elasticity and of the rigidity. In practice, for the digital images applications the problem must be discretized [Davison 2000]. Energies must be sampled at N equally spaced knots v_i around the edge C:

$$E_{\text{int}} = \frac{1}{2h}\sum_{i=0}^{N-1}\alpha_i\left|v_i - v_{i-1}\right|^2 + \frac{1}{2h^3}\sum_{i=0}^{N-1}\beta_i\left|v_{i-1} - 2v_i + v_{i+1}\right|^2$$

In general, the first curve is initialized by B-splines, widely described by [Blake 1998] and used, for instance, by [Stammberger 1999] for a magnetic resonance imaging application.

The image energy E_I depends on the variations of grey level $g(x_i,y_i)$.

$$E_I = k_I\sum_{i=1}^{N} g(x_i, y_i) \qquad E_{image} = -\nabla E_I$$

The energy-minimization is usually realized iteratively by a gradient descent algorithm until a minimum. Intuitively, a major drawback appears. Indeed, the contour C that depends on the initial position can be attracted by a local minimum, far-off the shape desired. A control of the final contour must be thus checked. Many approaches have been proposed to erase these problems in [Jain 1998], and in [Tsechpenakis 2004]. Moreover, the method fails sometimes for very complex shapes. Nevertheless, the method remains very powerful for image segmentation and its implementation is very fast.

Lastly, the use of colour information allows improving the performance of active contours. [Ngoi 1999] proposed thus a new active contour model for shape extraction of images acquired in outdoor conditions.

Parametric models

The active contours are based on an energy minimizing calculated from the coordinates of the pixels belonging to the contour; the basis of the parametric models is the study of the shape deformation by using geometrical parameters. The model needs now more specific a-priori knowledge of the shape. We can differentiate two parametric models [Jain 1998]:

- Analytical deformable models which are defined by analytical curves
- Prototype-based deformable templates defined an “average” shape of a class of objects.

- Analytical deformable templates

In those methods, templates are defined by parametric models such as ellipses, or circle parametric function. The model, which possesses only few degrees of freedom, fit the desired shape by energy minimizing applied to the model parameters. The most popular example of such a method is the eye template of [Yuille 1992]. In this model, the parameters for which the variations are carried out are the centre and the radius of the circle and the coefficients of the parabola. In this model, we distinguish also two kinds of energy, the internal energy, which is defined by a parametric function characterizing a shape, and the external energy, which represents the features of the image. Minimizing energy algorithm is then used.

Because of the parametric function is chosen previously, the analytical deformable templates are required for segment objects with a known shape.

- Prototype-based deformable templates

For the prototype-based deformable templates, a particular model is previously built according to the shape we want to extract (for example a model of a sculpture or a model of a building). The performance of the prototype template depends, obviously, on the description of the shape. Recent research works have adopted learning method from a set of samples. So as to do it, [Cootes 1994] and [Cootes 1995] have thus used this kind of method. From those samples templates, a mean shape is calculated and used as the generic model and the variations are determined by eigenvectors of the covariance matrix. Other deformable templates based on a prototype have been also described in [Jain 1998].

Digital Silk Roads archives contain high-resolution colour images acquired in outdoor (high luminosity, reflections, shadows). If the natural conditions of imaging and the variability of the conditions complicate the segmentation, the images to be segmented present objects with particularly simple shapes (doors, roofs, mosaics, arch…), making the use of deformable models easier. The main advantages of active contours are the speed and the flexibility. Specific a-priori knowledge is not required. On the other hand, it seems that the deformable templates are very adapted for locating specific structures in the images but this method needs a more specific knowledge about the shape we want to extract. Another difficulty is to define accurately objects. The segmentation tool must be precise for segment small objects such as frieze; tiles, lock and writings without over

segment the images by defining objects without any architecture signification. A quite "supervised" process is thus preferable.

4. Conclusion

This paper introduced the ASPICO platform as an archive and semantic advanced management system. It described the knowledge structure and interface based on multilingual ontological management. It investigated the possible solutions for automatic shape identification and then motivated the appropriate strategy to be adopted by the Digital Silk Roads project in order to perform indexing and efficient retrieval on digital cultural content through ASPICO.

In a next step, it will integrate color calibrated multi-resolution image management for both 2D and 3D objects. Finally, the platform is extensible and is able to evolve over time – allowing for extra modules, for example, for image content-control and secure remote printing (e.g. DPJ's VFZ modules).

Bibliography

[Blake 1999] Blake A. and Isard M. (1999). Active contours. Springer. 352 p.

[Cootes 1994] Cootes T.F., Hill A., Taylor C.J. Haslam J. (1994). Use of active shape models for locating structures in medical. In: Image Vision Computing 12(6), p 355-366.

[Cootes 1995] Cootes T.F., Taylor C.J., Cooper D.H., Graham J. (1995). Active shape model – their training and application. In: Computers Vision Image understanding 61(1). P 38-59.

[Davison 2000] Davison N.E., Eviatar H., Somorjai R.L. (2000). Snakes simplified. In: Pattern Recognition 33.p 1651-1664.

[Jain 1988] Jain A.K., Zhong Y., Dubuisson-Jolly M-P. (1998). Deformable template models: A review. In: Signal Processing 71, p 109-129.

[Godard 2003] Godard, J., Andres, F., Grosky, W., Ono, K.(2003). Management of Geomedia Content: Context-dependent Data Access, NII Journal No. 7, pp.9-17.

[Kass 1988] Kass M., Witkin A., Terzopoulos D. (1988). Snakes: active contours model. In: International Journal of computers vision 1 (4) (1988), p 321-331.

