Twenty-eight International Conference

INFORMATION AND COMMUNICATION TECHNOLOGIES AND PROGRAMMING

Varna, Bulgaria
June 23-25, 2003

ICT&P 2003

Proceedings

FOI-COMMERCE
SOFIA - 2003

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-eight 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
All Rights Reserved

ISBN 954-16-0026-3
C\o Jusautor, Sofia, 2003

TABLE OF CONTENTS

Invited Papers

Papers

Texts for Panel Discussions

Special session on multimedia semantics

Information about

Preface

The 28th Conference "Information and Communication Technologies and Programming" – ICT&P'03 is taking place in the resort "St. St. Konstantin and Elena" in the Varna District.

The Conference is devoted to:

- 100th anniversary of the birthday of the Father of the computer – John Atanasoff.
- 40th anniversary of the formation of the Department of Information Research at the Institute of Mathematics and Informatics
- 10th anniversary of the International Journal "Information Theories and Applications."

Other conferences are holding in the same place, in parallel to this conference to mark the 10th anniversary of the International Journal "Information Theories and Applications."

Within the framework of ICT&P'03 there will be a special session on Multimedia Semantics.

This proceedings contain texts or abstracts of the invited talks, several regular papers reviewed and accepted by the International Program Committee, and materials for the two discussions, devoted to E-Government and Mobile Communication and Positioning Systems.

The papers are included in the proceedings following the order of the Conference Program.

The proceedings also include information materials about John Atanasoff, about the Department of Information Research at the Institute of Mathematics and Informatics, about the Association for the Development of the Information Society, and about the International Journal "Information Theories and Applications."

I would like to present our 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'03. Special thanks go to the Association for the Development of the Information Society and the Institute of Mathematics and Informatics.

I would like to thank Petar Stanchev for organizing the special session 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."

I thank all the people who helped with the organization, and especially
Mrs. Detelina Stoilova – the secretary of the Organization Committee.

Petar Barnev

Invited Papers

E-GOVERNMENT- A NEW CHAPTER OF DEMOCRACY

Tibor Vamos

Computer And Automation Institute,
Hungarian Academy
11 Lagymanyosi, Budapest, Hungary, 1111
vamos@sztaki.hu

Abstract.

E-government is not the machine version of existing administration-governance practice. Deep philosophical and legal issues must be reconsidered in the relations of citizen and governance, the meaning of servicing state, privacy and openness relations, empowering. Data unification is a key legal and technical issue. Systems should survive technological changes and be adaptive for changes in regulations. Artificial intelligence, especially linguistic tools, natural language understanding, decision support offer challenging tasks. Review of recent experiments and results is given.

LEGAL ASPECTS OF E-GOVERNMENT

Vihar Kiskinov

Assoc. Prof. in Sofia University “St. Kliment Ohridski”

The study presents methods for the development of E-government, whose implementation ensures the unity between law, state organisation and information technologies.

The analysis of the international practice reveals that there are not strategies and projects for E-government, which place in the very centre of their attention the law as the means for determining the content of the information technologies. Respectively there is not a method, which considers the legal normative means for regulation of the organisational structures and processes as a basis for the development of an electronic government.

It is necessary to consider these phenomena together and to find a method, which can resolve organisational, legal and information problems at once. If there is not a method, which takes into account all these dependencies, the concepts, projects and realisations of the information technologies in the electronic government will be exposed to a number of risks. One of them is examination, modelling, presenting and consequently multiplying of informal and legally unregulated governing structures and units.

A few models for developing E-government are proposed. The ideal method ensures identity between the legal regulations, characteristic features of the public bodies and the information technologies. The other methods are its compromise versions. The disadvantages of the confined methods are un derlined, as well as the conditions for continuity between the results of these methods and the ideal one.

Under the conditions of an established method for the development of E-government the tasks for the law, state and the information technologies are outlined.

The consequences of the operation of E-government, established through the suggested methods, are versatile. A complete and consistent application of the principle of the rule of law is achieved under the new conditions of E-government. Evolution of the characteristic features of the state is expected under the impact of the E-government. The characteristic features and the expected changes are looked at. The correlation between the three forms of the existence of state — legal regulations, public institutions and the information technologies of the electronic government are studied.

The perspectives in relation to the problems of E-government are first of all related to the creation of knowledge bases of the E-government. Formalisms for presenting of management knowledge in their predominant part are formalisms for presenting legal knowledge. This is why the problem of creation of legal knowledge bases with the models of organisational structures, authority, proceedings, outcomes they contain are another important issue in the contiguous study of the general problems of the E-government.

STARTING FROM SCRATCH: CREATING AN INFORMATION TECHNOLOGY INFRASTRUCTURE FOR MEMS-RELATED RESEARCH AND DEVELOPMENT

Jeff LaFrenz, Giorgio Gattiker,
Karan V.I.S. Kaler, Martin P. Mintchev

Department of Electrical and Computer Engineering,
University of Calgary, Calgary, Alberta, Canada T2N 1N4
lafrenz@enel.ucalgary.ca, gattiker@enel.ucalgary.ca,
kaler@enel.ucalgary.ca, mintchev@enel.ucalgary.ca

Abstract: Micro Electro Mechanical Systems (MEMS) have already revolutionized several industries through miniaturization and cost effective manufacturing capabilities that were never possible before. However, commercially available MEMS products have only scratched the surface of the application areas where MEMS has potential. The complex and highly technical nature of MEMS research and development (R&D), combined with the lack of standards in areas such as design, fabrication and test methodologies, makes creating and supporting a MEMS R&D program a financial and technological challenge. A proper information technology (IT) infrastructure is the backbone of such research and is critical to its success. While the lack of standards and the general complexity in MEMS R&D makes it impossible to provide a "one size fits all" design, a systematic approach, combined with a good understanding of the MEMS R&D environment and the relevant computer-aided design tools, provides a way for the IT architect to develop an appropriate infrastructure.

Keywords: MEMS, Information Technology, Computer-Aided Design

INTRODUCTION

Micro Electro Mechanical Systems (MEMS) are highly miniaturized microscale structures, devices or completely integrated microsystems that are comprised of electrical and mechanical components fabricated using modified integrated circuit (IC) batch-processing techniques. The promise that MEMS holds is remarkable, from implantable devices for sensing, monitoring and control of bodily functions [1], to miniature positioning systems [2], accelerometers [3, 4], microfluidic pumps [5-7], microactuators [8, 9], etc. Many comparisons have been made between the early days of the Very Large Scale Integrated (VLSI) semiconductor devices and circuits industry, and the MEMS industry of today[10]. Certainly, MEMS-related research and development (MEMS R&D) has the potential for a similarly, if not even more, spectacular growth, but the realization of such progress and growth is hampered by the lack of standardization in design, fabrication and test methodology, another characteristic of the early VLSI industry

MEMS R&D covers a wide and diverse range of application areas. However, with the limited standardization between MEMS fabrication facilities, it is typical for each facility to be specialized in particular, and sometimes proprietary, techniques or technology. This means that, in general, no single facility is capable of handling the fabrication requirements for every MEMS application. As with the VLSI industry of yore, this is changing as the industry and technology matures, but reflects the reality facing MEMS researchers today.

On the surface it may seem best for an organization considering initiating and developing a MEMS R&D program to also consider developing their own MEMS fabrication process and facility. However, the cost of setting up such a facility is very high (a basic MEMS fabrication facility can cost upwards of $15 million, just for the initial equipment outlay), and if the MEMS industry continues to follow the trend that the VLSI industry has laid out, then “private” state-of-the-art MEMS fabrication facilities will not be viable in the long term, except for very specialized cases. The most practical and cost effective alternative for most organizations, and particularly in academia, is shared access with other organizations to a MEMS fabrication facility (and typically to more than one, based on the current specializations of these facilities).

By its very nature, MEMS R&D is a domain in which collaborative efforts between organizations and individuals are a necessity. The classical boundaries between various research areas are non-existent in MEMS R&D, and it is not unusual for a project to require the input and collaboration of experts in a wide range of classical areas of both science and engineering, including microbiology, chemistry, electrical engineering, mechanical engineering, chemical engineering, physics, material sciences, computing sciences and others. As with the individual, it is impractical, if not impossible, for an organization to have all the necessary expertise in house. This often leads to the need for communication and collaboration between non-collocated individuals and organizations.

The collaborative and geographically distributed nature of MEMS R&D, establishes the need to interface with a wide range of external organizations, and the computationally intensive nature of the work, makes selection of an appropriate information technology (IT) infrastructure a critical element of a successful MEMS R&D program. However, as with much else in the MEMS industry, there is an identifiable lack of standardization when it comes to the design and deployment of the associated IT infrastructure. The very nature of the industry prevents a “one size fits all” type of approach, but, as will be shown, a thorough understanding of the needs of MEMS R&D programs allows the IT architect to employ a standard systematic approach to the design and deployment of an appropriate IT infrastructure.

The aim of this paper is to provide IT architects with a better understanding of the complex requirements of MEMS R&D in order to assist them in such design efforts.

MEMS DESIGN PROCESS

The Concept or Idea

In addition to the number of different device areas where MEMS is being applied, it also encompasses a wide range of application areas including aerospace, automotive, military, storage devices, biology, medicine, etc. The wide range of application areas and types of devices makes for a rather diverse range of research ideas.

The process of determining the feasibility of any particular research idea involves sketching out the concept at a high level, assessing the basic functionality, and confirming the availability of appropriate resources (design tools, collaborators, fabrication facilities, testing tools, IT infrastructure, etc.). Although more directly relevant in the commercial environment than perhaps the academic environment, the additional factors of market potential, competitive landscape and organizational capabilities must also be taken into account.

Literature search into existing work in an application area can often eliminate months, if not years, of preliminary research. As more and more of the published literature is becoming available over the Internet, one of the first requirements on an IT infrastructure is that it provide researchers with the ability to do on-line literature searches, and save, sort and manage the results. Various bibliographic tools exist, including EndNote [11] and Reference Manager [12], to support such tasks. Extensive utilization of research databases such as Medline, Compendex, Inspec, ACM Digital Library, CISTI Source, and IEEE Xplore (see e.g., http://www.ucalgary.ca/library/gateway/indabs.html) is essential. For more industrially oriented research, patent searches (see e.g., http://www.delphion.com) are also a necessity.

Modeling

The ability to work and design with standardized high level abstractions based on process-independent design elements has been one of the instrumental reasons for the accelerated growth of the VLSI industry [10]. Unfortunately, with limited exceptions, such standardized high-level abstractions are lacking in the MEMS industry today, and thus the MEMS researcher must not only consider overall device functionality, but also take into account how the complexities of the fabrication processes, and its impact on the design. However, in many cases this information can only be obtained at the end of the fabrication and testing phases, requiring a time consuming reiteration through the whole design process.

Although standardized abstractions are lacking, it is still necessary for the MEMS designer to model MEMS designs at various levels of abstraction. Common abstraction levels are variously referred to as System, Network (or Link) and Geometry (or Physical) [13, 14]. The geometry level is the closest to the “real” device, and modeling at this level requires numerical solvers such as Finite Element Analysis (FEA). IntelliSuite [15], CoventorWare [16], Ansys [17], Abaqus [18] and others, are software tools available for this purpose. Modeling at this level is very computationally intensive [19].

Modeling at the Network level is commonly done by describing the design as systems of differential algebraic equations representing circuit equivalents of the various design elements. There are also a number of software tools available for modeling at this level, including MEMS Pro [20], SPICE, PSPICE, Sugar [21] and MATLAB [22]. Processing power is still highly important for network level modeling, although typically not as computationally intensive as at the geometry level.

System level modeling involves representing the design as a combination of various functional blocks and applying system level simulations. The biggest difficulty facing the MEMS designer is that there is no general systematic way of creating system models from physical models [14]. This means that the MEMS designers must work at all levels of abstractions and iterate between them in order to fully model and adequately simulate their design [23].

Design, Analysis and Simulation

Design of a MEMS device consists of the definition of structures that physically represent the device. The outcome of the design process is a set of masks and fabrication steps that will be used in the fabrication process. As with any Computer-Aided Design (CAD), MEMS design is highly graphical in nature. However, MEMS design must also take into account fabrication details, and so iteration between design, analysis and simulation is very common. This puts requirements on the associated computing resources for high computing performance, as well as for high performance and quality graphics processing. There are several software CAD tools available for MEMS design which also provide analysis and simulation capabilities, including MEMS Pro [20], IntelliSuite [15] and CoventorWare [16].

Iteration back to modeling, and even to the original design idea is not uncommon, as design fabrication limitations may make it impossible to implement a particular design that was initially considered feasible. This re-emphasizes the importance of taking into account fabrication details and material considerations early in the design, and of modeling these fabrication details and material properties at the geometric level.

Another point that is often missed is the need to test the final device. Test considerations must be part of the design at all phases, since fabricating a design only to find out that it doesn’t work without being able to find out why, essentially means having to start the whole design process over again.

Implementation

Transfer of the final design to a fabrication facility is typically done by electronic means (in file formats such as CIF, GDS and DXF) although physical transfer of files is also a possibility. In either case secure and reliable transfer is a requirement. IT architects must fully understand the nature and the availability of various communication alternatives (which are typically set by the MEMS fabrication facility) in order to determine which one is best suited to their particular requirements.

The MEMS fabrication facility will return the final product to the designer once the fabrication steps have been completed. This could be a complete silicon wafer, a wafer that has been diced into individual chips, or chips that have been packaged in some manner depending on the abilities of the fabrication facility and the desires of the designer. In some cases a fabrication facility will not be able to package a chip to the designer’s requirements, and so the designer must transfer the chips to another facility for packaging or perform the task him/herself if local facilities are available.

Testing

Once the final packaged device is available, it must be fully tested and characterized. Testing environments for MEMS devices are as diverse as their application areas. The IT architect must work closely with the MEMS researchers to fully understand their testing needs, and the implications on the IT infrastructure. In a general sense there will be one or more sets of testing equipment, frequently under computer control, and thus it is necessary for the computing resources to be able to support the various control interfaces that will be present. Testing and evaluation to assess operational reliability and safety are critically important to the potential deployment of MEMS devices focused on in vivo applications.

INFORMATION TECHNOLOGY (IT) INFRASTRUCTURE FOR MEMS

Functional Breakdown

Due to the lack of standardization in MEMS R&D and the variety of tools that may be used, it is necessary to tailor the associated IT infrastructure to suit the needs of the individual program. However, evaluation of these needs at a functional level allows for a systematic approach to such customization.

Functionally a MEMS R&D IT infrastructure can be broken down into general purpose workstations, modeling workstations, design and analysis workstations, testing workstations, license servers, file servers and network. Each of these may represent one or more physical entities, or may be combined together on a single physical machine, depending on the needs of the MEMS R&D group.

Functional Entity Descriptions

General-purpose workstations handle literature search, sort and management as well as general administrative tasks such as email, documentation creation, etc.

These workstations must have Internet access, as well as word processing, bibliographic, 2D or 3D drawing and diagramming software and other related tools.

Since modeling tasks are usually very computationally intensive operation, modeling workstations must be optimized for performance. It may be desirable in such instances to use high end multiprocessor workstations for modeling tasks. In many instances modeling software will represent derived models in a graphical manner, placing a strong need for high quality graphical interfaces on these machines as well.

Requirements for the design and analysis workstations are very similar to those for modeling workstations, and in a small MEMS R&D group one set of workstations is often used for both purposes. Emphasis is again placed on the provision of high performance and high quality graphical interfaces, both in resolution and speed.

Testing workstations must interface to the equipment used in the testing. The selection of such workstations relies heavily on the interfacing needs of the test equipment. High performance can be a requirement for certain types of test environments, especially when large quantities of data are being sent or received.

While it is necessary to purchase sufficient copies of each software tool to correspond to the simultaneous use expected, frequently the number of workstations will exceed the number of concurrent users of any one such tool. The most cost effective method of dealing with this is through the use of a license server. Provided that appropriate security measures are taken in the network design, this also has the advantage of allowing remote users and casual users to use the software tools, without having to provide each of them with their own licensed copy.

Performance requirements on a license server are typically fairly low, although consideration must be given to the I/O performance if many simultaneous users are expected.

The need to share files and access the same files from many different workstations points towards the need for a file server. From a system administration viewpoint, having a single location for all design and personal files, makes backing up and archiving the data much simpler as well.

Obviously, storage space and file access speed are important considerations for a file server. In addition input/output (I/O) performance and appropriate sizing and speed of backup capabilities must be considered.

As every IT architect knows, network design is critical to a smoothly running operation. Due to the often proprietary nature of MEMS R&D, security of data and systems is paramount. Use of a hardware firewall and/or proxy server with a software firewall can be considered a general requirement.