[Ngoi 1999] Ngoi K.P., Jia J.C. (1999). An active contour model for colour region extraction in natural scenes. In: Image and vision computing 17, p. 955-966.

[Protégé 2000] http://protege.stanford.edu/

[Ono 2001] K. Ono. Proceedings of the Tokyo Symposium for Digital Silk Roads, National Institute of Informatics, Tokyo, Japan, December 11-13, 2001, ISBN 4-86049-007-X (2002)

[Ono 2003] K. Ono. Proceedings of the Nara Symposium for Digital Silk Roads, Nara, Japan, December 10-12, 2003, ISBN 4-86049-024-X (2004)

[Stammberger 1999] Stammberger T., Eckstein F., Michaelis M., Englmeier K-H., Reiser M. (1999). Interobserver reproducibility of quantitative cartilage measurements: comparison of B-spline snake and manual segmentation. In: Magnetic Resonance imaging, 17(7). p. 1033-1042.

[Tsechpenakis 2004] Tsechpenakis G., Rapantzikos K., Tsapatsoulis N., Kollias S. (2004). A snake model for object tracking in natural sequences. In: Signal processing: Image communication 19. p219-238.

[Yuille 1992] Yuille A., Hallinan P., Cohen D. (1992). Feature extraction from faces using deformable templates. In: International Journal of Computer Vision, 8(2), p 99-111.

IMAGESPACE: AN ENVIRONMENT FOR IMAGE ONTOLOGY MANAGEMENT

Shiyong Lu, Rong Huang, Artem Chebotko, Yu Deng, and Farshad Fotouhi

Department of Computer Science, Wayne State University, Detroit, MI 48202, USA; {shiyong; f10272; artem; yudeng; fotouhi}@cs.wayne.edu

Abstract: More and more researchers have realized that ontologies will play a critical role in the development of the Semantic Web, the next generation Web in which content is not only consumable by humans, but also by software agents. The development of tools to support ontology management including creation, visualization, annotation, database storage, and retrieval is thus extremely important. We have developed ImageSpace, an image ontology creation and annotation tool that features (1) full support for the standard web ontology language DAML+OIL; (2) image ontology creation, visualization, image annotation and display in one integrated framework; (3) ontology consistency assurance; and (4) storing ontologies and annotations in relational databases. It is expected that the availability of such a tool will greatly facilitate the creation of image repositories as islands of the Semantic Web.

Keywords: Ontology, visualization, annotation, Semantic Web, DAML+OIL, ontology storage, ontology-based retrieval.

1. Introduction

More and more researchers have realized that ontologies will play a critical role in the development of the Semantic Web, the next generation Web in which content is not only consumable by humans, but also by software agents. [1,5]. Undoubtedly, images will be major constituents of the Semantic Web, and how to share, search and retrieve images on the Semantic Web is an important but challenging research problem. Unlike other resources, the semantics of an image is implicit in the content of an image. Although this is not a problem to human cognition, it imposes a challenge on image searching and retrieval based on the semantics of image content. Manual annotation of images provides an opportunity to make the semantics of an image explicit and richer. However, different

annotators might use different vocabulary to annotate images, which cause low recall and precision in image search and retrieval. We propose an ontology-based annotation approach, in which an ontology is created for a particular domain so that the terms and their relationships are formally defined. In this way, annotators of a particular image domain, say, the Family Album domain, will use the same vocabulary to annotate images, and users will search images guided by the ontology with greater recall and precision.

In [11, 12], we have briefly described *ImageSpace*, an image ontology creation and annotation tool, and our experience of annotating linguistic data using *ImageSpace* for the preservation of endangered languages [13, 17, 18]. This paper extends these results with ontology visualization, the storage of ontologies and annotations in relational databases, and ontology-based information retrieval. In summary, the contributions of this paper are:

- *ImageSpace* supports the functionality of ontology creation. In particular, it facilitates the creation of classes, properties, and relations between classes and relations between properties. It also provides ontology consistency assurance;
- *ImageSpace* provides full support for the standard web ontology language DAML+OIL [2];
- *ImageSpace* supports the visualization of an ontology to enable users to navigate, zoom-in and zoom-out various portions of an ontology.
- *ImageSpace* supports ontology-driven annotation of images.
- *ImageSpace* supports the storage of ontologies and annotations in a relational database.
- Finally, we have developed a simple web-based image retrieval system to search images.

Organization. The rest of the paper is organized as follows. Section 2 describes related work. Section 3 gives a brief primer for the DAML+OIL ontology language. Section 4, section 5 and section 6 present how to create and visualize an image ontology, and annotate images based on the created ontology using *ImageSpace*. Section 7 describes our approach to store ontologies and annotations in relational database. Section 8 gives an overview of a prototype image retrieval system. Finally, Section 9 concludes the paper and presents some future work.

2. Related work

Extensive research has been conducted on the processing, searching and retrieval of images [6]. Recently, due to the vision of the Semantic Web [1, 5] and the important role of ontologies, there is an increasing interest in ontology-based approaches to image processing and early results show that the use of ontologies can enhance classification precision [8] and image retrieval performance [7].

Numerous ontology creation tools have been developed. Among them, *Protégé* (http://protege.stanford.edu/), developed at Stanford University, and *OntoEdit* [9] are two well-known representatives. While some of these tools provide partial support to DAML+OIL, *ImageSpace* provides full support of this language, and integrate image ontology creation, image annotation and display in one framework. The tool is built particularly with image support in mind and features a user-friendly interface support for image display and ontology-driven annotation capabilities.

Recently, independently and concurrently, *Protégé* has released five publicly accessible plugins that provide capabilities for ontology visualization: *ezOWL*, *Jambalaya*, *OntoViz*, *OWLViz*, and *TGViz*. *ezOWL* supports graphical ontology building. *ezOWL* and *OntoViz* have *ERWin*-like views of ontology classes (rectangles with names) with their properties and restrictions (“attribute” fields in rectangles). *Jambalaya* [10] provides nested interchangeable views and nicely implements three zooming approaches: geometric, semantic and fisheye zooming. *OWLViz*, and *TGViz* have graph-like views of ontologies. *OWLGraph* shares a lot of features with these tools but it provides a richer set of views and layouts. For more details of the features of *Protégé*, the reader is referred to http://protege.stanford.edu/plugins/domain_visualization.html.

3. A primer on DAML+OIL

DAML+OIL is a semantic markup language for publishing and sharing ontologies on the World Wide Web. It is developed as an extension of XML [14], RDF [15] and RDF Schema (RDF-S) [16] by providing additional constructs along with a formal semantics. DAML+OIL uses 44 constructs (or XML tags) to define ontologies, classes, properties, individuals, data types and their relationships. In the following, we present a brief overview of the major constructs and refer the reader to [2] for more details.

Classes. A class defines a group of individuals that share some

properties. A class is defined by *daml:Class*, and different classes can be related by *rdfs:subClassOf* into a class hierarchy. Other relationships between classes can be specified by *daml:sameClassAs, daml:disjointWith*, etc. The extension of a class can be specified by *daml:oneOf* with a list of class members or by *daml:intersectionOf* with a list of other classes.

Properties. A property states relationships between individuals or from individuals to data values. The former is called *ObjectProperty* and specified by *daml:ObjectProperty*. The later is called *DatatypeProperty* and specified by *daml:DatatypeProperty*. Similarly to classes, different properties can be related by *rdfs:subPropertyOf* into a property hierarchy. The domain and range of a property are specified by *rdfs:domain* and *rdfs:range* respectively. Two properties might be asserted to be equivalent by *daml:samePropertyAs.* In addition, different characteristics of a property can be specified by *daml:TransitiveProperty, daml:UniqueProperty,* etc.

Property restrictions. A property restriction is a special kind of class description. It defines an anonymous class, namely the set of class of all individuals that satisfy the restriction. There are two kinds of property restrictions: *value constraints* and *cardinality constraints*. Value constraints restrict the values that a property can take within a particular class, and they are specified by *daml:toClass, daml:hasClass, etc*. Cardinality constraints restrict the number of values that a property can take within a particular class, and they are specified by *daml:minCardinality, daml:maxCardinality, daml:cardinality*, etc.

Recently, DAML+OIL [2] has been revised into OWL, which is a Web ontology language that has become a W3C recommendation [3].

4. Creating an image ontology

ImageSpace provides a user-friendly interface to the user to create image ontologies. Figure 1 shows a snapshot of creating an image ontology *FamilyAlbum*. The four tabs, labeled by *Ontology*, *Class*, *Property* and *Instance,* facilitate the specification of these components and their relationships in a graphical fashion.

As shown in Figure 1, when the *Class* tab is enabled, the left frame displays the class hierarchy, and the right frame shows the relationships of this class with other classes including restriction classes. With this interface, one can easily insert, delete, and update a class. In addition, using the right

frame, one can specify the relationships of this class with other classes. At the right-bottom corner of the right frame, is a panel that corresponds to property restrictions, where a user can specify both value constraints and cardinality constraints. Note that those shaded property restrictions are automatically inherited from their parent classes unless they are overridden. Also, note that, since a class might have multiple parents, other parent classes are shown in the *SubClassOf* field.

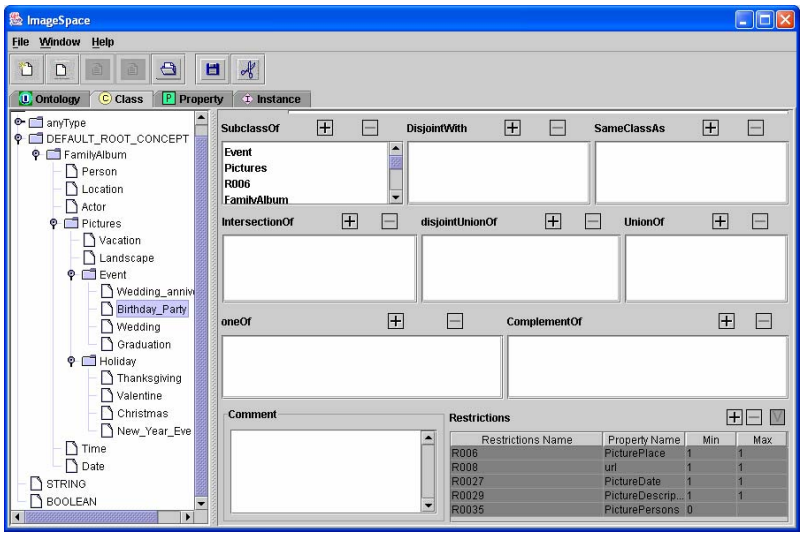


Figure 1. A snapshot of creating image ontology FamilyAlbum
(http://www.cs.wayne.edu/~shiyong/ontology/FamilyAlbum.daml)

When the user enables the *Property* tab, similarly, the left frame shows the property hierarchy, in which parent-child relationship associates the *subPropertyOf* relations between properties. On the right frame, one can specify the type, domain, range of a property. In addition, one can relate a property to other properties in the fields of *InverseOf* and *SamePropertyAs*. Also, note that, since a property might have multiple parents, other parent properties are shown in the *SubPropertyOf* field.