If remote users will need to access internal resources, secured access through the firewall, through such means as a Virtual Private Network (VPN) connection, will be necessary. Higher end hardware firewalls support VPN serving directly, but use of a computer as a VPN server is also an option.

Where the need exists to connect multiple labs together through a network with traffic unrelated to the MEMS R&D work (common in an academic environment) several options are available. Each lab may have its own firewall, with IP tunneling set up between them. Although slightly less secure, if intelligent switches (which support Virtual Local Area Network (VLAN)) are used to connect the labs to a common backbone, then a firewall can be established in one lab, and other labs can “share” its security. Physically connecting networks between labs is the most secure option, although in many cases difficult to implement in practice.

Example

An example of an IT infrastructure which could be used for a smaller sized MEMS R&D group is shown in Figure 1. This infrastructure was designed for multiple general purpose workstations, shared workstations for both the modeling and the design and analysis of MEMS devices, a combined license/file server, multiple test workstations and a combined Web/VPN server.

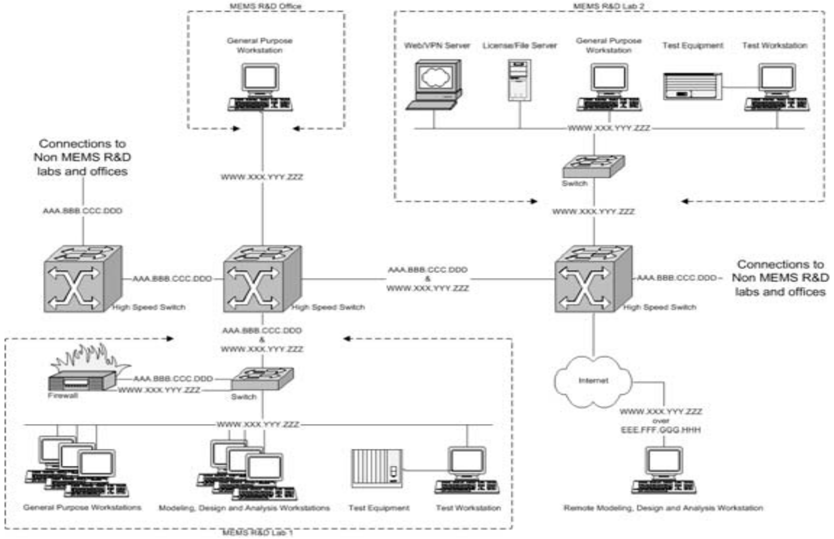

MEMS R&D Office
General Purpose Workstation
MEMS R&D Lab 2
Web/VPN Server
License/File Server
General Purpose Workstation
Test Equipment
Test Workstation
WWW.XXX.YYY.ZZZ
Switch
WWW.XXX.YYY.ZZZ
Connections to Non MEMS R&D labs and offices
WWW.XXX.YYY.ZZZ
AAA.BBB.CCC.DDD
AAA.BBB.CCC.DDD
High Speed Switch
High Speed Switch
AAA.BBB.CCC.DDD & WWW.XXX.YYY.ZZZ
AAA.BBB.CCC.DDD
Connections to Non MEMS R&D labs and offices
High Speed Switch
AAA.BBB.CCC.DDD & WWW.XXX.YYY.ZZZ
AAA.BBB.CCC.DDD
WWW.XXX.YYY.ZZZ
Firewall
Switch
WWW.XXX.YYY.ZZZ
Internet
WWW.XXX.YYY.ZZZ over EEE.FFF.GGG.HHH
General Purpose Workstations
Modeling, Design and Analysis Workstations
Test Equipment
Test Workstation
Remote Modeling, Design and Analysis Workstation
MEMS R&D Lab 1

Figure 1: Example of an Academic IT Infrastructure for MEMS R&D

The environment is academic, with a general-use backbone and intelligent switches. A port-based VLAN is set up between the various switches enabling secured access with a single firewall. Using such a configuration the physical locations of the two MEMS R&D labs and of any non-collocated collaborators becomes irrelevant, providing a truly distributed collaborative research and development environment.

DISCUSSION

Micro Electro Mechanical Systems (MEMS) is an emerging area of research, which holds the great promise of combining electrical and mechanical macro-system features on a miniature scale for a wide variety of applications ranging from microfluidics to process monitoring and control.

Since this novel research area is quite dynamic and at the same time highly technological, establishing an adequate information technology (IT) infrastructure for it represents a challenge, both structurally and financially. However, properly conducted preliminary assessment of the technology and prospective design needs reveals that relatively modest IT investment can quickly lead to an adequate MEMS design environment.

For an academic unit such a modest investment may be sufficient, as design and simulation is usually enough to penetrate the global market of ideas. However, a broader-range ambition of creating a corporate or academic MEMS laboratory capable of delivering industrially-viable MEMS products is inevitably related to the availability of adequate MEMS fabrication and testing facilities. To some degree the creation of a MEMS testing facility may be easier in an academic environment than in a corporate environment, as most academic establishments have micro and macro electronic and material testing capabilities that may also be adequate for MEMS testing. However, there is a significant difference between just exploring ideas in MEMS, and producing actual MEMS devices, and the work and commitment involved in implementing this should not be underestimated.

In any case, building, or even simply utilizing, a MEMS fabrication facility represents a substantial financial and logistic challenge and could be a major stumbling block in the completion of a full MEMS development cycle. This re-emphasizes the importance of exploring various avenues for collaborative distributed research, and fully considering the impact on the MEMS-related IT infrastructure.

CONCLUSIONS

Creating and supporting a MEMS R&D program in a corporate or academic environment can be a substantial financial and technological challenge. Establishing a proper IT infrastructure for such a program is critical to its success.

While the MEMS industry suffers from a lack of standardization in design, fabrication and test methodologies (preventing the ability to utilize a “one size fits all” approach to deploying the associated IT infrastructure) a systematic approach, with a detailed understanding of the MEMS R&D environment, provides a way for the IT architect to develop an appropriate infrastructure.

A distributed collaborative environment, incorporating standard design and modeling software systems, can be regarded as a quick and efficient method of providing MEMS design exposure in an academic environment. However, industrially-viable MEMS products can only be developed with access to appropriate MEMS fabrication and testing facilities and this must be taken into account when developing the associated IT infrastructure.

REFERENCES

[1] Ishihara, K; Tanouchi, J; Kitabatake, A; Kamada, T; and Kishimoto, S, "Noninvasive and precise motion detection for micromachines using high-speed digital subtraction echography (high-speed DSE)," In: Proceedings of IEEE Micro Electro Mechanical Systems (MEMS '91), Nara, Japan, pp. 176-181, 1991.

[2] Faulkner, NM; Cooper, SJ; and Jeary, PA, "Integrated MEMS/GPS navigation systems," In: Proceedings of Position Location and Navigation Symposium, 2002 IEEE, Palm Springs, CA, USA, pp. 306-313, April, 2002.

[3] Lee, KI; Takao, H; Sawada, K; and Ishida, M, "A three-axis accelerometer for high temperatures with low temperature dependence using a constant temperature control of SOI piezoresistors," In: Proceedings of The Sixteenth Annual IEEE International Micro Electro Mechanical Systems Conference, Kyoto, Japan, pp. 478-481, January, 2003.

[4] Chang, DT; Kubena, RL; Stratton, FP; Kirby, DJ; Joyce, RJ; and Kim, J, "Wafer-bonded, high dynamic range, single-crystalline silicon tunneling accelerometer," In: Proceedings of IEEE Sensors, Orlando, Florida, USA, pp. 860-863 vol.2, June, 2002.

[5] Mizoguchi, H; Ando, M; Mizuno, T; Takagi, T; and Nakajima, N, "Design and fabrication of light driven micropump," In: Proceedings of IEEE Micro Electro Mechanical Systems (MEMS '92), Travemunde, Germany, pp. 31-36, February, 1992.

[6] Yun, K-S; Cho, I-J; Bu, J-U; Kim, C-J; and Yoon, E, "A surface-tension driven micropump for low-voltage and low-power operations," *Journal of Microelectromechanical Systems*, vol. 11, pp. 454-461, 2002.

[7] Zeng, S; Chen, C-H; Mikkelsen, JC, Jr.; and Santiago, JG, "Fabrication and characterization of electrokinetic micro pumps," In: Proceedings of The Seventh Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITHERM 2000), Las Vegas, Nevada, USA, pp. 31-36 vol. 2, May, 2000.

[8] Sassolini, S; Del Sarto, M; and Baldo, L, "Electrostatic microactuator for future hard disk drive," In: Proceedings of Asia-Pacific Magnetic Recording Conference, Singapore, pp. WE-P-09-01-WE-P-09-02, August, 2002.

[9] Milanovic, V, "Multilevel beam SOI-MEMS fabrication and applications," In: Proceedings of 9th International Conference on Electronics, Circuits and Systems, Dubrovnik Croatia, pp. 281-285 vol.1, September, 2002.

[10] Antonsson, EK, "Executive Summary," In: Proceedings of Structured Design Methods for MEMS, Pasadena, CA, USA, pp. iii-iv, November, 1996.

[11] Software Program: *EndNote*, ver: 6, ISI ResearchSoft, http://www.endnote.com/.

[12] Software Program: *Reference Manager*, ver: 10, ISI ResearchSoft,

http://www.refman.com/.

[13] Neul, R, "Modeling and Simulation for MEMS Design, Industrial Requirements," In: Proceedings of 2002 International Conference on Modeling and Simulation of Microsystems, San Juan, Puerto Rico, U.S.A, pp. 6-9, April, 2002.

[14] Senturia, SD, "Simulation and design of microsystems: A 10-year perspective," *Sensors and Actuators A: Physical*, vol. 67, pp. 1-7, 1998.

[15] Software Program: *IntelliSuite*, Corning IntelliSense, http://www.intellisense.com.

[16] Software Program: *CoventorWare*, ver: 2003, Coventor, http://www.coventor.com/.

[17] Software Program: *Ansys*, ver: 7, Ansys Inc., http://www.ansys.com/.

[18] Software Program: *Abaqus/CAE*, ver: 6.3, Abaqus, http://www.abaqus.com/.

[19] Mukherjee, T and Fedder, GK, "Structured Design of Microelectromechanical Systems," In: Proceedings of Annual ACM IEEE Design Automation Conference, Anaheim, California, United States, pp. 680 - 685, June, 1997.

[20] Software Program: *MEMS Pro*, ver: 3.2, MEMSCAP, http://www.memscap.com/.

[21] Clark, JV; Bindel, D; Kao, W; Zhu, E; Kuo, A; Zhou, N; Nie, J; Demmel, J; Bai, Z; Govindjee, S; Pister, KSJ; Gu, M; and Agogino, A, "Addressing the needs of complex MEMS design," In: Proceedings of The Fifteenth IEEE International Conference on Micro Electro Mechanical Systems., Las Vegas, Nevada, USA, pp. 204-209, January, 2002.

[22] Software Program: *MATLAB*, ver: 13, The MathWorks, Inc., http://www.mathworks.com/.

[23] Bushyager, N; Dalton, E; Papapolymerou, J; and Tentzeris, M, "Modeling of Large Scale RF-MEMS Circuits Using Efficient Time-Domain Techniques," In: Proceedings of Applied Computational Electromagnetics Society (ACES) Conference, Monterey, CA, USA, pp. 219-224, March, 2002.

Acknowledgement: This study was supported in part by the Natural Sciences and Engineering Research Council of Canada.

THE SUBCLASSING ANOMALY IN COMPILER EVOLUTION

Atanas Radenski

Chapman University, One University Drive, Orange, California 92866, USA

radenski@computer.org, http: //www.chapman.edu/~radenski/

Abstract. Subclassing in collections of related classes may require re-implementation of otherwise valid classes just because they utilize outdated parent classes, a phenomenon that is referred to as the subclassing anomaly. The subclassing anomaly is a serious problem since it can void the benefits of code reuse altogether. This paper offers an analysis of the subclassing anomaly in an evolving object-oriented compiler. The paper also outlines a solution for the subclassing anomaly that is based on alternative code reuse mechanism, named class overriding.

1. Introduction

Object-oriented applications are collections of related classes. For example, a typical compiler incorporates (1) a set of mutually recursive syntax trees and (2) translation operations on such trees; in an object-oriented compiler, such mutually related trees are implemented as mutually related classes.

As the requirements for an object-oriented application evolve, so should do the applications itself. For example, a programming language may need to be enhanced with new linguistic features, or it may need to have existing features modified. Consequently, an object-oriented compiler for such language may need to have some of its classes adequately adapted.

Subclassing is the principal object-oriented programming language feature that provides code adaptation. (Many patterns have evolved as more robust alternatives to straight forward subclassing for adaptation purposes, but in this paper we are interested in a discussion of linguistic primitives.) Subclassing allows the derivation of new classes from existing ones through extension and method overriding. A subclass can inherit variables and methods from a parent class, can extend the parent class with newly declared variables and methods, and can override inherited methods with newly declared ones.

When a class that needs to be updated belongs to a collection of classes but is independent from all other classes from the collection, the functionality of that class can be easily updated through subclassing and method overriding. Subclassing is a straightforward code adaptation mechanism in the case of independent classes.

Unfortunately, subclassing may not properly support code adaptation when there are dependencies between classes. More precisely, subclassing in collections of related classes may require re-implementation of otherwise valid classes just because they utilize outdated parent classes, a phenomenon that has been termed as the *subclassing anomaly* (Radenski 2002). The subclassing anomaly is a serious concern since it can largely invalidate the benefits of inheritance altogether.

The goal of this paper is to offer an analysis of the subclassing anomaly as it appears in an object-oriented compiler (Section 3). This analysis is preceded by an overview of the subclassing anomaly in domain-independent manner (Section 2). The paper outlines a solution to the subclassing anomaly based on a alternative code reuse mechanism called class overriding (Section 4), and concludes with an discussion of related work (Section 5).

2. Overview of the Subclassing Anomaly

Subclassing in a collection of dependent classes may require re-implementation of otherwise valid classes just because they depend on the parent class. The need to re-implement such otherwise valid classes is referred to as the subclassing anomaly. The subclassing anomaly needs to be understood because it may seriously affect code reusability. This section is devoted to a brief overview of the subclassing anomaly. A more detailed analysis of the subclassing anomaly in a problem independent manner is presented in (Radenski 2002).

Depending on the programming language, a collection of classes can be represented as a namespace (in C#), a stateless package (in Java), or as a package with a state (in Ada 95). In this paper we utilize C# as sample language in order to provide clarity of discussion. However, all results presented in the paper can be applied equally well to virtually any compiled object-oriented language.

Let us assume that in a collection of related classes, a *container* class instantiates and utilizes an object of a *constituent* class. Let us also assume that at a later point of the existence of the collection of classes, the constituent class needs to be adapted to changing requirements, while the container class remains valid, meaning that it still provides relevant functionality and needs no changes.

Subclassing of the constituent produces an evolved constituent subclass of the original constituent class, which is then incorporated in the evolved collection. The problem is that the integrity of the evolved collection is violated, since in the evolved collection the container class still instantiates and utilizes an object of the old parent constituent class, rather than an object of the evolved constituent class. Even though the container class is assumed to provide relevant functionality, it needs to be re-implemented (which is anomaly), so that it creates an object of the evolved constituent class and thus maintains the integrity of the evolved collection.

Classes may depend on each other in various ways. Some dependencies do not cause anomalies, while others do. The so-called monomorphic dependencies, as defined below, trigger the subclassing anomaly.

Object-oriented languages allow two types of references to classes: polymorphic references and monomorphic references. A *polymorphic reference* to a class *C* stands (1) for *C* itself and (2) for all possible subclasses of *C*. A *monomorphic reference* to a class *C* stands for *C* only but not for any subclasses of *C*.

Polymorphic references to a class *C* occur in:

- parameter, variable, and constant declarations, e.g.: void f (C x); C x;
- type tests, e.g.: if (y is C) …; if (y instanceof C) …;
- type casts, e.g.: x = (C) y;

Monomorphic references to a class *C* occur in:

- constructor invocations, e.g.: *x = new C ();*
- static member access, e.g.: *C.staticMethod ();*
- subclass definitions, e.g.: *class C1 : C {…} ; class C1 extends C {…};*

A class *A* depends monomorphically on class *C* if the definition of *A* contains a monomorphic reference to *C*; further on, we skip the word monomorphically and simply say that *A depends on C*. A class *A* depends on *C* when *A* invokes the constructor of *C*, when *A* extends *C*, or when *A* refers to a static member of *C*.

The subclassing anomaly is triggered by monomorphic dependencies within a collection of classes. When the collection evolves, subclasses can be defined in order to adapt the collection to the changing environment. However, no matter how subclassing is applied, a monomorphic reference continues to stand for the outdated base class in the evolved collection. Thus, all classes that contain monomorphic references must be re-implemented, often in textually equivalent form, as members of the evolved collection. Such re-implemented classes must be recompiled so that monomorphic references are bound to up-to-date subclasses. In contrast to monomorphic references, polymorphic references to outdated base classes do not necessarily require re-implementation of the referring classes - because polymorphic references stand not only for the base class (as monomorphic references do), but for all of its subclasses as well.

3 Analysis of the Subclassing Anomaly in an Evolving Object-Oriented Compiler

This section is devoted to an analysis of the subclassing anomaly in an evolving object-oriented compiler. Our goal is to provide a non-trivial example of the subclassing anomaly as defined in the previous section and to reveal various kinds of class references that trigger the anomaly.

This is not an artificially constructed design example: it is derived from a popular book on object-oriented compilers (Watt, 2000). Depending on one's personal perspective, this design might be considered bad or good (we consider it good), but what is more important, is that it is *common* design which exhibits the subclassing anomaly.

The sample compiler has the usual three phases of syntactic analysis, contextual analysis, and code generation. As shown on Fig. 1, the three phases are implemented as *Parser*, *Checker*, and *Encoder* objects. The parser, checker, and encoder take one pass each, communicating via a syntax tree that represents the source program.

Syntax trees are defined as a hierarchical collection of interfaces and classes. On the top of the hierarchy, a *SyntaxTree* interface encapsulates methods common for all abstract syntax trees (such as a visitor method implemented by both the contextual analyzer and the code generator). Any multiple-form non-terminal symbol is represented by a single interface and several classes that implement this interface, one for each form. For example, statements are represented by the *Statement* interface and several classes that implement this interface, such as *WhileStatement*, *IfStatement*, etc.

The recursive-descent *Parser* class consists of a group of methods *parseN*, one for each non-terminal symbol *N*. The task of each *parseN* method is to performs syntactical analysis of a single *N*-form, and build and return its syntax tree. These parsing methods cooperate to perform syntactical analysis of a complete program. For example, *parseWhileStatement* performs syntactical analysis of a single *WhileStatement,* and creates and returns an instance of a *WhileStatement* syntax tree (Fig. 1).

Anomaly triggered by constructor invocation. Suppose that a developer needs to enhance all syntax trees classes with a *display* method, thus converting the *CompilerCollection* into an *UpdatedCompiler* (Fig. 1). One approach is to use subclassing in order to extend with a *display* method all original syntax tree classes, such as *WhileStatement*, *IfStatement*, etc.

Unfortunately, subclassing of the syntax tree classes does not affect any other classes form the *CompilerCollection* and all *parseN* methods from the *Parser* class continue to instantiate the old syntax tree classes. For example, the *parseWhileStatement* method form the *Parser* class, as defined in the *CompilerCollection*, instantiates class *WhileStatement* which is also defined in the *CompilerCollection*.

To effectively update the *CompilerCollection* and convert it into an *UpdatedCompiler*, the developer needs to re-implement the otherwise valid *Parser* class. The re-implementation of the *Parser* class is textually identical with the old one and only needs to be encapsulated within the *UpdatedCompiler*.

```
namespace CompilerCollection {
    public class Compiler {
            public static void compileProgram (...)
            {       Parser parser = new Parser (...);
                    Checker checker = new Checker (...);
                            Encoder encoder = new Encoder (...);
                            Program syntaxTree = parser.parse (…);
                            checker.check (syntaxTree);
                            encoder.encode (syntaxTree);
            }
    }
    public abstract class SyntaxTree {… }
    public abstract class Statement : SyntaxTree {…}
    public class WhileStatement : Statement
    { Expression e; Statement s; …}
    …
    public class Parser {…
            public Statement parseWhileStatement ()
            {       … Expression e = parseExpression ();
                    … Statement s = parseStatement ();
                    … return new WhileStatement (e, s);
            }
            ….
    }
    public class Checker {…}
    public class Encoder {…}
}
```

```
using CompilerCollection;
namespace UpdatedCompiler {
    public class WhileStatement : CompilerCollection.WhileStatement
     { public void display () {…} }
    public class IfStatement : CompilerCollection.IfStatement
    { public void display () {…} }
    …
    public class Parser {… // identically re-implemented
            public Statement parseWhileStatement ()
            {       … Expression e = parseExpression ();
                    … Statement s = parseStatement ();
                    … return new WhileStatement (e, s);
            }
            ….
    }
}
```

Figure 1. Subclassing anomaly in an evolving compiler.

The necessity to re-implement a valid *Parser* class is triggered by the subclassing of the syntax tree classes, and this phenomenon is an example of the subclassing anomaly. Note that each *parseN* method from the *Parser* class instantiates an object of class *N*, where *N* is the syntax tree representation for *N*.

These monomorphic dependencies of class *Parser* on classes *N* trigger the inheritance anomaly.

Anomaly triggered by subclass definition. A developer who needs to enhance all syntax trees classes with a *display* method needs to start with their parent class. Technically, the developer should use subclassing in the *UpdatedCompiler* in order to extend the *SyntaxTree* abstract class with a display method. Unfortunately, the *Statement* subclass of the original *SyntaxTree* class is not affected by this subclassing and remains without a *display* method, as originally defined in *CompilerCollection*. Therefore, the developer needs to re-implement in the *UpdatedCompiler* the otherwise valid *Statement* class. The re-implementation of the *Statement* class is textually identical with the old one and only needs to be encapsulated within the *UpdatedCompiler*.

The necessity to re-implement a valid *Statement* class is triggered by the subclassing of the *SyntaxTree* class, and this phenomenon is an example of the subclassing anomaly. Note that the *Statement* class is defined in the *CompilerCollection* as a subclass of *SyntaxTree*. This monomorphic dependency of the *Statement* class on the *SyntaxTree* class triggers the inheritance anomaly.

Anomaly triggered by static member access. A compiler may use classes with static members for various purposes. For example, all token kinds may be specified as static members of a *Token* class and all operation codes can be encapsulated as static members of a *Machine* class. Parser methods need to access static members of the *Token* class, e.g. *Token.While*. Suppose now that a developer of an *UpdatedCompiler* needs to enhance the *Token* class with a new token, such as *Repeat*. One approach is to use subclassing in order to extend *Token* with a *Repeat* static member. Unfortunately, subclassing of the *Token* class does not affect any of the static references to the original *Token* class, as defined in the *CompilerCollection*. All inherited parser methods continue to use the old version of the *Token class*, while new parser methods that are developed in the *UpdatedCompiler* utilize the updated *Token* class. If the developer wants to have the parser utilize the same *Token* class, the developer must re-implement the whole *Parser* class in the *UpdatedCompiler*. The re-implementations of all inherited parser methods are textually identical with the old ones and only need to be encapsulated within the updated *Parser* class.

The necessity to re-implement valid parser methods is triggered by the subclassing of the *Token* class, and this phenomenon is an example of the subclassing anomaly. Note that the parser methods access static members of *the Token* class. This monomorphic dependency of the *Parser* class on the *Token* class triggers the subclassing anomaly.

4 Elimination of the Subclassing Anomaly with Class Overriding

Class overriding, an object-oriented language feature that is complementary to subclassing, can be used to eliminate the subclassing anomaly (Radenski 2002). In contrast to subclassing, class overriding does not create a new and isolated derived class, but rather extends and updates an existing class. Class overriding is not limited to a single class but propagates across a collection of related classes: it updates all classes from the collection that refer to the class being overridden. Thus, class overriding preserves the integrity of a collection of classes by guaranteeing that any update to a class replaces the previous version of the class within the whole collection.

The definition of class overriding is based on the concept of *replication*. Replication consists in embedding a replica of each class from an existing collection of classes (*the replicated collection*) into a newly created collection of classes (*the replicating collection*). In addition to class replicas, the replicating collection can be further extended with newly defined classed or subclasses.

Replication changes class membership: while all original classes are members of the replicated collection, the class replicas become members of the replicating collection. Except for class membership, class replication preserves all other class properties, including names and access levels. In the replicating collection, each class replica is referred to by the same name and incorporates the same public, protected, and private access levels as the original class in the replicated collection.

A class replica can be overridden (meaning replaced) across the entire replicated collection with its own extension. Similarly to a subclass, the overriding class:

- inherits all data and method members of the class replica
- can override some of the inherited methods
- can extend the replica with additional data and method members

The overriding class replaces the class replica across the entire replicated collection, meaning that all classes from the replicated collection are updated to use the overriding class instead of the replica. Technically this is achieved by *late class binding*: class references are bound to particular class definitions late, at class loading time, rather than early, at compile time. This is in contrast to traditional compiled languages, such as C#, which use late binding only for methods but limits monomorphic class references to early static binding.

C#, and likewise, various other object-oriented languages, can be enhanced to support class overriding. In C#, collections of classes can be represented as namespaces. Therefore, C# is to be extended with a namespace replication statements and with class overriding definitions.

```
namespace CompilerCollection {
        public class Compiler { … }
        public abstract class SyntaxTree {… }
        public abstract class Statement : SyntaxTree {…}
        public class WhileStatement : Statement { Expression e; Statement s; …}
        …
        public class Parser {… }
        public class Checker {…}
        public class Encoder {…}
}

namespace UpdatedCompiler {
        replicate CompilerCollection;
        override public class WhileStatement  { public void display () {…} }
        override public class IfStatement { public void display () {…} }
        …
}
```

Figure 2. Elimination of the subclassing anomaly by namespace replication and class overriding.

A C# outline of a compiler that is updated by means of namespace replication and class overriding – thus avoiding the subclassing anomaly - is presented in Fig. 2. Class overriding updates the *WhileStatement* and *IfStatement* across the entire replicated *CompilerCollection.* No re-implementation of valid classes is needed.

5 Conclusions

This extensibility problem (Findler, 1999; Flatt 1999) appears when a recursively defined set of data and related operations are to be extended with new data variants or new operations. A set of recursive data and related operations can be straightforwardly represented as a collection of dependent classes. Thus, compiler extensibility can be viewed as a special case of recursive class extensibility. Although extensibility can be achieved through subclassing, it requires extensive use of type casts and cumbersome adaptation code, a necessity that is referred to as the extensibility problem.

The compiler extensibility problem can be avoided by following design patterns that are targeted specially at extensibility, such as the extensible visitor (Krishnamurthi et al., 1998), the generic visitors (Palsberg and Jay, 1997), and the translator pattern (Kühne, 1997). Using such patterns implies serious penalties. In the case of the extensible visitor and the translator patterns, the penalty is the significant programming effort needed for an extension. In the case of the generic visitors, the penalty is the significant run-time overhead imposed by the utilization of reflectivity.

Several known linguistic techniques can be applied to attack the compiler extensibility problem, as for example the extensible datatypes with defaults of Zenger and Odersky (2001) and the evolving open classes of Clifton et al. (2000). None of the known language-level mechanisms seems to offers a silver bullet solution for software evolution. Compared to other approaches, class overriding is simpler and easier to use method to eliminate the subclassing anomaly.

Acknowledgement: This research has been partially supported by a National Science Foundation grant, award number 0243284.

REFERENCES

1. Clifton, C., G. Leavens, C. Chambers, T. Millstein, 2000. MultiJava: Modular Open Classes and Symmetric Multiple Dispatch for Java. OOPSLA'00, Minneapolis, Minnesota, October 2000, ACM Press, New York, 130-145.

   http://www.cs.iastate.edu/~cclifton/multijava/papers/TR00-06.pdf
2. Findler, R., M. Flatt, 1999. Modular Object-Oriented Programming with Units and Mixins. ACM SIGPLAN International Conference on Functional Programming (ICFP '98), 34(1), 94-104. http://www.cs.utah.edu/plt/publications/icfp98-ff/icfp98-ff.pdf
3. Flatt, M., 1999. Programming Languages for Reusable Software Components. PhD thesis, Rice University, Houston, Texas.

   http://cs-tr.cs.rice.edu/Dienst/UI/2.0/Describe/ncstrl.rice_cs/TR99-345/
4. Krishnamurthi, S., M. Felleisen, D. P. Friedman, 1998. Synthesizing Object-Oriented and Functional Design to Promote Reuse. ECOOP'98, Brussels, Belgium, July 1998, Springer, Berlin, 91-113.
5. Kühne, T., 1997. The Translator Pattern - External Functionality with Homomorphic Mappings. In Ege, R., M. Singh, and B. Meyer (Eds.), The 23rd TOOLS conference USA 1997, 48-62.
6. Palsberg, J., C. B. Jay, 1997. The Essence of the Visitor Pattern. Technical Report 05, University of Technology, Sydney, Australia.
7. Radenski, A., 2002. Anomaly-Free Component Adaptation with Class Overriding, Journal of Systems and Software, Elsevier Science (under print).
8. Zenger, M., M. Odersky, 2001. Extensible Algebraic Datatypes with Defaults. International Conference on Functional Programming, ICFP 2001, Firenze, Italy, September, 2001. http://lamp.epfl.ch/~zenger/papers/icfp01.pdf
9. Watt, D., D. Brown, 2000. Programming Language Processors in Java: Compilers and Interpreters, Prentice Hall, New York, New York.

TOWARDS BUILDING BULGARIAN WORDNET: LANGUAGE RESOURCES AND TOOLS[1]

George. Totkov

Plovdiv University, 24 Tsar Asen Str., Plovdiv 4000, Bulgaria totkov@pu.acad.bg

Abstract: WordNet (WN) is an on-line lexical database of English language. Word meanings are represented by synonym sets of words (synsets). Different lexical and semantic relations link the synsets. The basic idea in the development of Bulgarian WN (BWN) is to perform a recording of already available resources and tools that were useful to the development of the monolingual WN. In this work, new tools and language resources have been developed along the specifications and methodology set. The first part of the paper presents the methodology for the selection of lexical resources and tools, the second one– a number of machine dictionaries and tools (REBUS –system for creating regular grammars and corresponding analyzers); BulMorph 2.0 –morphological processor); editor for semi-automatic extraction and forming Bulgarian synsets using WordNet synsets and own lexical resources), the third part examines the experimental tools under development. The applications of the WordNet in the fields of information retrieval, information classification, and natural language processing are very important, so we believe that the development of the BWN will be a long-term project.

Keywords: Natural language processing, WordNet

Introduction

Lexical resources used in natural language processing (NLP) have evolved from handcrafted lexical entries to machine-readable lexical databases and large corpora. Much effort is being applied to the creation of electronic lexicons and electronic linguistic resources in general.

WordNet [4, 9] – a lexical database with semantic relations between English words, was developed in the Cognitive Science Laboratory at the University of Princeton. It can be used both as an on-line dictionary or thesaurus for reference purposes, and as a taxonomic lexical database. The building block of *WordNet* (WN) is a synonym set (synset) of words that expresses a given concept. WN gives definitions (explanatory glosses) and sample sentences for its synsets. The basic semantic WN relations are *hyponymy* (X is a kind of Y), *hypernymy* (this is a kind of X), *meronymy* (part of this X), *holonymy* (this is a part of X), *entailment* for verbs (like *meronymy* for the nouns), *antonymy*, *synonymy*, etc. *WordNet* has been used in various applications including Information Retrieval, Word Sense Disambiguation, Machine Translation, Conceptual Indexing, Text and Document Classification and many others.

The success of *WordNet* has determined the emergence of several projects that aim the construction of WN for other languages than English or to incorporate WN in various NLP applications. Following the initial design of WN, the *EuroWordNet* (EWN) project [30, 31] resulted in a multilingual lexical database for eight European languages.

BalkaNet[2] [22] is an EC funded project (IST-2000-29388) that aims to develop in accordance with EWN philosophy a multilingual lexical database with semantic networks for the following Balkan languages: Bulgarian, Czech, Greek, Romanian, Serbian and Turkish. Each monolingual *WordNet* that has been developed separately will be incorporated in the *BalkaNet* database, which in turn will be linked to EWN, thus resulting in a global semantic database.

The applications of the *WordNet* in the fields of information retrieval, information classification, and natural language processing in general are very important, so we believe that the development of the BWN will be a long-term project.

Methodology

The *Bulgarian WN* (BWN) has been developed as a cooperative task involving the Plovdiv University and the Department for Computer Modelling of Bulgarian Academy of Sciences. Each team performed a recording of already available resources and tools that were useful to the development of the monolingual WN. In addition, new tools and language resources have been developed along with the specifications and methodology set.