Creating restrictions is a part of the definition of a class. It creates an anonymous class. For example, we define a class *Pictures*. Every instance of a class *Pictures* must have a *PicturePlace*. In this case, we define a

restriction. Each restriction must have a property called *onProperty*. In other words, that means the restriction is imposed on that property. We can also define a local range using *toClass*, and *hasClass*, and the number for range (*cardinality*, *minCardinality*, *maxCardinality*). The definition of qualification is a part of the restriction. It has *hasClassQ* and the number for range (*cardinalityQ, minCardinalityQ, maxCardinalityQ*). Because restriction is an anonymous class, we represent a restriction with the relation (*subClassOf*, *complementOf*, *unionOf*, *disjoinWith*, *disjointUnionOf*, *sameClassAs*, *intersectionOf*) within the class. For example, when defining a class *Pictures*, *SubClassOf* field contains a restriction on the property *PictureDate*. Its range (*toClass*) is *dateTime* and the number for that range (*cardinality*) is 1. In order to keep the consistency of ontology, we check whether *maxCardinality* is larger than *minCardinality*. If we define the *toClass*, it should not have *hasClass* and qualification; the reverse should agree as well. *Birthday_Party* class (shown on figure 1) has inherited all restrictions from its parent *Pictures*.

The consistency of an ontology is essential and special cares must be taken in order to create a consistent ontology. For example, if Class A is specified as the parent class of Class B, then Class A cannot be in the *complementOf* class list of Class A. *ImageSpace* uses the following four mechanisms to ensure creating only consistent ontologies: (1) *No action.* If an insert, delete or update of a component will violate the consistency of the whole ontology, then the action is cancelled with a warning given to the user to indicate the reason of such cancellation. (2) *Cascaded action.* When an offending action occurs, it triggers another or a series of other recovering actions to occur so that the consistency of the ontology is maintained. For example, when a class is deleted, then all references to the class will be deleted as well provided that such cascaded deletion will not cause inconsistency of the ontology. (3) *Using a filter.* To prevent consistency violating action from occurring, a filter is used to restriction the actions that a user can perform. For example, in the *disjointWith* field of a class, a filter is used so that no ancestor classes of this class can be chosen as a class in the *disjointWith* list. (4) *Validation before submission.* This mechanism is used, for example, in the *instance* interface. After an image is annotated, constraints such as cardinality constraints are checked, and if some inconsistencies occur, then the submission is cancelled, with an error message prompted to the user. The submission will not be committed until all constraints are satisfied. A detailed description of all the consistency

checks that are performed by *ImageSpace* is beyond the scope of this paper. Interested readers are referred to [4] for details.

5. Ontology visualization

Ontology visualization plays an important role in understanding and maintaining the structure of large knowledge bases. Ontology creation tools usually have many tabs and dialog windows, because of complex relationships and dependencies among classes, properties, and restrictions. As a result, one of the problems that users experience while navigating large ontologies is disorientation.

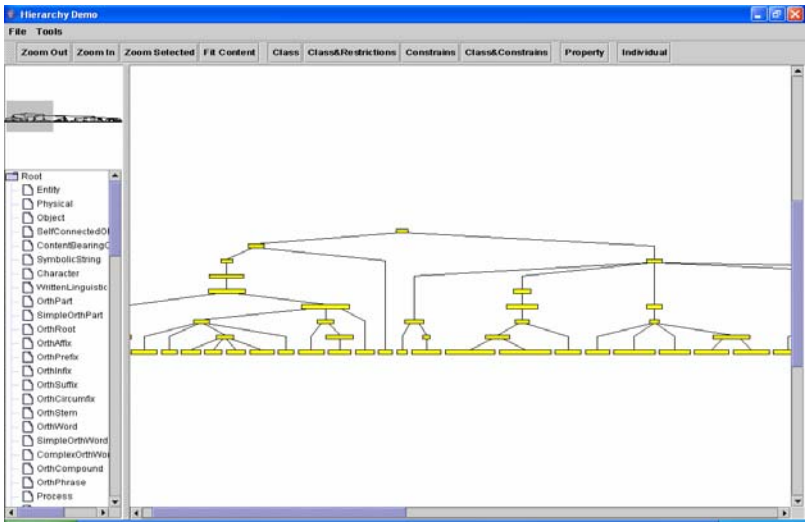


Figure 2. A snapshot of an ontology visualization
(class view, hierarchical layout)

We have developed a tool for ontology visualization that can work as a stand-alone application as well as an *ImageSpace* plugin. It provides simple and user-friendly interface for graphical navigation through ontology.

Figure 2 shows a snapshot of a sample ontology visualization. The tool main window has a menu, a toolbar, and 3 frames: left upper frame shows preview of a whole ontology graph; right frame shows main view of an ontology; left bottom frame shows a list of classes. A user can use all 3 frames to navigate an ontology.

The visualization plugin supports the following views/hierarchies: class; class and restrictions; constrains; class and constrains; property; and individual. Various concepts (class, property, individual) have different colouring scheme. In addition, a user can experiment with 3 highly customizable layouts: hierarchical, orthogonal, and organic. Figure 2 shows a class view of an ontology displayed with hierarchical layout.

Finally, we provide support of such common features like zooming (in, out, selected content, frame fitting) and manual layout of graphical primitives.