The basic idea for development of BWN is to create automatically new linguistic resources and to check and improve the automatically built results. The *methodology* proposed has been realized in *6 steps*:

A. Research and gathering of lexical resources (particularly in English) and methods used by previous projects;
B. Collection, creation and systematization of electronic linguistics resources such as Bulgarian Dictionary of Synonyms (BDS), Explanatory Dictionary of Bulgarian (EDB), Bilingual Bulgarian – English Dictionary (BBED), Word Formation Dictionary of Contemporary Bulgarian (WFDCB), etc.;
C. Compilation, collection and systematization of tools developed in the field of NLP;
D. Evaluation of the linguistics and software resources we have and the options they offer for semi-automatic extraction of BWN;

[1] The work was supported by the 5th Framework programme (Project IST-2000 29388-BalkaNet) and the Plovdiv University Grant M-30/2003.

[2] Further information can be obtained from the project's web site http://dblab.upatras.gr

E. Creation of new software tools for automatic extraction of synsets prototypes and basic concepts and taxonomy (with the use of BDS, BBED and EDB);
F. Design and working out of tools for editing the BWM prototypes, appropriate for lexicographers.

The methodology used is based mainly on the *merge model* [30]: the selection is done in our linguistics resources; synsets and language-internal relations are first developed separately, afterwards the equivalence relations to EWN are generated and researched. The merge approach includes different semi-automatic verifications of coded semantic relations in existing Bulgarian language resources.

Language Resources and Tools

In order to make the process as reliable as possible we have assembled a number of machine dictionaries (synonymous, bilingual, explanatory, derivational, etc.) and tools that the lexicographers and programmers should use in making the decisions necessary for building the core BWN. In the following paragraphs there is a brief account of these tools, described elsewhere:

- Pre-processing tools for Bulgarian texts (part of speech tagger, tokenizer, sentence splitter and paragraph splitter, procedure for clause extraction, noun phrase extraction, procedures for anaphora resolution, procedure for heading identification, acquisition of features of unknown language elements etc.) [25, 26];
- Bulgarian morphological processor (*Bulmorph 1.0*[3]) (morphological analysis and synthesis, robust analysis of unknown words, lemmatizer, containing approx. 69,000 stems, etc.) [23, 24].

Language Resources

One delivery of the *BalkNet* is the large word-form lexicon used by *BulMorph 2.0* (more than 80,000 entries) containing triples [word-form, lemma, morpho-syntactic code]. Another extremely useful lexical resource we relied on is the *Bulgarian Dictionary of Synonyms* (BDS) [11], which was digitized and encoded as an ACCES database. The dictionary consists of 24,699 entries in alphabetical order. The relations of synonymy in the dictionary are presented in more than 36,000 synonymy sets. Every word belongs to 1.4 synsets (on average). Every word has 3.4 word synonyms and 0.2 phrase synonyms (on average).

At present our electronic *English-Bulgarian Dictionary* consists of 58,000 English headwords and corresponding Bulgarian (one or more) translation equivalents and 42,500 headwords for the *Bulgarian-English Dictionary*. The reference dictionary we used is the *Explanatory Dictionary of Bulgarian* (EDB) [1]. This authoritative

[3] The *BulMorph 1.0* is being distributed by ELDA (European Language Distribution Agency) since 1990. The ELDA catalogue is available on address *www.icp.grenet.fr/ELRA/cata/tabtext.html*.

lexicographic source for contemporary Bulgarian was digitized, corrected, edited (using the morphological processor *BulMorph 2.0*) and converted into a lexical database (Access encoded). The EDB (in electronic form) consists of 52,434 Bulgarian headwords and corresponding Bulgarian (one or more) explanatory notes. The derivational relations in the Bulgarian WN will be defined and handled within *Word Formation Dictionary of Contemporary Bulgarian* (WFDCB) [16].

Finally, an extremely valuable resource was the *ILI* of the English *WordNet*, exported in XML format by means of the *VisDic* editor produced by the Masaryk University of Brno [15].

Tools

Some additional tools have been developed and used for the construction of the BWN: *REBUS* – dialogue system for creating of regular grammars and corresponding analyzers [28]; *BulMorph 2.0* – morphological processor and lexical database (more than 80,000 base forms and about 1,750,000 corresponding word forms [29]; tools for *semi-automatic extraction and forming Bulgarian synsets* (using EWN synsets, BBED and BDS) [27].

REBUS

The regular expressions and finite states transducers [6, 10, 13, 17, 21, 32] are at the appropriate level of abstraction for thinking about finite-state languages and finite-state relations. The created Regular Expressions Builder and Parser for Unicode Systems (*REBUS*) is a tool for creating nondeterministic finite automata (NFA). Using *REBUS*, every regular expression can be transferred into a graph with no ε-transitions. The memory needed by *REBUS* Nondeterministic Automata (RNA[4]) can be up to times lower than the one needed by NFA. In RNA-graph the information for the transitions and characters that match every node is actually stored in one place. When using depth-tracking this makes it possible to check all child-nodes that match the next character at once. The ability of the system to work with external modules allows also the parsing of non-regular content.

Using its own engine the performance of parsing and replacing of *REBUS* is better than most of the similar systems. The experiments show that *REBUS*-algorithm really is linear and that the speed is proportional to the lengths of the expressions and the parsed text. What is more, here is a result of *Grep*[5] and *REBUS* on a 2,000,000 words with 2,500 words in the grammar: *Grep* – 13,029 seconds (3h

[4] Formal definition of RNA, including the varieties and definition of computation are expected in the very near feature.

[5] The version of *Grep* (one of the most common programs used for general regular expression parsing for the past years) used is *2.4b* [33].

37m 9s) and *REBUS* – 502 seconds (8m 22s). Unlike *Grep*, *SED*[6] is about two times faster than REBUS. *REBUS* on the other hand does not have a problem with the grammar size when trying to parse a file with a grammar larger than 64 KB (or for example bigger than 10,000 words). In this case *Grep* and *SED* would note “Regular expression too big” or “Regular expression too large”.

BulMorph

The first morphology analysers of Bulgarian texts are the *BulMorph 1.0* [23, 24] and *MorphoAssistant* [14]. Some morphology analyzers are rather slow. For example, *MORPHY* [7] analyzes 300 German word forms[7] per second on a fast PC and the speed achieved by *BulMorph 1.0* is about 900 word forms per second (on 133 MHz processor).

Some other suitable approaches for modelling grammatical dictionaries are based on acyclic finite state automata with labels of finite states and acyclic finite state transducers [5, 18-21]. These representations are used for development of grammatical systems for Polish and Bulgarian languages [2, 8]. In spite of good practical results demonstrated by those systems, we can enumerate some disadvantages of the approaches applied: the used structures do not correspond naturally to the grammatical phenomena modelled; the presentations are not convenient (perhaps it is not possible at all) to obtain any grammatical information for an ‘unknown’ word form; we have to develop a completely different structure (another finite state automaton), if we need to model another kind of analysis or dictionary (for example derivational dictionary), etc. Those are the reasons to propose more compact representations, faster algorithms and enrichment of lexical resources [2, 3, 8, 18, 19].

The morphology analyzer *BulMorph 2.0* is object of similar goals – to achieve significant advancement in overcoming those disadvantages in the case of Bulgarian language. The *BulMorph 2.0* solves the outlined problems, using some modifications of the mentioned models. The model of computation (and more specially the introduced bipartite Finite State Transducers – bPFST[8]) is similar to a standard finite state transducer (FST), but the transitions are managed by two FSTs. bPFSTs are deterministic finite transducers with the property of being able to jump during a transition from one FST to the other. The new solutions in *BulMorph 2.0* are:

- distributing the language information between two FSTs integrated in one bPFST;
- incremental method of the construction of FSTs that combines the process of construction with the process of their minimization[9];
- using the constructed bPFST as analyser, synthesizer, lemmatizer and unknown word guesser simultaneously.

The underlying idea of the *BulMorph 2.0* is based on an economical representation of the lexical corpus by regular expressions called patterns. In essence, this means dividing the representation of the source corpus into two parts – an *invariant*, called Morphological Kernel (MK) and a *variant*, called Morphological Shell (MS). The MK encompasses a part of the corpus that is invariant and remains unchanged for all word forms of the same word paradigm. While the MK includes one and only one entry (the pattern) for each one of the word paradigms, the MS conversely, represents the inflectional phenomena and word-forming properties of the words in the source corpus. As a result, the complete morphological information for initial 1,500,000 word forms in the *BulMorph 2.0* takes less than 1 MB disk space. In comparison, the text representation of the same lexicon requires about 106 MB.

The *BulMorph 2.0* consists of three modules[10]: the morphological *Analyzer* (incl. unknown-word *Guesser*), the *Synthesizer* (module generating the possible word forms) and the *Lemmatizer* (system for determining of the base word form).

The speed achieved is very high – more than *150,000 word forms analysed per second (on 1.6 GHz processor)*. For the evaluation of the robust analysis a corpus (comprising about 11,600 new base forms and the corresponding 226,730 word forms) was used. Rare words, archaic and dialect words or words of foreign origin constitute for the most part the rest of the corpus. That is the reason, why the received grades of the precision (93.5%) and the recall (79.6%) of the robust analysis can be considered to be improved[11].

Automatic Improving and Forming BWN Synsets

The main purpose of this research is to propose methods and tools for semi-automatically mapping of Bulgarian synsets to English WN entries and, therefore, building a BWN. The first attempt to build a core of BWN (only for nouns) using semi-automatic extraction from the original EWN is described in [12].

In the process of building BWN we used the following resources: original EWN dictionary – over 175,000 entries accompanied with its synonyms, glosses and WN identifier number; new English-Bulgarian Dictionary (EBD) – about 165,000 entries and BDS (about 25,000 synonym paradigms).

[6] The version of *SED* (a famous *search-and-replace* program well-known by the Linux world) used is *Super-sed v3.48”* [34]

[7] We must not forget that German is a highly inflectional language.

[8] Formal definition of bPFST, including the varieties and definition of computation are expected in the very near feature.

[9] Similar methods are proposed in [2, 3, 8] in the case of one FST.

[10] The current version has been implemented with Visual C++ 6.0 on Windows XP platform.

[11] In reality the precise and the recall of the morphological analysis on standard Bulgarian texts (with a relatively low percentage of words belonging to the aforementioned types) are not less than 99 % and 95% respectively.

Our first goal was to convert all three dictionaries into compatible formats, where each line corresponds to one entry in the dictionary. All fields in the dictionary that are not needed are stripped, so that less effort is needed to process the dictionary in the next stage.

The process of automated improving and forming BWN based on BDS and EBD may be split into *three steps* that are fairly independent: automatic improvement of BDS; finding a correspondence between rows of EBD and EWN synsets; forming synsets of the BWN.

The *first step* (automatic improvement of synonym dictionary) includes discovering different synsets representing one concept; synsets, which contain words for two or more different concept (mixed synsets); synsets with missed or incorrect synonyms placed in them, etc. In order to locate and remove various types of errors, gaps and discrepancies, a distance between arbitrary two synonym rows is introduced and a special *editor for splitting/merging synonym rows* was created and tested.

The *second step* (finding a correspondence between rows of EBD and EWN synsets) is to match each EWN row with its respective EBD one. With this view, each EWN line is provided with an identification number, which is transferred into a special field in the EBD. The problem is tackled with by the creation of two Access tables – the first one represents EBD and the second one – EWN. Then the EBD rows are joined with the EWN rows so that the joined fields from both tables are the corresponding English words. The so-related tables are processed by professional translator who transfers EWN numbers to the corresponding column in EBD table. In case that a Bulgarian signification is missing, it is added as a new row in EBD with a respective number and translation. The result is the formation of more than 55,000 correspondences between the EWN and Bulgarian entries. The additional effect is the enrichment and updating of the EBD.

The *third step* (forming BWN) in the process of setting up the corresponding BWN requires Bulgarian synonyms to be separated in each row of EBD in the translated Bulgarian equivalent of the given English word. Any such line contains a list of meanings for the target word, separated by commas. Our goal is to design an efficient algorithm that determines which of the commas are separators for the different meanings and which are not. An experimental tool based on a sequence of rules[12] is under development.

Perspectives

The construction of BWN requires the supply of additional lexical resources and the creation of new software tools. The digitalization of the EDB and the BBED is finished and procedures for their automated alignment are being worked out. The Bulgarian glosses will be automatically attributed to their respective synonym lines in the BWN and a completely mirrored Bulgarian-English-Bulgarian dictionary can be created. Synsets could be enlarged automatically relying on the data from the BSD and EDB. If we compare the data from these dictionaries, we can observe some common elements and taxonomic operators in the vocabulary definitions. Very often the synonym is explained by the differences with the main word in the group (the "mother" concept from a higher level of the lexical hierarchy). Searching for these “key words” (usually nouns) we can support building the hierarchy in the Bulgarian WN.

A problem which remains open to this stage is the usage of derivative links between separate words (synsets respectively) in BWN and EWN. From the word formation point of view, various suffixes/affixes tend to alter the word root, but these transformations are not presented in the electronic dictionaries we used. So, in Bulgarian synsets, we omit a lot of adverbial nouns because of incomplete dictionary input. The WFDCB [Penchev’2000] can be used to enlarge the Bulgarian synsets, especially those consisting of nouns derived from verbs, also prefixed verbs, adverbs, adjectives, etc.

[12] The order of the rules does matter, since it is possible that a general rule will override a more specific one.

REFERENCES

1. Andrejchin L. (ed.), *Bulgarian Explanatory Dictionary.* Sofia, Nauka i Izkustvo, 1999 (in Bulgarian).
2. Daciuk J., *Incremental Construction of Finite-state Automata and Transducers and their Use in Natural Language Processing.* Ph.D. thesis, Technical University of Gdansk, 1998.
3. Daciuk J., S. Mihov, B. Watson, R. Watson, *Incremental Construction of Minimal Acyclic Finite State Automata*. Computational linguistics, 26 (10), April 2000, pp. 3-16.
4. Fellbaum C. (ed.), *WordNet: An Electronic Lexical Database.* The MIT Press, Cambridge, London, England, 1998.
5. Karttunen L., *Constructing lexical transducers*. COLING-94, Kyoto, Japan, 1994.
6. Karttunen I., J.-P. Chanod, G. Grefenstette, A. Schiller, *Regular Expressions for Language Engineering*. Journal of Natural Language Engineering, 2:4, 1996, 305-328.
7. Lezius W., R. Rapp, M. Wettler, *A Morphology System and Part of Speech Tagger for German.* in D. Gibbon (ed.), Natural Language Processing and Speech Technology, Results of the 3rd KONVENS Conference, Belfield, October 1996, Mouton de Gruyter, Berlin, 1996.
8. Mihov S., *Direct Building of Minimal Automation for Given List.* Annuaire de l’Université de Sofia “St. Kl. Ohridski”, Faculté de Mathematique et Informatique, vol. 91, livre 1, February, 1998.
9. 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.

10. Mohri M., *Finite-State Transducers in Language and Speech Processing*. Computational Linguistics 23 (2) 1997, 269-312.
11. Nanov L, A. Nanova, *Bulgarian Synonym Dictionary.* Sofia, Hejzal, 2000 (in Bulgarian).
12. Nikolov T., *Building a Core for Nouns for Bulgarian WordNet.* Diploma thesis, Sofia University, 2000 (in Bulgarian).
13. Noord G., D. Gerdemann, *An Extendible Regular Expression Compiler for Finite-state Approaches.* in Natural Language Processing, WIA 1999.
14. Paskaleva E., K. Simov, M. Dimitrova, M. Slavcheva, *The Long Journey from the Core to the Real Size of the Large LDB*. In Acquisition of Lexical Knowledge from Text (eds. Boguraev, Pustejovski), Columbus, Ohio, 1993, 161-169.
15. Pavelek T., Pala K., *VisDic: A new Tool for WordNet Editing*. in Proceedings of the 1st International Wordnet Conference, Mysore, India, 21-25 January 2002, 21-25.
16. Penchev J. (ed.), *Word Formation Dictionary of Contemporary Bulgarian.* Sofia, 2000 (in Bulgarian).
17. Perrin D., *Finite Automata*. In J. Van Leuwen (ed.), Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics, Elsevier, Amsterdam. 1990, 1-57.
18. Revuz D., *Dictionnaires et lexiques: methodes et algorithms*. Ph.D. thesis, Institut Blaise Pascal, Paris, France, 1991.
19. Revus D., *Minimisation of Acyclic Deterministic Automata in linear time*. Theoretical Computer Science, 92, 1, 1992.
20. Revus D., *Dynamic Acyclic Minimal Automation*. 5th Int. Conf. on Implementation and Appl. of Automata, CIAA 2000, London, Canada, pp. 226-232.
21. Roche E., Y. Schabes, *Finite-State Language processing.* Bradford Book series, MIT Press, Cambridge, Massachusetts, USA, 1997.
22. 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.
23. Totkov G., Hr. Kruschkov, M. Ivanova, *Formalisation of Bulgarian Language and the Development of a Linguistic Processor*. Universite de Plovdiv, Travaux scientifiques, Mathematique, 26 (3), 1988, pp. 301–311 (in Bulgarian).
24. Totkov G., *Robust Methods for Analysis of Bulgarian Texts and the Development of a Linguistics Processor.* The 19th Conference of the UBM, 1990, Proceedings ‘Mathematics and Mathematical Education’, Sofia, BAS, 1990, pp. 295–303 (in Bulgarian).
25. Totkov G., *Resources and Tools for Computerization of Bulgarian Language (1988-2000)*. Artificial Intelligence, No.3, 2000, 573-577.
26. Totκov G., Ch. Tanev, *LINGUA – Architecture for Robust Text Processing in Bulgarian*. in A. Narin’iyani (eds.), Comp. Ling. and its Applications, Proc. of the Int. Workshop DIALOGUE’2002, Protvino, 6-11 June 2002, 582-589.

27. 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 (in print).
28. Totkov G., D. Blagoev, R. Dokov, *Regular Expressions Builder and Parser for Unicode Systems.* 1st Balkan Conference in Informatics, Thessalonica, Greece, Nov. 21-23, 2003 (presented).
29. Totkov G., R. Doneva, *Bipartite Finite State Transducers as Morphology Analyser, Synthesizer, Lemmatizer and Unknown-Word Guesser*. 1st Balkan Conference in Informatics, Thessalonica, Greece, Nov. 21-23, 2003 (presented).
30. Vossen P. (ed.), *EuroWordNet General Document*. EuroWordNet (LE2-4003, LE4-8328), Final Document, 1998, 108p.
31. Vossen P. *Building a multilingual database with wordnets for several European languages. http://www.hum.uva.nl/~ewn*, 1999.
32. Watson Br., *Taxonomies and Toolkits of Regular Language Algorithms*. Ph.D. thesis, Eindhoven University of Technology, the Netherlands, 1995.
33. *www.gnu.org/software/grep/grep.html*. Grep’s on-line page.
34. *www.gnu.org/software/sed/sed.html*. Sed’s on-line page.

Papers

A NEW CYBER THREATS HIGH-SPEEDS COMPUTER WORMS ATTACKS

Eugene Nickolov

National Laboratory of Computer Virology - BAS
1113 Sofia, Acad. G. Bonchev, Building 8, Office 104
eugene@nlcv.bas.bg

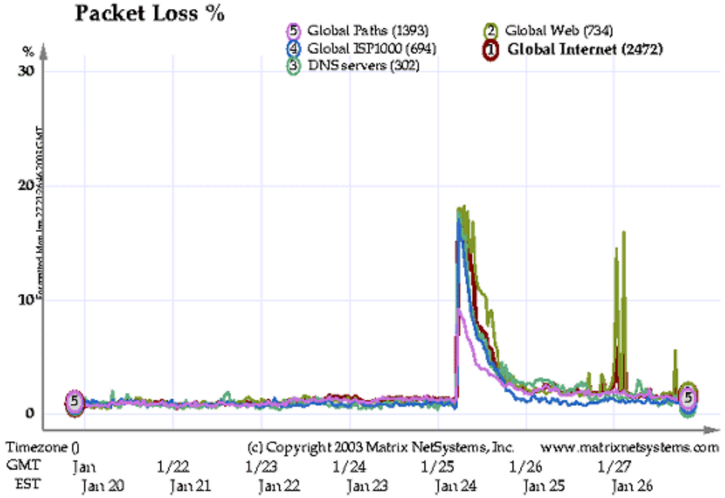
Abstract: The report includes short history, attacked platforms, network overload, known alias, and worm details about Slammer worm. Accent has been put on his evaluation in comparison with high-speed computer worms CodeRed and Nimda. New cyber threats are commented. E-Government solutions are discussed about the strategic objectives and critical priorities for the future National Strategy to Secure Cyberspace on Bulgaria.

Keywords: Cyberspace, Cybersecurity, Cyber attacks, Cyber Incidents, Viruses, Worms, Trojans, Back Doors, Threats, and Vulnerabilities.

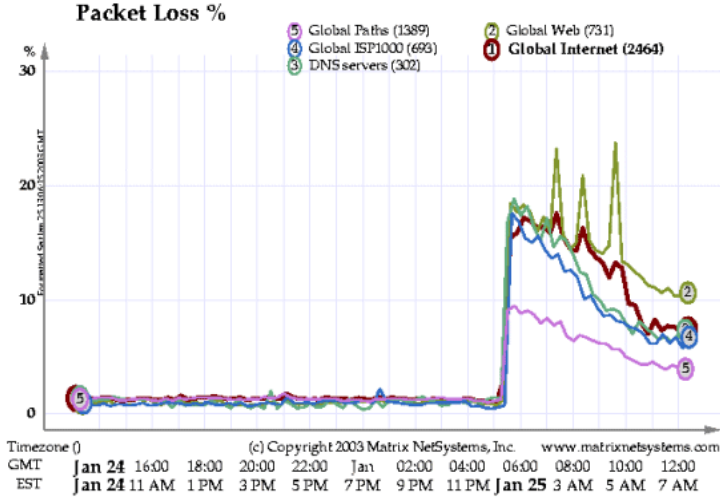
Short History. Although its traces were found already on 20/01/2003, the outbreak of the worm was detected in the Internet on 25/01/2003 (05:30 GMT). It spreaded out very quickly (10 min) and in 4 hours has been detected in most countries around the world. The worm generates massive amounts of network packets, overloading servers and routers and slowing down network traffic. The next day, (26/01/03) five of the thirteen Internet root name servers were down.

Attacked Platforms. This worm does not infect typical end user machines at all. End users might only notice this worm because of network slugginess. Slammer only infects computers running MS SQL Server 2000. However, the SQL server is embedded inside MS Data Engine 2000. In addition, there are lot of other applications that might silently install MS SQL Server or MSDE 2000. Examples of such software are: MS Biztalk Server, MS Office XP Developer Edition, MS Project, MS SharePoint Portal Server, MS Visio 2000, MS Visual FoxPro, MS Visual Studio.NET, MS .NET Framework SDK, Compaq Insight Manager, Crystal Reports Enterprise, Dell OpenManage, HP Openview Internet Services Monitor, McAfee Centralized Virus Admin, McAfee Epolicy Orchestrator, etc.

Network overload. The 3 load peaks (1 / Sat and 2 / Mon) can be seen in this graph from Matrix NetSystems taken on Monday the 27th.



The effects of the network overload can clearly be seen in this graph from Matrix NetSystems taken on Saturday the 25th.



Known Alias. The worm *Slammer* is known also as *Sapphire, W32.Slammer, SQLSlammer, W32.SQLExp.Worm, W32/SQLSlam-A, W32/SQLSlammer.worm, WORM_SQLP1434.A, Worm.SQL.Helkern*.

Worm Details. The worm code is 376 bytes, written using the Assembly language. It is not a mass mailer: it does not send any e-mail. The worm only spreads as an in-memory process: it never writes itself to the hard drive. It exists only as network packets and in running processes on the infected computers. In this respect Slammer is similar to the worm CodeRed detected in July 2001. Slammer uses UDP port 1434 to exploit buffer overflow vulnerability in MS SQL Server 2000 (MS02-039). When the SQL server receives the malicious request, the overrun in the server's buffer allows the worm code to be executed. After the worm has entered the vulnerable system, first it gets the addresses to certain system functions and starts an infinite loop to scan for other vulnerable hosts on the Internet. Slammer uses GetTickCount() function from the Win32 API to initialise its random number generator. It uses the random numbers as IP addresses to search for vulnerable hosts. The worm does not try IP addresses sequentially - it gets a new random address for every infection attempt. The worm uses a socket connected to the UDP port 1434 on the remote machine to send itself to the vulnerable computers. If the computer has MS SQL Server 2000 it runs on that port. Sometimes the random generator returns numbers that are broadcast addresses (x.y.z.0 or x.y.z.255) causing all the hosts on the particular network to receive the malicious packet. In many scenarios, this could infect several SQL Servers in the subnet at the same time. This makes the spreading routine even more aggressive. At the very least hitting broadcast addresses multiplies the network load generated by the worm. Since the worm code does not have any delay in the scanning loop it generates massive amount of network traffic as a side effect. Slammer does not have any intentional payload or strings inside. Since the worm does not reach the disk on the infected computer it disappears when the server is restarted. On the other hand the server might get over again infected if the security problem is not fixed. The worm might have been started off with a hitlist of a thousand or so vulnerable machines (along the lines of Warhol worms) to make initial spreading faster.

CodeRed Details. The Code Red worm is malicious self-propagating code using a security hole ("Unchecked Buffer in the Index Server ISAPI Extension") in MS Internet Information Server (IIS) to spread and does not pose a threat to end users. The worm infected more than 250,000 systems in just 9 hours. Analysis estimates that starting with a single infected host, the time required to infect all vulnerable IIS servers with this worm could be less than 18 hours. More than 2,000 new hosts were infected each minute. 43% of all infected hosts were in the USA, while 11% originated in Korea followed by 5% in China and 4% in Taiwan. The .NET Top Level Domain (TLD) accounted for 19% of all compromised machines, followed by .COM with 14% and .EDU with 2%. We also observed 136 (0.04%) .MIL and 213 (0.05%) .GOV hosts infected by the worm. Its activity on a compromised machine is time sensitive; different activity (4 phases) occurs based on the day of the month of the system clock. *Infection phase*: The infected host will attempt to connect to TCP port 80 of randomly chosen IP addresses in order to further propagate the worm. Infected systems may experience web site defacement as well as performance degradation as a result of the propagating activity of this worm. This degradation can become quite severe and may cause some services to stop entirely. The worm also drops a trojan program to '\explorer.exe' that modifies different some IIS settings. It disables the System File Checker (SFC) functionality in Windows which is responsible for checking the integrity of system files. *Flood mode*: A packet-flooding denial-of-service attack will be launched against a specific IP address embedded in the code. *Termination* (after the 27th day): The worm remains in memory but is otherwise inactive. There are three variants of Code Red known: CRv1 - using a specially crafted string sent to HTTP servers over the Internet, the worm manages to overwrite a variable in the module named "idq.dll" thus, forcing the system to jump to an incorrect address, executing the worm code; CRv2 - a second, fixed variant appeared in the wild few days after the first variant of the worm. It shared almost all of its code with the first version, but spread much more rapidly; CodeRedII - which installs a backdoor into systems it infects. As a result, any web surfer can now execute commands on any infected www site just by typing suitable URLs to the web location.

Nimda Details. Nimda is a complex virus with a mass mailing worm component, which affects Windows 95/98/Me/NT4/2000 users. This is the first worm that modifies existing web sites to start offering infected files for download. Also it is the first worm to use normal end user machines to scan for vulnerable web sites. This technique enables Nimda to easily reach intranet web sites located behind firewalls - something worms such as Code Red couldn't directly do. Nimda uses the Unicode and the MIME exploits to infect IIS web servers by four methods: 1) *File infection*. Nimda locates EXE files from the local machine and infects them by embedding each victim file inside a worm copy. This new file replaces then the originally clean file and when executed, the worm will extract the original clean file to a temporary file and execute it along with itself; 2) *Mass mailer*. Nimda contains a mass-mailing routine, which is executed every 10 days. The worm locates e-mail addresses via MAPI from the e-mail client as well as by searching local HTML files for additional addresses. The worm uses its own SMTP server and the configured DNS entry to obtain a mail server record (MX record). The mails contain an attachment called README.EXE, which might be executed automatically on some systems; 3) *Webworm*. Nimda starts to scan the Internet, trying to locate www servers using randomly generated IP addresses. Once a web server is found, the worm tries to infect it by using several known security holes. If this succeeds, the worm will modify random web pages on the site. End result of this modification is that the worm will automatically infect web surfers browsing the site. The JavaScript causes visitors who open infected pages to be presented with Readme.eml, which was created by the worm. Thus, simply browsing the infected Web page may infect a computer; 4) *LAN propagation*. The worm uses backdoors on IIS servers such as the one CodeRedII installs. It scans random IP addresses for these backdoors. When a host is found to have one the worm instructs the machine to download the worm code from the host used for scanning. After this it executes the worm on the target machine this way infecting it. The worm changes Windows Explorer settings to not show hidden files and known file extensions. After that the worm adds a 'guest' account to infected system account list, activates this account, adds it to 'Administrator' and 'Guests' groups and shares C:\ drive with full access privileges.

World cyber threats. The threats have many parameters but the speed of the attacks (less then 15 min for the Web), the number of the stricken servers (more then 100 000), the dropping out of the almost the half of the world Internet Root Name Servers are sufficient cause to realize the dimension of the danger. The dramatic increase of discovered and used vulnerabilities for the period 2000 – 2002 (from 1090 to 4129) and the growth of the cyber incidents for the period 1999 – 2002 (from 10 000 to almost 90 000) show that the losses are immense but its difficult to mention reliable evaluations.

NATIONAL STRATEGY TO SECURE CYBERSPACE ON BULGARIA.

STRATEGIC OBJECTIVES. Prevent cyber attacks. Reduce vulnerability to cyber attacks. Minimize damage and recovery time from cyber attacks. *CRITICAL PRIORITIES.* Response system. Threat and vulnerability reduction program. Awareness and training program. Securing Governments' cyberspace. International cyberspace security cooperation. *NATIONAL LEVEL TASKS.* Comprehensive national plan for securing the key resources and critical infrastructure. Crisis management in response to attacks on critical information systems. Technical assistance to emergency recovery plans for failures of critical information systems. Provide specific warning information and advice about protective measures and countermeasures. Performing and funding research and development to new scientific understanding and technologies.

Conclusion

There is a pressing need to create National Strategy to Secure Cyberspace in Bulgaria formulating Strategic Objectives, Critical Priorities and National Level Tasks.

Acknowledgments: In the report are used many information sources from some world famous web sites like www.f-prot.com, www.f-secure.com, www.command.com, www.symantec.com, www.nai.com, www.mcafee.com, www.kaspersky.com, www.cert.org, www.sans.org, www.whitehouse.gov, www.dhs.gov, etc. I would like to express my acknowledgments to all persons, which have the responsibility for creating and supporting these sites.

PLANNING TECHNOLOGIES FOR THE WEB ENVIRONMENT: PERSPECTIVES AND RESEARCH ISSUES

A. Milani, S. Marcugini

Department of Mathematics and Informatics,
University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy;
e-mail: milani@unipg.it, gino@dipmat.unipg.it

Abstract: This work will explore and motivate perspectives and research issues related with the applications of automated planning technologies in order to support innovative web applications. The target for the technology transfer, i.e. the web, and, in a broader sense, the new Information Tecnologies (IT) is one of the most changing, evolving and hottest areas of current computer science. Nevertheless many sub-area in this field could have potential benefits from Planning and Scheduling (P&S) technologies, and, in some cases, technology transfer has already started. This paper will consider and explore a set of topics, guidelines and objectives in orde to implement the technology transfer an new challenges, requirements and research issues for planning which emerge from the web and IT industry.

Sample scenarios will be depicted to clarify the potential applications and limits of current planning technology. Finally we will point out some new P&S research challenge issues which are required to meet more advanced applicative goals.

Keywords: Planning, Web, IT, technology transfer

Introduction

The Information Society (IS) is announced by a set of interrelated emerging technologies where the web it is certainly one of the most apparent and popular elements. These technologies envision new relationships of the individuals between his/her own tasks and the new tools.

Individuals are forced to develop new methods of work in order to exploit the ITs at their best, new tools and application should reflect and model this new methods in order to be effective.

Despite of the successful buzzword "web" (and popular e-*something* terms such as: "e-commerce", "e-business" etc.), it is important to focus on a wider vision of the potential role of planning and scheduling technologies (P&S) in the Information Society not limited to the web applications. An exponential number of internet based services, low cost mobile devices over GSM and UMTS networks, tools which integrate traditional and knowledge based systems, represent promising application fields where P&S can have a primary or a supporting role.

P&S is a research area which dates back to [Fikes1971], scientific and technological results are due to a wide community of researchers and scientific programs on the topic.