6. Annotating an image

One attractive feature of *ImageSpace* is that, it nicely integrates annotation of images into one framework. The *Instance* tab corresponds to this functionality. Figure 3 displays a snapshot of annotating an image using *ImageSpace*. The left frame shows the class hierarchy and instances (shown by I-icons) associated with the classes to which they belong. The interface on the right frame is ontology-driven. In other words, for different ontologies and different classes, the interface will be generated dynamically based on the properties, cardinality constraints specified for the ontology. For example, for the *FamilyAlbum* ontology, the interface will contain fields *PicturePersons*, *PictureDate*, *PictureDescription* (hidden), etc. While *PictureDate* and *PictureDescription* are *DatatypeProperties*, *PicturePersons* is an *ObjectProperty* that relates an image to a list of actors. Here, an *actor* models a particular snapshot of a person in a particular picture. In the example given, there are two actors. The + button on the right of *PicturePersons* field allows a user to pop up a dialogue window to choose from a list of actors, in which the +/- buttons facilitates the insert/delete of actors in this list. This nested dialogue interface greatly facilitates a user to create instances in an on-the-demand fashion. For example, the insert of an actor might require a person to be inserted first, the nesting order of the dialogue windows ensures that a referenced instance is inserted before a referencing instance is inserted. In our example, the *FamilyAlbum* ontology will enable a user to model that two actors, say *Kathleen-actor1* and *Kevin-actor1*, exist in the picture, that these two actors are for persons *Kathleen* and *Kevin*, and *Kathleen-actor1* hugs *Kevin-actor1* in the picture. In this way, an intelligent semantic search such as "*return all the vacation pictures in which Kathleen hugs Kevin*" can be supported.

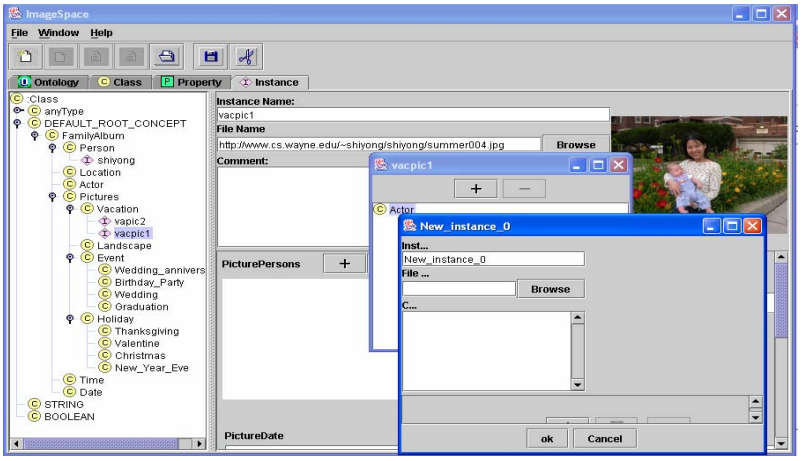


Figure 3. A snapshot of annotating an image

7. Storing ontologies and annotations in a relational database

Both ontologies and annotations are saved in a relational database for the support of ontology-driven search of images. We describe our database design in terms of the following tables that we create where primary keys are underlined:

- Ontology(OntologyID, versionInfo, comment)
- Import(OntologyID, importedOntologyID)
- Class(classID, ontologyID, type, label, comment)
- SubClassOf(classID, parentClassID)
- DisjointWith(classID, otherClassID)
- DisjointUnionOf(classID, otherClassID)
- UnionOf(classID, otherClassID)
- SameClassAs(classID, otherClassID)
- IntersectionOf(classID, otherClassID)
- ComplementOf(classID, otherClassID)
- OneOf(classID, instanceID)
- Property(propertyID, ontologyID, type, comment)
- SubPropertyOf(propertyID, parentPropertyID)
- PropertyDomain(propertyID, classID)
- PropertyRange(propertyID, classID)
- SamePropertyAs(propertyID, otherPropertyID)
- InserseOf(propertyID, classID)
- Restriction(restrictionID, onProp, toClass, minC, maxC, C)

- HasClass(restrictionID, classID)
- HasValue(restrictionID, value)
- HasClassQ(restrictionID, classID, minC, maxC, C)
- Instance(instanceID, classID)
- InstanceRelationship(instanceID, propertyID, value)
- DifferentInvividualFrom(instanceID, otherInstanceID)
- SameIndividualAs(instanceID, otherInstanceID)

As an example, consider an image where Kathleen smiles and hugs Kevin, and Kevin cries. An appropriate annotation can be stored in relational tables *Instance* and *InstanceRelationship*, which are shown in table 1 and table 2 correspondingly. In practice, for efficiency concerns, we split *InstanceRelationship* table to set of tables with names that correspond to *propertyID* attribute value and with attributes *subject* (corresponds to *instanceID*) and *value*. Thus, the final schema for our example will contain the following tables (instead of *InstanceRelationship*):

- hasActor (subject, value)
- hugs (subject, value)
- hasAction (subject, value)
- hasName (subject, value)
- isSnapshotOf (subject, value)

Table 1. Relational table Instance

instanceID	classID
Kathleen	Person
Kevin	Person
http://www.cs.wayne.edu/example.jpg	Vacation
Kathleen-actor1	Actor
Kevin-actor1	Actor

Table 2. Relational table InstanceRelationship

instanceID	propertyID	value
http://www.cs.wayne.edu/example.jpg	hasActor	Kathleen-actor1
http://www.cs.wayne.edu/example.jpg	hasActor	Kevin-actor1
Kathleen-actor1	hugs	Kevin-actor1
Kathleen-actor1	hasAction	smiles
Kevin-actor1	hasAction	cries
Kathleen	hasName	Kathleen
Kevin	hasName	Kevin
Kathleen-actor1	isSnapshotOf	Kathleen
Kevin-actor1	isSnapshotOf	Kevin

8. Ontology-based image retrieval

Based on this database schema presented in the previous section, we have developed a simple web-based image retrieval system to search images. The system provides an interface to allow the user to navigate to images under different categories. In addition, a user can specify a list of "triples" as the search criterion to retrieve images. For example, one can specify a search criterion such as return all the images under the "vacation" category such that

- Kathleen hugs Kevin, and
- Kathleen smiles, and
- Kevin cries.