P&S can convey the flexibility typical of AI technologies in order to answer to expectations and requirements of new methods of work and for the exploitation of the new I

There are some important elements which facilitate the application of P&S technologies to the web:

- *Digital content*, the web is machine readable and it provides a large quantity of machine readable data and software entities to interact with; automatic knowledge acquisition can be potentially realised in the web environment;
- *Virtual is real*, the web offers the chance having "real" planning application bypassing the complex robot machinery needed in traditional planning application fields such as aerospace or robotics, for example the actions "browse a list of available books" and "buy-a-book" can be easily executed in the digitalised world thus producing a real and useful effect;
- *Scalable complexity*, web applications can have different degree of complexity, simple applications can exists for simple existing planning models. As an example it is worth citing machine translation as a case of a technology which is not mature (i.e. for literature translation) but it could be worthwhile in niches applications (i.e. web pages automated translation)

Planning and scheduling technologies represent a key factor in the framework of the service technologies for the global new IS especially with respect to the current scenario of web applications, they provide:

- *Knowledge based flexibility*, P&S convey the flexibility typical of knowledge based technologies in support of new modalities of work, production, commerce, entertainment, education etc.;*Models of change and interaction*, P&S provides models of change and interaction, this research area has developed systems which provide automated generation of plans, tasks monitoring, plans checking etc.

Another key factor which motivates the application of P&S technologies to the web is its dimension: **web is** *large*. In other words, a massive audience of personal end-users and business end-users is developing the need of customised versions of services, and consequenly the need for web industry of tools and technologies to support them. This scenario is made more complex by the increasing diffusion of personal mobile devices which fosters the development of new modalities of work and interaction between the individuals and the business organisations, and it tends also to modify the traditional modality of B2B interaction.

For this purpose the relevant issues to be supported and managed by services and applications over the information infrastructure are in general: *autonomy, adaptation, distribution, mobility, agent interaction, automatic collaborative support* etc.; which are also typical issues deeply related with P&S research.

Another important element has been the proposal of Semantic Web [Berners-Lee 2001], which has the ambitious goals of allow reasoning on a web of knowledge and meanings, with respect to the initial web made of presentation tags (i.e. HTML) and a web of syntactic structures and terms (i.e. XML). The data on the semantic web are defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. Semantic Web is of great important in the future of knowledge bases applications for the web, because of its success among the research community, the growing number of available tools and application and the standardisation factor, since it is supported and coordinated by W3C Consortium. Despite of the initial focus on ontologies and relationships, research started under the Darpha Agent Markup Language (the DAML program [DAML 2000] has lead to models of more dynamical aspects of the web. For example, DAML-S is oriented DAML-based Web Service Ontologies in order to facilitate the automation of Web service tasks including automated Web service discovery, execution, interoperation. It is worth mentioning the DAML-PDDL translators [PDDL2.1][DAML-PDDL] [McIlrait 2002] which apply P&S to the semantic web.

Defining the Scenario: The Web as an Environment

In this view the *web is a planning environment*,i.e. web entities are the object of P&S systems.

Web entities exists (a simple *web planning domain* is made of elements such as web pages, emails, files etc.) on which typical *actions* can be executed by users (e.g. pay, subscribe, supply, order, browse, look for, find, download etc.) or provided by service systems (e.g. web page servers, search engines, mail servers, messenger servers etc.). Moreover user and systems are pursuing *web tasks*, i.e. tasks which involve web entities (e.g. find a book, buy it and download), and in general user transactions and user activities over the web (e.g. informative tasks, educational tasks, distance work tasks etc.). Sometimes these tasks are not given automatic support, i.e. they rely on the user decisions, or, if automatic support is given, then it uses a too rigid and procedural approach, which does not satisfy the wide variety of user/business needs.

Planners can develop plans to act on the web virtual world in order to reach goals on *web entities*, a key points is that web entities do not necessarily have a "real" counterpart, and they are not necessarily designed as a part of a single distributed system.

Consider for example the following completely *virtual* plan for the *virtual* goal of promoting a web site: *buy disk space, transfer your old site to a popular web portal in order to obtain a better click rate*. The synthesis of this plan would require the ability of describing a model of change (e.g. defining the meaning of acting on the web), goals (defining concepts such "a better visibility"), and a model of the actions to be taken to execute it (i.e. pay, subscribe, download etc.), i.e. web entities should be represented as part as part of a planning domain.

In addition there are some traditional planning phases that needs to be reconsidered in the web domain:

- *domain knowledge acquisition* can be realised by activities of information gathering, information discovery and comparison (about existing portals, rates and prices), with respect to more traditional planning domain in which domain knowledge has not frequent variations;
- *action execution* in the web could imply to take into account of a dynamical scenario, in which not only actions can have unpredictable failure, but the domain can change during execution.

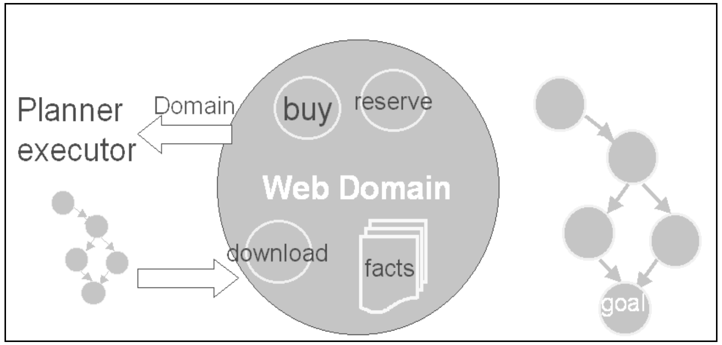


Fig.2 The Web as a Planning Environment

A more general scenario is characterised by activities which take place only *partially* on (or through) the web.

A simple example is that of a personal planner assistant, which suggests the user to buy a textbook online as *part of* the plan "successfully preparing an exam", which contains some other *not web* steps such as *going to lessons* and *doing exercises* . User tasks and goals are in general related with real world activities and should interact and be coordinated with actions and plans which act in the web domain.

When more production related activities become available on the web (for example: *suppliers chains, customers, markets, delivery, payments etc.*) the manufactures planning activity will need to model the web as part of the production plans.

A Glossary of Terms for Mapping Planning on the Web

In this section it will be shown how the basic features of planning domain model can be mapped into corresponding web entities, and contribute to solve and support web tasks and goal oriented activities on the web. On the other hand the mapping allows to point out the limits of the current planning models and the research issues which are required to be solved in order to give a complete account of operating on the web domain.

Glossary Terms: State, Initial State, Goal State.

The notion of state is central to most planning models. In the web the concept of state is represented by the sum of the states of the various component of the web domain:

- *user internal state*, the set of features and facts describing the user state (e.g. maintenance conditions, general constraints, preferences, but also info about users such as identities and passwords);
- *web-state facts*, the state of info as available on the web, in the simplest form web-state is represented by the set of current web pages, existing files etc., more likely the state description would address the content and the semantic of available information (e.g. stock quotations, available item on e-markets etc.)
- *web actors state,* this is represented by the internal state of interactive services, consider, for example, activities which require multiple steps in order be carried out, such as online reservation of a flight (it is necessary to be able to represent the current step of a given transaction), or consider services which require access authorisation(it is necessary to represent that certain existing resources are/are not available for use).

The concept of *state,* that is used in planning to model initial states, goals state, and to model change as modification of state, will need reformulation in the framework of web applications, because of the inadequacy of current planning model to fully describe the state of a web domain in all its aspects.

Research Issue: the main drawback of the current planning models of states derive from the fact that the web is a *vast and dynamical environment of active resources*, on the other hand planning domains are usually characterised by domain states which are fully knowledgeable and somewhat static, i.e. the planner is the only one (or one of the few) agent in the domain.

Summarising, the main P&S research issues that should be addressed in order to fully capture the main features of a *web state* are:

- *managing uncompletness*, web state is vast and not completely knowledgeable, the web domain is inherently incomplete there is no hope to model inside the planner *a complete* description of all the web, then mechanisms and strategies are required in order to do knowledge acquisition and to *circumscribe* the domain description for the planning problem at hand ;

- *managing events*, web state changes independently of the planning actor, this is usually addressed in planning models by the concept of *events*, i.e. state changes which are out of the control of the planner [Ghallab 1998]; in the web domain the amount of *planning time* appear to be a crucial factor, since dynamical changes can occur during planning;
- *managing inconsinstency*, web state can contain contradictory information coming from different sources, consider for example info about the weather conditions used by a travel planner; concepts such as trust and believes and should be modelled [Ambite 1998];
- *managing richer knowledge representation,* P&S models usually focus more on actions and plans than on the adequacy of the knowledge representation formalism to capture the *static* aspect of the domain; since the web offers information in different but equivalent forms, the planners should be aware of ontologies and mechanisms for relating information from different sources, consider for example the problem of representing the concept of *price*, in the sense of amount of money to be payed for buying a good, an effective representation requires that *stock exchange market quotations*, *monthly renting rates*, *price of a book* and *currency exchange rates* would represent instances of the same concept of price [Berners-Lee 2001],

Glossary Terms: Actions and Operators

The other fundamental concept in planning models is the notion of *action* and *operators* , which represents the basic elements to model change in the domain.

Actions (i.e. istances of *operators*) represent the state transiction which occurs during the execution. Operator and action are usually modelled in term of precondition/effects.

In a web domain actions are represented by the available services, which change the user/web state according to a preconditions/effects model. For example available services on the web such as *buying a book, reserving a meeting room, downloading a satellite image of a town, moving a remote webcam* , *sending an email*, are actions in the sense that are allowed to take place only under certain preconditions on the web state (e.g. *having enough money on the account, availability of the tow sat picture etc.*) and produce effects on it (e.g. *changing book ownership, meeting room reservation state, a copy of the picture locally available etc.*).

Research Issues: it is worth noticing that an effective model of *web actions* should take into account the elements formerly pointed out about web states (uncompletness, inconsistency in particular), but it requires, in addition, to consider some *web specific* P&S research issues about the actions/operators model:

- *operators can change, appear and disappear*, new services become available and changes over time, issue: dynamical discovery, monitoring and maintaining of domain models;
- *actions execution time,* actions on the web take time (depending on bandwith and servers overhead factors) , issue: timing constraints and failure recovery;

Glossary Term: Plan

“A *plan* is a set of ordered actions, that, if executed in the initial state will transform it in the goal state”. This definition of solution plan is common to most planning models since [Fikes 1971], and it is easy to see that the definition easily applies also to the web domain. Currently most web tasks and goal oriented activities which take place on the web can be described basically as sequences of actions, i.e. on the web, *plans are sequences of interrelated services requests made on web entities*, (consider for example a typical user driven web-plan such as*: look for items sellers on search engine, browse and compare prices, order one and pay for it*).

Research Issues: although plan as sequences or partially ordered sequences are a formalism that is sufficient for planning in the web at a basic level, it is worth to investigate on planning models which

- *provide expressive and flexible models of plans*, few planning models have a satisfactory management of actions with duration, loops, conditionals [Lin 1995];
- *combine plan knowledge with task oriented languages*, planning models based on a hierarchical approach [Erol 1995], as well as of workflow management models [WorkFlow-Roadmap 2003] offer examples of task oriented formalisms which should be included into generative planning models.

The objective should be to reflect also in the structure of plans, some typical element of the web domain, we have already pointed out, such as contingency, non-determinism, and dynamical aspects of the domain.

Glossary Terms: Plan Monitoring, Execution, Sensing, Re-planning

Planning terms such as *plan monitoring, execution, sensing* and *re-planning* regard the phase that follow plan synthesis and aimed at actually reaching the goals by executing the plan in the real world. In the web domain (where *virtual and real* are somewhat overlapping) can characterised in a “web specific” way traditional planning concepts like *execution* and *sensing*:

•*Execution*, since the web is *large*, execution can involve complex decisions about choosing the service to invoke among a plurality of available and equivalent ones; moreover *execution takes time* also for processing and data transfer, then criteria are needed to combine the issue of execution time with the issue of bandwidth; finally note as the implementation of *actuators* would be greatly favoured by the diffusion of web services, but the use of *wrappers* or similar mechanism should also be considered as an intermediate solution for interacting with services initially designed for human users;

•*Sensing*, in the web environment sensing mainly means: *actively* looking for info, i.e. gathering info [Knoblock 1995] [Golden 1996b] [Naveen 1997] and results from web sources (for example *web pages, searching databases, streaming video, results from called services*)*;* main open problems in this area are related with the integration of information from different non-homogeneous sources, and with decisions to be taken about *what to sense* (consider for example criteria like: *time for sensing and processing vs bandwidth vs timing goals and execution constraints*).

It is also worth noticing that the dynamical nature of the web suggests the investigation of theoretical models [Haigh1996] [Friedman 1997], where the plan synthesis phase is interleaved with execution and sensing.

Conclusion

It is not easy to envision the future of technology transfer in an area where the target industry, i.e. the web and the broader IT society, is moving and developing at an high unprecedented pace toward hard to predict directions. For these reason, previsions and objectives in the long term are not very realistic, and should be limited only to short term and medium term scales.

On the basis of the current achievements and potentiality of P&S technologies depicted in the previous sections some conclusive guidelines for future perspectives and goals can be drawn along two main dimensions: End-user requirements and Research goals.

The first examples of the technology transfer between P&S and the Web have been in the area of:

- automation and autonomy for machine to machine services, (P&S, softbot and agents technologies Etzioni 1994]); task support, Travel Assistant [Ambite 2001], i.e. P&S based dynamic integration of web sources for travel organisation (managing and combining multiple sources for airfares, parking fares, timetables, weather conditions etc);
- automatic maintenance of web applications, web info and web services [Smith 2001] [McIlraith 2001] [Marcugini 2002][Thakkar 2002]; (the P&S knowledge based representation of change, constraint checking and failure repairing techniques has been applied to support automatic maintenance in not web domains [Chien, 1998]).

End-user requirements,

End-users are intended to be in most cases personal users, but in the following we also means by end-user a company, an organisation as a whole, (e.g. in organisational workflow management) or an industry using the technology as a component of more complex systems.

The application issues that are expected to be asreached in a short term are mainly in the area of

- support/automation of personalised services and user adaptive services on the web (educational, e-learning, recreation, information; automatic synthesis/monitoring of online courses and educational plans based on the personal skills/advancements; automatic synthesis of personalized newspaper based on user models);
- supporting mobility and distribution in organisation: scheduling activities & information workflow in distributed organisation through the web and new devices/communication media, supply chain management;supporting user web tasks, (e.g. cooperative recognition, automation, monitoring and repairing of user web tasks, consider, for example, a planning based systems which support the user in task such as “organize a vacation” or “organize a meeting”);
- automatic integration of Services/Web Services, goals directed synthesis and maintenance in specific domains.

There are potential applications of P&S as components for the new IT systems:

- *components for web servers and tools*, (e.g. applications to goal guided synthesis and maintenance of web sites/pages);
- *components for supporting online services*, (e.g. scheduling load distribution to online assistance operators, activation of assistance/emergency chains, online configuration of products);
- *component in systems for online supply chains management*, (issues: logistics, resources management, workflow management, tasks scheduling and assignment).

On a medium/long term horizon, we expect increasingly to met higher requirements on tools and techniques for automation of Web Services, and increasing requests for personal support services, either on the web or onboard of personal assistant devices, such as:

- robust automatic integration of Web Services, in broader application domains, support for automatic discovery of services and failure recovery;
- autonomous web agents, goals directed software agents which operate over the web on user’s behalf (e.g. auctioning systems, stockbot etc.);
- support to massive adaptative services on small personal devices, planning and scheduling support for daily activities (e.g. home devices);
- cooperative and distributed planning which integrates personal devices/web

services/web agents for global user goals;*online P&S services*, general or specialized tools which plan or schedule on demand (e.g. for checking/rescheduling activities) and are available on the web.

*Research goals*The following research goals are prerequisites for supporting the initial and next steps of the technology transfer, on the other hand it seems to exist an increasing autonomous interest of the research community on topics such as modelling dynamic domains, time management, interleaving planning/execution and others, which can be exploited for modeling the web environment, the main topics related to this purpose include:

- modeling domain discovery during planning
- contingency and sensing in web environment
- time duration, failure recognition, and repair
- expressive and robust model of execution plans, (loop iterations)
- mixed initiative planning models, task support models
- incorporating web oriented extension in PDDL-like language (see [KE-Roadmaps] and [WorkFlow-Roadmap])
- wrappers and planning operator wrappers

In a second phase advancements in the previous topics are needed, as well as an increasing the research efforts toward more technology oriented topics such as *portability*, *interoperability, scalability* and advanced planning models, these objectives include:

- portability for planning algorithms, strategies, heuristics, preprocessing;
- interoperability: mapping between planning models;
- models and measures of planning scalability (small devices/ small domains);
- models for cooperative distributed planning, (domains/problem partitioning/integration).

It is generally expected that knowledge based technologies (the semantic Web proposal for instance) can give effective contribution to web applications. We have pointed out that planning technologies, in particular, can positively exploit on the web their ability of modelling action and changes. We also point out the potential scalability of the technology transfer of P&S to the web in a short/medium term perspective from simple applications which use ready-off-the-shelf planning technologies to more complex applications which need appropriate innovative models. On the other hand web and IT applications represent a source of interesting open research issues such as mananging of uncomplete and dynamical domains; interoperability among domains, planners and planning tools.

BIBLIOGRAPHY

[Ambite 2001] J.L. Ambite, *Heracles: Building Planning and Information Assistants* , Proceedings of the ECP/PLANET Workshop on Automated P&S in New Methods of Electronic, Mobile, and Collaborative Work, Toledo, Spain, September 2001,

[Berners-Lee 2001] Berners-Lee T., Hendler J., Lassila O., The Semantic Web*, Scientific American, May 2001*

[Chien,1998] S. Chien, F. Fisher, H. Mortensen, E. Lo, R. Greeley, A. Govindjee, T. Estlin, X. Wang, *Using Artificial Intelligence Planning Techniques to Automatically Reconfigure Software Modules*, Artificial Intelligence and Knowledge-based Systems, Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998

[DAML 2000] Burke M. manager, *The Darpha Agent Markup Language Project DAML* , www.daml.org

[DAML-S 2002]The DAML Services Coalition (Anupriya Ankolenkar et al.), *DAML-S: Web Service Description for the Semantic Web*, The First International Semantic Web Conference (ISWC), Sardinia (Italy), June, 2002.

[DAML-PDDL] Mc Dermott D., Dou D., Qi P., *PDDAML:An Automatic Translator Between PDDL&DAML* http://www.cs.yale.edu/homes/dvm/daml/pddl_daml_translator1.html

[Erol 1995] K. Erol, D. Nau, J. Hendler, *HTN Planning: Complexity and Expressivity,* in AAAI-94, Seattle, July, 1994

[Etzioni 1994] Etzioni O., Weld D. *A Softbot-Based Interface to the Internet*, Communications of the ACM, 1994

[Fikes 1971] Fikes, R. and Nillson, N., *STRIPS: A new approach to the Application of Theorem Proving to Problem Solving*, Artificial Intelligence 2, 1971, pp. 189-208.

[Friedman 1997] Friedman M., Weld D.S., *Efficiently Executing Information Gathering Plans* Proceeding of International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 1997

[Ghallab 1998] Ghallab M. *Chronicles as a practical representation for dealing with time, events and actions*, Proceeding of the 6th Italian Conference on Artificial Intelligence (AIIA'98), Padova (Italy), Lecture Notes on AI, pp.6-10, September 1998

[Golden 1996a] Golden K., Etzioni O., Weld D.S., *Planning with Execution and Incomplete Information*, Technical Report, University of Washington, n.UW-CSE-96-01-09, 1996

[Golden 1996b] Golden K., *Leap before you look: information gathering in the PUCCINI Planner*, Proceedings of AIPS 1998

[Haigh 1996] Haigh K.Z., Veloso M., *Interleaving Planning and Robot Execution for Asynchronous User Requests,* AAAI Spring Symposium on Planning with Incomplete Information for Robot Problems, 1996

[KE-RoadMap 2003] T.L. McCluskey editor, *Knowledge Engineering for Planning Roadma*, PLANET-II Technical Coordination Unit on Knowledge Engineering, March 2003, available on line at http://www.planet-noe.org/TCUs

[Knoblock 1995] Knoblock G.A., *Planning Executing, Sensing and Replanning for Information Gathering*, Proceeding of the International Joint Conference onAI, IJCAI 1995

[Laithwaite 2001] Laithwaite B., *BritishTelecom Workforce Management System*, Proceedings of the ECP/PLANET Workshop on Automated P&S in New Methods of Electronic, Mobile, and Collaborative Work, Toledo, Spain, September 2001,

[Lin 1995] Lin, S.-H., Dean, T. (1995). *Generating optimal policies for high-level plans with conditional branches and loops*. In Proceedings of the Third European Workshop on Planning, pp. 205--218.

[Marcugini 2002] Marcugini S., Milani A., *Automated Planning of Web Services,* Proceedings of the 3rd International Conference on Electronic Commerce, ICEC 2002, Hong Kong, October 2002

[McDermott 2002] McDermott D., *Estimated-Regression Planning for Interactions with Web Services*, Proceedings of the AI Planning Systems Conference (AIPS'02), June 2002.

[McIlraith 2001] McIlraith, S., Son, T.C. and Zeng, H., *Semantic Web Services* , IEEE Intelligent Systems. Special Issue on the Semantic Web. 16(2):46--53, March/April, 2001

[McIlrait 2002] McIlrait S., Fadel R., *Planning with Complex Actions,* Proceedings of AIPS 2002 Workshop on Exploring Real World Planning, Tolouse, France, April 2002

[Naveen 1997] Naveen A., Knoblock G. A., Levy A.Y., *Information Gathering Plans with Sensing Actions,* in Recent Advances in AI Planning: Proceedings of European Conference on Planning ECP97, Springer-Verlag, New York, 1997

[PDDL2.1] Fox M., Long D. editors, *PDDL 2.1: Planning Domain Description Language* http://www.dur.ac.uk/d.p.long/IPC/pddl.html

[Sahuguet 1999] Sahuguet A., Azavant.F. *WysiWyg Web Wrapper Factory*. In WWW8, 1999.

[Smith 2001] Stephen F. Smith S.F., Hildum D., Crim D. *Toward the Design of Web-based Planning and Scheduling Services*, Proceedings of the ECP/PLANET Workshop on Automated P&S in New Methods of Electronic, Mobile, and Collaborative Work, Toledo, Spain, September 2001,

[Thakkar 2002] Thakkar S., Knoblock G.A., Ambite J.L., and Shahabi C., *Dynamically Composing Web Services from On-line Source,* Workshop on Intelligent Service Integration, AAAI02, Edmonton, Alberta, Canada, 2002 , *http://www.isi.edu/info-agents/dotnet/aaaiworkshop2002.pdf*

[WebTCU] A.Milani editor, *Website of the PLANET-II Technical Coordination Unit on Planning and Scheduling for the Web*, http://www.planet-noe.org/TCUs

[WorkFlow-Roadmap 2003] D. Borrajo Editor, *Planet Workflow Management R&D Roadmap*,

PLANET-II Technical Coordination Unit on Workflow Management, February 2003, available on line at http://www.planet-noe.org/TCUs

TELEHEALTH APPROACH FOR GLAUCOMA PROGRESSION MONITORING

Mihaela Ulieru and Alexander Grabelkovsky

Department of Electrical and Computer Eng., University of Calgary
Calgary, Alberta, Canada T2N 1N4
e-mail: ulieru@enel.ucalgary.ca

Abstract: This paper presents a Web-Centric [3] extension to a previously developed glaucoma expert system that will provide access for doctors and patients from any part of the world. Once implemented, this telehealth solution will publish the services of the Glaucoma Expert System on the World Wide Web, allowing patients and doctors to interact with it from their own homes. This web-extension will also allow the expert system itself to be proactive and to send diagnosis alerts to the registered user or doctor and the patient, informing each one of any emergencies, therefore allowing them to take immediate actions.

The existing Glaucoma Expert System uses fuzzy logic learning algorithms applied on historical patient data to update and improve its diagnosis rules set. This process, collectively called the learning process, would benefit greatly from a web-based framework that could provide services like patient data transfer and web-based distribution of updated rules [1].

Keywords: Neuro-Fuzzy Diagnosis and Prediction, Web-Centric Expert System, Telehealth, HL7 standard

1. INTRODUCTION

Glaucoma is a progressive eye disease that damages the optic nerve, usually associated with increased intraocular pressure (IOP). If left untreated, it can lead to blindness.

An expert system would be useful to help doctors to make consistent diagnosis by integrating different experiences into the knowledge database. Besides diagnosis, other important outputs of the system could be the potential to evaluate the risk of glaucoma occurrence, its progression and when and how should the patient be treated (for example when to do surgery taking into account alcohol breath, problems using drops, timing, etc).

2. OBJECTIVES AND SYSTEM REQUIREMENTS

The objectives of this work are:

- Increase the accuracy and consistency of diagnosing the progression of glaucoma
- Centralize data from different type of tests

- Develop a Fuzzy Inference Machine to analyze patient data
- Allow worldwide doctors and patients access to Glaucoma Expert System

The major system requirements are defined as follows:

- Ability to access the system from any part of the world
- Use of Fuzzy-logic learning algorithms to update and improve diagnosis
- SSS (Stability, Scalability and Security)
- Support one of the most widely used E-Health standards
- Support 2 types of users: on-line and off-line
- Notification and Alert messaging service
- Clientless configuration for on-line users
- System architecture should ensure delivering of various types of data, such as text, images, binary data, etc

3. DATA FLOW AND ANALYSIS

The proposed Web-Centric glaucoma diagnosis system serves as a framework upon which medical services can be offered over the World-Wide Web (WWW). The most important of these services are provided by the Web-Centric Expert System via the inference engine (implemented using the Fuzzy Control Manager (FCM) software from Transfertech GmbH Germany), which has the ability to diagnose the severity of a case of Glaucoma based on patient data within the system. Data transfers normally performed informally between the doctor and the patient will now be managed by the web-centric glaucoma system, including data from expert eye examination devices. The Web-Centric extension focuses on implementing the data transfers that are required between the Internet and the Web-Centric Expert System, Fig. 1. The doctor is able to enter new patient data on the Web that is subsequently sent to the Web-Centric database. When a user (this is a doctor, patient or a registered user) requires a glaucoma diagnosis, the data that he submits through the Web, combined with the stored user profile, are entered into the FCM. The FCM will then, based on the rules and the submitted information, generate a diagnosis for that user. Finally, that diagnosis will be sent back to the user who will still be online. Alert messages are implemented so that they are sent out to the user as well as to the doctor (only if that user is a patient of a particular doctor) in emergency situations, so that immediate actions can be taken. Appropriate graphical user interface (GUI) is developed that allows the doctor to enter/change/retrieve patient data, Fig. 2. Another administrative GUI is designed in order to manage registered users who do not belong to a specific doctor. The patient or the registered user has access to his/her diagnosis but is not able to modify the data that was entered by the doctor or the system administrator. Finally, a Web-Centric extension will also be able to receive data from expert eye examination devices and update the Web-Centric database with these test results so that an updated diagnosis can be generated. Thus, the user and possibly the doctor (if that user is a patient of a particular doctor) can be informed about the new patient status.

4. HL7 version 3.0 SUPPORT

Health Level Seven (HL7) is one of several ANSI-accredited standards developed by the Standards Developing Organization (SDO) operating in the healthcare arena. HL7 domain is clinical and administrative data. "Level Seven" refers to the highest level of the International Standards Organization's (ISO) communications model for Open Systems Interconnection (OSI) - the application level. The application level addresses the definition of the data to be exchanged, the timing of the interchange, and the communication of certain errors to the application. The seventh level supports such functions as security checks, participant identification, availability checks, exchange mechanism negotiations and, most importantly, data exchange structuring.

Why we chose HL7 standard?

- An HL7 message definition, known as the Hierarchical Message Description (HMD) is technology neutral;
- Version 3 includes the idea of an Implementable Technology Specification (ITS) to define how a message can be instantiated from its definition;
- HL7 has chosen Extensible Markup Language (XML) as the target of the first (and so far the only) ITS specification;
- XML schemas and non-healthcare specific XML parsers can be used to parse inbound and outbound messages;
- Parsers are “free”;
- XML-related tools are readily available, e.g. viewing, testing, and conversion.

5. ARCHITECTURE

Figure 1 represents the present system architecture, which mirrors the majority of similar distributed web-based applications. This architecture includes a Web-Centric Production Server, Central Learning Expert System and 2 typical Web-Centric clients: On-line User, which has Internet browser only and a standalone user, which has his own Fuzzy Control Manager (FCM) and database. The Web-Centric Production Server and the Central Learning Expert System have the same architecture and modular structure, but they differ in the functionality of their manager objects.

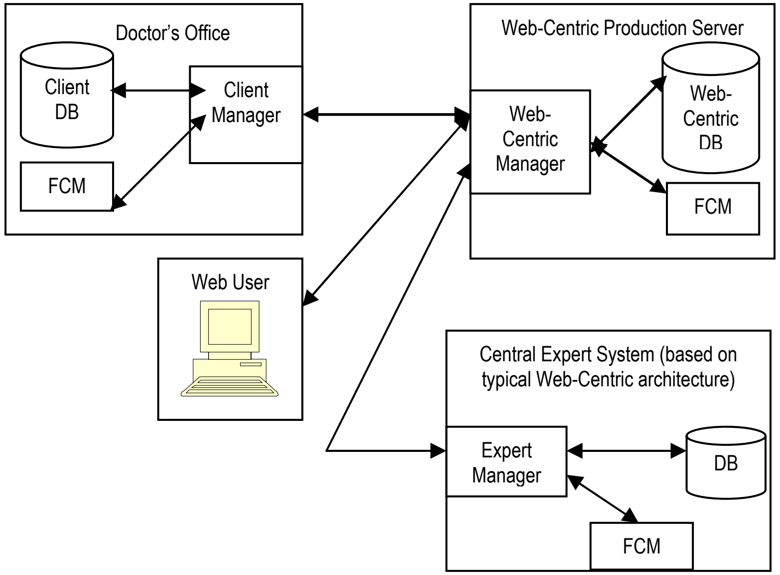


Figure 1. System Architecture

5.1 Web Centric Server

FCM Service

The FCM service interacts with the doctor via the functional server. This FCM is able to receive new patient/registered user data and give an updated diagnosis regarding his/her condition.

Database

This database holds all existing patient data, permits changes to that data and allows for addition of new patients and patient data. Also registered user data is stored in this database (users who do not have a doctor) so that these users also are able to obtain glaucoma diagnosis from the system. Database is implemented by usage of Oracle technology.

Web-Centric Manager

The manager service acts as a barrier between the Internet and the local web-centric database containing the sensitive patient data. It handles all incoming requests and ensures that they are authorized and have not been altered. All requests are verified to be unique and complete and if so, confirmation will be sent to the user to notify them of a successful transaction.

5.2 Central Expert System Server

FCM service

This subsystem performs the learning functions of the expert system. The FCM produces updated versions of the rules based on historical data [2].

Database

This database is capable of holding large amounts of data and provides functionality for fast traversal of this data. It contains only data regarding the development of glaucoma and does not hold any personal information about the patients involved.

Expert Manager

The functional server provides the means of interaction between the Central learning expert system and the Web-Centric Production Server to allow the acceptance of new/changed data from the Web-Centric database and the transfer of new rules to the web-centric system. There are no interactions between this subsystem and the doctors or users.

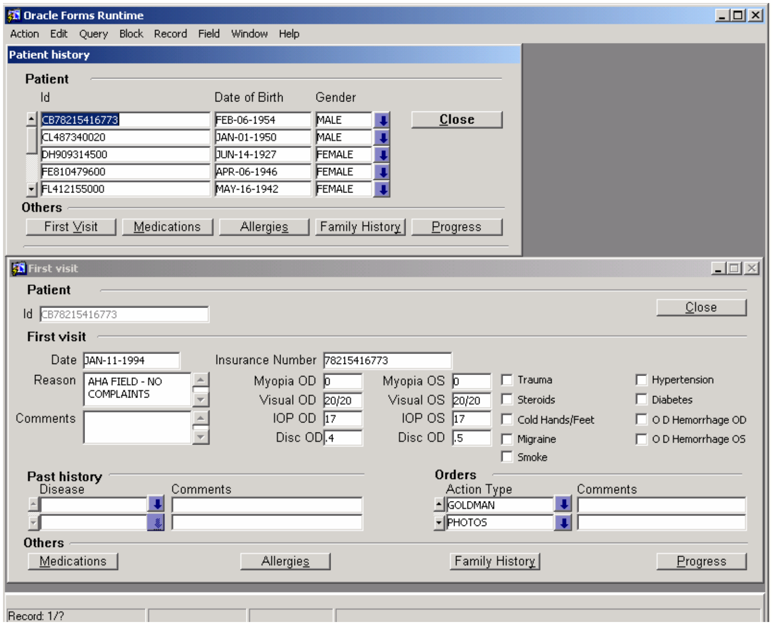
5.3 Doctor's Office Software and On-line User

Doctor's Office Software has the same structure as a typical Web-Centric node, but it is extended by ActiveX control, which is responsible for data delivering management between the local database, the Internet Browser and the Fuzzy Control Manager. The local database is currently developed using MS Access, but it may be implemented using any different relational database system (RDBS) technology, which supports open database connectivity (ODBC).

A typical on-line user has Internet browser with XML support only. A doctor is able to perform all on-line user operations, but all data will be stored in the Web-Centric Production database. On-line patient/guest is able to submit his symptoms and obtain on-line diagnostic results only.