The following datalog-style query will retrieve the needed photos where variables are prefixed by a '$':

```
Answer ($instanceID):-
        instanceOf ($instanceID, Vacation),
        hasActor ($instanceID, $A1),
        hasActor ($instanceID, $A2),
        isSnapshotOf ($A1, $P1),
        isSnapshotOf ($A2, $P2),
        hasName ($P1, "Kathleen"),
        hasName ($P1, "Kevin"),
        hugs ($A1, $A2),
        hasAction ($A1, smiles),
        hasAction ($A2, cries).
```

Finally, query is translated to the following sequence of SQL statements:

- Select all actors for “Kathleen” and store them into *KathleenActor*.

```
SELECT isSnapshotOf.subject
FROM isSnapshotOf, hasName
WHERE isSnapshotOf.value = hasName.subject AND hasName.value = ‘Kathleen’
```

- Select all actors for “Kevin” and store them into *KevinActor*.

```
SELECT isSnapshotOf.subject
FROM isSnapshotOf, hasName
WHERE isSnapshotOf.value = hasName.subject AND hasName.value = ‘Kevin’
```

- Select all “smiling” actors for “Kathleen” and store them into *SmilingKathleenActor*.

```
SELECT hasAction.subject
FROM KathleenActor, hasAction
WHERE KathleenActor.subject = hasAction.subject AND hasAction.value = ‘smiles’
```

- Select all “crying” actors for “Kevin” and store them into *CryingKevinActor*.

```
SELECT hasAction.subject
FROM KevinActor, hasAction
WHERE KevinActor.subject = hasAction.subject AND hasAction.value = ‘cries’
```

- Retrieve all images that satisfy all specified conditions.

```
SELECT H1.subject
FROM hasActor H1, hasActor H2, Hugs
     SmilingKathleenActor, CryingKevinActor
WHERE H1.subject = H2.subject AND
   H1.value = SmilingKathleenActor.subject AND
   H2.value = CryingKevinActor.subject AND
   Hugs.subject = SmilingKathleenActor.subject
   AND Hugs.value = CryingKevinActor.subject
```

All and only the images that satisfy this criterion will be returned (in our case, http://www.cs.wayne.edu/example.jpg). The reader is referred to [4] for more details about the ontology-driven image retrieval system.

9. Conclusions and future work

We have developed *ImageSpace*, an image ontology creation, visualization and annotation tool that fully supports the standard DAML+OIL ontology language and enables the storage of ontologies and annotations in a relational database. Future work includes:

- The development of *MultimediaSpace* that will not only support the annotation of images, but also other multimedia resources such as videos, audios, etc.
- Future version of *MultimediaSpace* will also support OWL, the successor of DAML+OIL.
- The development of graphical ontology building features to support by *MultimediaSpace* and visualization plug-in.

- Better optimization of SQL queries that are generated by image retrieval system.

Bibliography

[1] S. Lu, M. Dong and F. Fotouhi, “The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications”, International Journal of Information Research, 7(4), 2002.

[2] F. Harmelen, P. Patel-Schneider and I. Horrocks, “Reference Description of the DAML+OIL Ontology Markup Language”, http://www.daml.org/2001/03/reference, March 2001.

[3] S. Bechhofer, F. Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider and L Stein, “OWL Web Ontology Language Reference”, W3C Recommendation. http://www.w3.org/TR/owl-ref/. February, 2004.

[4] R. Huang, “ImageSpace: A DAML+OIL Based Image Ontology Creation and Annotation Tool”, master thesis, advisor: Dr. Shiyong Lu, Department of Computer Science, Wayne State University. December 2003.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. “The Semantic Web”, Scientific American. May 2001.

[6] Y. Rui, T. Huang, and S. Chang, “Image retrieval: current techniques, promising directions and open issues”, Journal of Visual Communication and Image Representation, Vol. 10, 39-62, March 1999.

[7] A. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, “Ontology-based Photo annotation”, IEEE Intelligent Systems, 16(3), pp. 66-74, 2001.

[8] A. Ponnusamy, C. Breen, L. Khan, and L. Wang, “Ontology-based image classification using neural networks”, in SPIE: The International Society for Optical Engineering, Boston, MA, USA, July 2002.

[9] Y. Sure, M. Erdmann, J. Angele, S. Staab, R. Studer, and D. Wenke, “OntoEdit: Collaborative Ontology Development for the Semantic Web”, Proc. of the first International Semantic Web Conference 2002 (ISWC 2002), June 9-12 2002, Sardinia, Italia

[10] M. Storey, M. Musen, J. Silva, C. Best, N. Ernst, R. Fergerson and N. Noy, “Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé”, appeared in "Workshop on Interactive Tools for Knowledge Capture", K-CAP-2001, October 20, 2001, Victoria, B.C. Canada.

[11] R. Huang, S. Lu and F. Fotouhi, "ImageSpace: An Image Ontology Creation and Annotation Tool", in Proc. of the 19th International Conference on Computers and Their Applications (CATA'2004), pp. 340-343, Seattle, WA, USA, March, 2004.

[12] S. Lu, R. Huang, and F. Fotouhi, "Annotating Linguistic Data with ImageSpace for the Preservation of Endangered Languages", in Proc. of the

19th International Conference on Computers and Their Applications (CATA'2004), pp. 193-196, Seattle, WA, USA, March, 2004.

[13] S. Lu, D. Liu, F. Fotouhi, M. Dong, R. Reynolds, A. Aristar, M. Ratliff, G. Nathan, J. Tan, and R. Powell, "Language Engineering For The Semantic Web: A Digital Library For Endangered Languages", International Journal of Information Research, vol.9, no.3, April 2004.

[14] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible Markup Language (XML) 1.0 (Third Edition) W3C Recommendation 04 February 2004". 2004.

[15] D. Beckett and B. McBride, "RDF/XML Syntax Specification", W3C Recommendation 10 February 2004.

[16] D. Brickley, R. Guha, and B. McBride, "RDF Vocabulary Description Language", W3C Recommendation 10 February 2004.

[17] J. Stefan, R. Reynolds, F. Fotouhi, A. Aristar, S. Lu, and M. Dong, "Evolution Based Approaches to the Preservation of Endangered Natural Languages", in Proc. of the IEEE International Congress on Evolutionary Computation, pp. 1980-1987, Canberra, Australia, December, 2003.

[18] W. Grosky, F. Fotouhi, A. Aristar, S. Lu, M. Dong, and R. Reynolds, "A Digital Library for Endangered Languages", the Nara Symposium for Digital Silk Roads (DSR), pp. 85-92, Nara, Japan, December, 2003.

Information Materials

In memoriam

Bulgarian computer science lost a prominent colleague.

Dimitar Petrov Shishkov

January 22, 1939 Varna – March 8, 2004 Sofia

D. Shishkov graduated mathematics at Sofia University in 1962. In the last year of his studies he started a specialization as a programmer at the Joint Institute of Nuclear Research – Dubna which lasted three years. Then he worked at the Institute of Mathematics for two years. In 1966 D. Shishkov together with a group of experts transferred to the newly created Central Laboratory for Information Technologies. In 1976 he defended his PhD dissertation. He has been an associate professor in computer science at Sofia University since 1985 and a professor in computer science since 2000.

His scientific interests and results were in the fields of computer architectures, computational linguistics, artificial intelligence, numerical methods, data structures, etc. He was remarkable with his teaching activities.

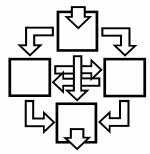
D. Shishkov was the creator of high-quality software for the first Bulgarian electronic calculator "ELKA" – one of the first calculators in the world as well as for the series of next calculators and for specialized minicomputers.

He was the initiator of the international project "Computerization of the natural languages".

He was a member of a range of international scientific organizations. Among his numerous activities was the organization of the First Programming Competition in 1979.

D. Shishkov was the initiator of sport dancing in Bulgaria (1967) and founder of the first sport-dancing high school education in the world.

D. Shishkov was a highly accomplished person with a diversity of interests, with a developed social responsibility and accuracy in his work.



ASSOCIATION FOR THE DEVELOPMENT OF THE INFORMATION SOCIETY

Acad. G. Bonchev St., block 8, Sofia 1113, Bulgaria
Tel. (+359-2) 979-3813, -3808, Fax (+359-2) 739-808
e-mail: ario@math.bas.bg, adis@einet.bg
http://www.adis.org

The Association for the Development of the Information Society (ADIS) was established in April 1997 and is an independent, non-government, non-profit organisation. The main goal of the Association is to assist in the development of the information society in Bulgaria and in the Balkans as part of the global information society. The Association has as members, besides individual persons, a number of organisations—collective members from various regions of Bulgaria: Technical University—Gabrovo, National Sports Academy (Sofia), the Police Academy, the Institute of Mathematics and Informatics, the Institute of Information Technologies, the Central Laboratory of Computer Security of the Bulgarian Academy of Sciences (Sofia), and other organisations. Societies in the cities of Plovdiv, Shoumen, and Bourgas have been formed as autonomous subsidiaries of the Association. Its membership and associated structures are growing quickly and already include foreign members. The Association has existed since recently but it unites people and organisations with several decades of experience in the field of computer science and information technologies.

The Association was established with the non-commercial objective to support the development of the information society. This objective is extensively defined in the Association's statute and includes:

- Interaction with individuals and organisations working for the development of the information society in Bulgaria and in the world.
- Support of the comprehensive utilisation of the capacity of the information infrastructure and information technologies by all layers of society and all ages and professions, as well as by unemployed, ethnic minorities, people with disabilities, etc.

- Development and implementation of national and international projects whose goal is establishing, developing, and governing the information society.
- Participation in the elaboration and implementation of educational, promotional, and demonstration programmes dedicated to information society issues.
- Participation in international activities on issues of the development of the information society, and maintenance of ties to and interaction with foreign and international organisations.
- Organisation of conferences, forums, workshops dedicated to the information society.
- Publishing of a newsletter distributed among the individual and collective members of the Association.

The Association for the Development of the Information Society has been for the last eight years the main organiser of the international conference *Information and Communication Technologies and Programming.*

Since 1999, the Association has organised monthly national seminars in the framework of the Forum Global Information Society. The seminars are devoted to the development of the information society in all fields of the human activities and aspects.

Other activities include implementing a project for training disabled (deaf) people to use computers and the Internet, a project for training secondary school teachers in a broad range of computer technologies, participation in the drafting of the Bulgarian national strategy for the Information Society, drafting of models and principals for creating, management and development of public centers for access to Internet, information and communication services and public e-information and e-services for the Bulgarian citizens as well as delivering of talks on Information Society issues at various national and regional events by members of the Association.

The Association gladly welcomes contacts with organisations from abroad whose activities are related to the development of the global information society.

International Journal
"Information Theories and Applications"
ISSN 1310-0513 www.foibg.com e-mail: foi@nlcv.net

I J I T A

Ten successful years !

Verba volant, scripta manent !

The progress in Information and Computer Sciences is fuelled by the results of research work and the accumulation of practical experience. The concept "Information Theories & Applications" (ITA) represents the synthesis of this knowledge. The field of ITA is progressing rapidly and is constantly creating new challenges for professionals involved in it.

It is clear that there is continuing need for international forums for exchange of knowledge, experience and creative inspiration among ITA professionals in order to share and stimulate solutions.

The "**International Journal on Information Theory and Applications**" (IJ ITA) has been established in 1993 as independent scientific media.

Founder and Editor in chief of IJ ITA is Krassimir Markov.

IJ ITA is edited by the Institute of Information Theories and Applications FOI ITHEA, Bulgaria.

IJ ITA Publisher is FOI-COMMERCE Co., Bulgaria.

For ten years IJ ITA became as well-known international journal. Till now more than 300 papers of more than 500 authors have been published in 10 volumes. IJ ITA authors are widespread in 34 countries all over the world: Armenia, Belarus, Belgium, Bulgaria, Canada, Czech Republic, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Latvia, Lithuania, Mexico, Moldavia, Netherlands, Poland, Portugal, Romania, Russia, Scotland, Senegal, Spain, Sultanate of Oman, Turkey, UK, Ukraine, and USA.

International Journal
"Information Theories and Applications"
ISSN 1310-0513 www.foibg.com e-mail: foi@nlcv.net

IJ ITA major topics of interest include, but are not limited to:

INFORMATION THEORIES

General Information Theory

Philosophy and Methodology of Informatics

Abstract Information Models

Artificial Intelligence

Knowledge Discovery

Knowledge Acquisition and Formation

Distributed Artificial Intelligence

Models of Plausible Reasoning

AI Planning and Scheduling

Natural Language Processing

Neuroinformatics

Theory of Computation

Cognitive Science

Cognitive Graphics

Information Models of Business Activities

Statistical Methods

Software Engineering and Quality of the Programs

APPLICATIONS

Computing

Hyper Technologies

Object and Cell Oriented Programming

Program Systems with Artificial Intelligence

Intellectualisation of Data Processing

Business Informatics

Information Systems

Pyramidal Information Systems

Intelligent Information Systems

Very Large Information Systems

Multimedia Systems

Business Information Systems

Graphics Systems

Communication Systems

Statistical Systems

Special Applied Systems

Computer Art and Computer Music

International Journal
"Information Theories and Applications"
ISSN 1310-0513 www.foibg.com e-mail: foi@nlcv.net

Call for papers

IJ ITA welcomes scientific papers connected with any information theory or its application.

Original and non-standard ideas will be published with preferences.

Papers must be written in English.

Responsibility for papers published in IJ ITA belongs to authors.

Please get permission to reprint any copyrighted material before you send it to IJ ITA.

IJ ITA rules for preparing the manuscripts are compulsory.

The rules for the papers for IJ ITA as well as the fees are given on
www.foibg.com/ijita .

The camera-ready copy of the paper should be received by e-mail:
foi@nlcv.net

IJ ITA ISSN 1310-0513

International Prize "ITHEA"

International Prize "ITHEA" is aimed to mark original and non-standard information theories and applications.

Prize "ITHEA" is established by FOI Institute for Information Theories and Applications.

Every year, an International Scientific Jury selects the works to be awarded by Prize ITHEA in following divisions: General Information Theory; Software Engineering; Artificial Intelligence; Business Informatics; Computer Art; Special Applied Systems.

The awarded persons are listed below (in alphabetical order):

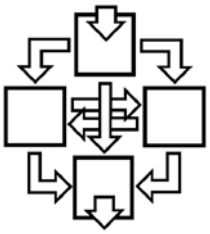
1995 Sandansky	K. Bankov, P. Barnev, G. Gargov, V. Gladun, R. Kirkova, S. Lazarov, S. Pironkov, V. Tomov
1996 Sofia	T. Hinova, K. Ivanova, I. Mitov, D. Shishkov, N. Vashtenko
1997 Yalta	Z. Rabinovich, V. Sgurev, A. Timofeev, A. Voloshin
1998 Sofia	V. Jotsov
1999 Sofia	L. Zainutdinova
2000 Varna	I. Arefiev, A. Palagin
2001 St.Peterburg	N. Ivanova, V. Koval
2002 Primorsko	A. Milani, M. Mintchev
2003 Varna	T. Gavrilova, A. Eskenazi, V. Lozovskiy, P. Stanchev

Invitation to Participation

Thirtieth Jubilee International Conference

INFORMATION AND COMMUNICATION TECHNOLOGIES AND PROGRAMMING ICT&P 2005

Knowledge-Based Society: Perspectives and Challenges



June 23-25, 2005

Contacts:
Sofia 1113, Bulgaria, Acad. G. Bonchev St., Block 8
Tel.: (+359-2) 979-2828, -3813 Tel./Fax: (+359-2) 739-808
e-mail: ictp@math.bas.bg