5.4. User Interface

The patients/registered users are not able to change any of their data. The only functionality that they have is to enter their symptoms into the system to obtain an updated diagnosis. Then that diagnosis will be sent back to this user



Patient

Id	Date of Birth	Gender
CB78215416773	FEB-06-1954	MALE
CL487340020	JAN-01-1950	MALE
DH909314500	JUN-14-1927	FEMALE
FE810479600	APR-06-1946	FEMALE
FL412155000	MAY-16-1942	FEMALE

Patient

Id CB78215416773

First visit

Date JAN-11-1994 Insurance Number 78215416773

Reason AHA FIELD - NO COMPLAINTS

Comments

Myopia OD 0 Myopia OS 0

Visual OD 20/20 Visual OS 20/20

IOP OD 17 IOP OS 17

Disc OD .4 Disc OD .5

□ Trauma □ Hypertension
□ Steroids □ Diabetes
□ Cold Hands/Feet □ O D Hemorrhage OD
□ Migraine □ O D Hemorrhage OS
□ Smoke

Past history

Disease	Comments

Orders

Action Type	Comments
GOLDMAN	
PHOTOS	

Record: 1/?

Fig 2: GUI for patient (first visit) information screen

The doctor has the most control over the system. The doctor will have to enter/delete/modify/search detailed patient information. When a new patient arrives to the eye clinic or to the doctors office, the patient has to fill in a form containing personal information (complete name, date of birth, etc), the specific reason for the appointment, previous history of eye problems or injury, eye surgeries, allergies, medications the patient is on, medical history (diabetes, high blood pressure, kidney, etc), family history (diabetes, glaucoma, blindness, retinal detachment), personal and family eye health history, and surgeries. Then a nurse measures the patient's intraocular pressure.

5.5. Eye Test Machine Data

The eye test machine data consists of information diagrams stored in the Production and Central Expert databases containing information from the ophthalmic test machines (the Humphrey Field Analyzer (HFA) and the Humphrey and Heidelberg Retina Tomograph (HRT)). For the communication we use an Internet platform (since our system is a typical distributed web application), which performs XML-message exchange over the HTTP/HTTPS protocol

CONCLUSIONS

The proposed Telehealth system for glaucoma progression monitoring supplies (via usage of Web-centric technology) the ability to access our glaucoma expert system from any part of the world. Use of Fuzzy-logic learning algorithms offers a possibility to update the knowledge base continuously [1] and by this improve previous diagnosis, and share self-learning process data to other customers. We developed a generic interface to the diagnostic system that enables access from any part of the world. The system includes integrated Web server, secured data storage and transfer, delivering of data within XML messages, which are prepared according to HL7 v.3 standard. The proposed architecture of Glaucoma Progression and Monitoring System might be adapted for any E-Health System, which requires supporting of widely used E-Health standards.

REFERENCES

[1] Mihaela Ulieru, "Internet-Enabled Soft Computing Holarchies for e-Health Applications", in New Directions in Enhancing the Power of the Internet, (L.A. Zadeh and M. Nikravesh – Editors), Springer Verlag, Berlin, 2003

[2] Ulieru M. and Pogrzeba, G. Integrated Soft Computing Methodology for Diagnosis and Prediction with Application to Glaucoma Risk Evaluation, Proceedings of 6th IASTED International Conference on Artificial Intelligence and Soft Computing, July 17-19, 2002, Banff, Canada, pp. 275-280, ISBN: 0-88986-346-6

[3] Ulieru, M., Web-Centric Diagnostic and Prediction System for Global manufacturing, Proceedings of IFSA-NAFIPS 2001, Vancouver, BC, July 24-29, 2001.

ON STATISTICAL HYPOTHESIS TESTING VIA SIMULATION METHOD

B. Dimitrov, D. Green Jr., V.Rykov, and P. Stanchev

Department of Science & Mathematics, Kettering University, Flint, Michigan, USA
1700 West Third Avenue, Flint, Michigan 48504-4898, USA
e-mail {bdimitro, dgreen, vrykov,pstanche}@Kettering.edu

Abstract: A procedure for calculating critical level and power of likelihood ratio test, based on a Monte-Carlo simulation method is proposed. General principles of software building for its realization are given. Some examples of its application are shown.

1. Introduction

In this paper we show how the present day fast computer could solve non-standard old statistical problems. In most cases statisticians work with approximations of test statistics distributions, and then use respective statistical tables. When approximations do not work the problem is usually tabled. We propose such simulation approach which we do believe could be helpful in many case.

The problem of statistical hypothesis testing is very important for many applications. In the notable but rare case, it is possible to find some simple test statistic having a standard distribution. However, in the general case the statistics based on the *Likelihood Ratio Test* (LRT) does not usually have one of the known standard distributions. The problem could be overcome with the help of an appropriate simulation method. This method was first used in [3] for a specific case of almost lack of memory (ALM) distributions. In this paper we propose a general approach for using the method, describe its general principles and algorithms, show how to build up an appropriate software, and illustrate with examples it application.

2. LRT and the Simulation Approach

It is well known according to Neyman-Pearson theory [7], that the *most powerful test* for testing a null hypothesis H_0 : *f(x)* = f_0 *(x)* versus an alternative H_1 : *f(x)* = f_1 *(x)* is the LRT. For this test, the *critical region W* for a sample x_1 ,..., x_n of size *n* has the form

$$W=\left\{ (x_1,\ldots,x_n) : w(x_1,\ldots,x_n) = \frac{f_1(x_1,\ldots,x_n)}{f_o(x_1,\ldots,x_n)} > t \right\},$$

where $f_0(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n})$ and $f_1(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n})$ are joint probability densities of the distributions of observations (the *likelihood functions*) under hypotheses H_0 and H_1 with probability density functions (p.d.f.) $f_0(.)$ and $f_1(.)$ respectively. The notation

$$\mathrm{w}(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n}) = \frac{f_1(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n})}{f_o(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n})}$$

is used for *test's statistic*. For independent observations this statistic can be represented in the form

$$\mathrm{w}(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n}) = \frac{f_1(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n})}{f_o(\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{n})} = \prod_{1\le i\le n}\frac{f_1(x_i)}{f_0(x_i)}.$$

Considering the observations x_1 ,..., x_n as independent realizations of random variable (i.i.d. r.v.) *X* with p.d.f. $f_0(.)$ the *significance level* of the test is

$$\alpha = P_{H_0}\{W\} = P_{H_0}\{w(X_1,\ldots,X_n) > t_\alpha\}. \qquad (1)$$

Here an appropriate *critical value* t_α for any given significance level α is the smallest solution of equation (1).

Analogously, considering the same observations x_1 ,..., x_n as independent realizations of random variable *Y* with p.d.f. $f_1(.)$, *the power of the test* is

$$\pi_\alpha = P_{H_1}\{W\} = P_{H_1}\{w(Y_1,\ldots,Y_n) > t_\alpha\}. \qquad (2)$$

Thus, to find the critical value for a given significance level α and the power of the test $\pi\ (\alpha)$, a statistician needs to know the distributions of the test statistic w under hypotheses H_0 and H_1.

For parametric hypothesis testing the problem becomes more complicated because in such cases one has to be able to find a free of parameters distribution of this statistic.

To avoid calculations of these functions we propose to use the simulation method. This means that instead of searching for exact statistical distributions, we will calculate appropriate empirical distributions as their estimations. This method gives desired results due the fact (based on the *Strong Law of Large Numbers*) that the empirical distribution function of the test statistic converges with probability one to the theoretical distribution.

In the following, due to numerical reasons, instead of statistic *w* we will use its natural logarithm, and for simplicity we will denote this statistic with the same latter, w,

$$w = \ln \prod_{1\le i\le n}\frac{f_1(x_i)}{f_0(x_i)} = \sum_{1\le i\le n}(\ln f_1(x_i) - \ln f_0(x_i)). \qquad (3)$$

Due to additional statistical reasons, instead of the cumulative distribution functions (CDF) of the statistic w under hypotheses H_0 and H_1, we will use their tails,

$$\overline{F_o}\ (t) = P_{H_0}\{w(X_1,\ldots,X_n) > t\ \}, \qquad (4)$$

and

$$\overline{F_1}\ (t) = P_{H_1}\{w(Y_1,\ldots,Y_n) > t\ \}. \qquad (5)$$

For large size samples, *n*>>1, it is possible to use a simplier approach based on the *Central Limit Theorem*. It is well known that this theorem provides a normal approximation of the distribution for sums of i.i.d. r.v.'s under conditions of existence of finite second moments. This would allow one to calculate and use only two moments of the test's statistic w and then to calculate the appropriate significance level and power of the test making use of the respective normal approximation.

To show how it works, let us denote by *U* and *V* the r.v.'s

$$U = \ln f_1(X) - \ln f_0(X), \qquad V = \ln f_1(Y) - \ln f_0(Y),$$

where *X* and *Y* are taken from distributions with densities $f_0(.)$ and $f_1(.)$ respectively, corresponding to hypotheses H_0 and H_1. Denote by μ_U, μ_V and σ_U^2, σ_V^2 their expectations and variances respectively, when they exist. Then, for large samples, *n*>>1, under null hypothesis, the test's statistic w has approximately normal distribution with parameters $n\,\mu_U$, and $n\,\sigma_U^2$. This means that the significant level t_α for given value of α can be found from the equation

$$\alpha = P_{H_0}\{w(X_1,\ldots,X_n) > t\ \} = P_{H_0}\{\frac{w - n\mu_U}{\sigma_U\sqrt{n}} > \frac{t_\alpha - n\mu_U}{\sigma_U\sqrt{n}}\} \approx 1 - \Phi\left(\frac{t_\alpha - n\mu_U}{\sigma_U\sqrt{n}}\right),$$

or equivalently

$$\frac{t_\alpha - n\mu_U}{\sigma_U\sqrt{n}} \approx z_{1-\alpha}\,.$$

Here $z_{1-\alpha}$ is the (1-α)-quantile of the standard normal distribution. Thus, the critical value t_α for the test statistic w at a given significance level α is

$$t_\alpha \approx n\,\mu_U + z_{1-\alpha}\,\sigma_U\sqrt{n}\,. \qquad (6)$$

The power of the test equals

$$\pi_\alpha = P_{H_1}\{w(Y_1,\dots,Y_n) > t_\alpha\} = P_{H_1}\{\frac{w-n\mu_V}{\sigma_V\sqrt{n}} > \frac{t_\alpha - n\mu_V}{\sigma_V\sqrt{n}}\} =$$

$$= 1 - \Phi\left(\frac{t_\alpha - n\mu_V}{\sigma_V\sqrt{n}}\right) = 1 - \Phi\left(\frac{\mu_U-\mu_V}{\sigma_V}\sqrt{n} + z_{1-\alpha}\frac{\sigma_U}{\sigma_V}\right). \qquad (7)$$

From this equality it is possible to see that the power of the test mainly depends on the difference in expectations of the r.v.'s *U* and *V*.

In some cases the parameters μ_U, μ_V and σ_U^2, σ_V^2 can be calculated in closed (explicit) form. In general it is possible to estimate them also with the help of Monte-Carlo techniques and then use the respective estimated values instead of the exact ones. Appropriate algorithms for calculating the empirical cumulative distribution functions (CDF) of the test's statistic under hypotheses H_0 and H_1 for both cases are described below.

3. Algorithms

In this section two algorithms for calculation of the tails of CDF of LRT's statistic w under both null and alternative hypothesis (the null H_0 : *f(x)* = f_0 *(x)* and the alternative H_1 : *f(x)* = f_1 *(x)*), based on Monte-Carlo method are proposed. One algorithm can be applied for any sample size *n*. The second algorithm should be used for large samples, *n*>>1, mainly when the parameters μ_U, μ_V and σ_U^2, σ_V^2 are finite.

Algorithm 1. LRT for any sample size

Begin. Select the p.d.f.'s $f_0(.)$ and $f_1(.)$, and the sample size *n*.

Step 1. Generate a sequence of *N* random samples $(x_1^{(j)},\dots,x_n^{(j)})$, *j=1,…,N,* from a distribution with p.d.f. $f_0(.)$, and calculate *N* values of the test's statistics

$$w_j = w(x_1^{(j)},\dots,x_n^{(j)}) = \sum_{1\le i\le n}(\ln f_1(x_i^{(j)}) - \ln f_0(x_i^{(j)})),\ j=1,\dots,N. \qquad (8)$$

Step 2. Calculate the complementary empirical distribution function

$$\overline{F}_{0,N}(t) = \frac{1}{N}\{\text{number of } w_j\text{'s} > t\}, \quad t > 0.$$

Step 3. Calculate the critical value t_α for the test statistic w at a given significance level α as the smallest solution of the equation $\overline{F}_{0,N}(t) = \alpha$.

Step 4. Generate a sequence of *N* random samples $(y_1^{(j)},\dots,y_n^{(j)})$,*j=1,…,N,* from a distribution with p.d.f. $f_1(.)$, and calculate the values of the analogous to (8) test statistics w_j with $y_i^{(j)}$'s instead of $x_i^{(j)}$'s.

Step 5. Calculate the complementary empirical distribution function for the new sample $\overline{F}_{1,N}(t) = \frac{1}{N}\{\text{number of } w_j\text{'s} > t\}$, $t > 0$.

Step 6. Calculate the power of the test statistic w at the given significance level α from the equation $\overline{F}_{1,N}(t_\alpha) = \pi_\alpha$.

Step 7. **Enter the application's data:** For a given user's sample $(x_1,\dots,x_n)$, calculate the test statistic

$$w = w(x_1,\dots,x_n) = \sum_{1\le i\le n}(\ln f_1(x_i) - \ln f_0(x_i)).$$

Calculate the *p*-value for testing the null hypothesis H_0 : *f(x)* = f_0 *(x)* versus the alternative H_1 : *f(x)* = f_1 *(x)* by making us of the Likelihood Ratio Test from the equation $\overline{F}_{0,N}(w) = p - \text{value}$.

Make a decision by comparing the calculated *p*-value and α. Alternatively, reject the hypothesis H_0 if the inequality $w > t_\alpha$ holds.

Calculate the *probability of committing an error of type II* (when testing the null hypothesis H_0 : *f(x)* = f_0 *(x)* versus the alternative H_1 : *f(x)* = f_1 *(x)* by making use of the Likelihood Ratio Test by the simulation method) from the equation

$1-\overline{F}_{1,N}(w)=\beta$ – the probability of type II error.

Step 8. Print results:

- The chosen null hypothesis H_0: *f(x)* = f_0 *(x)* and alternative hypothesis H_1 : *f(x)* = f_1 *(x),* the selected significance level α, and the sample size *n* .
- The *p*-value of the test
- The power of the test, $\pi_\alpha = 1-\beta$.
- The calculated value of the test statistic w , and the calculated by simulation critical value t_α
- The graphs of the tails of the empirical CDFs $\overline{F}_{0,N}(t)$ and $\overline{F}_{1,N}(t)$.

End.

For large size samples when the second moments of the r.v.'s *U,* and *V* exist, it is possible to modify and simplify the simulation algorithm as shown below.

Algorithm 2. LRT for large samples.

Begin. Select the p.d.f.'s $f_0(.)$ and $f_1(.)$, and the sample size *n*.

Step 1. Generate a sequence of *N* random variables $(x_1,\dots,x_N)$, from a distribution with p.d.f. $f_0(.)$, and calculate *N* values of the statistics

$$u_j = u\,(x_j) = \ln f_1(x_j) - \ln f_0(x_j)\,,\quad j=1,\dots,N. \qquad (9)$$

and its sample mean $\overline{u}$, and sample variance s_u^2 according to

$$\overline{u} = \frac{1}{N}\sum_{1\le j\le N} u_j\,,\quad s_u^2 = \frac{1}{N}\sum_{1\le j\le N}(u_j-\overline{u})^2 = \frac{1}{N}\sum_{1\le j\le N} u_j^2 - (\overline{u})^2\,. \qquad (10)$$

Step 2. Calculate the critical value t_α for the test statistic w at a given significance level α from the equation

$$t_\alpha \approx n\overline{u} + z_{1-\alpha}\cdot s_u\cdot\sqrt{n}\,, \qquad (11)$$

where $z_{1-\alpha}$ is the (1-α)-quantile of the standard normal distribution.

Step 3. Generate a sequence of *N* random variables $(y_1,\dots,y_N)$, from a distribution with p.d.f. $f_1(.)$, and calculate *N* values of the statistics

$$v_j = v\,(y_j) = \ln f_1(y_j) - \ln f_0(y_j)\,,\quad j=1,\dots,N. \qquad (12)$$

and its sample mean $\overline{v}$, and sample variance s_v^2 according to rules (10) for the data (12).

Step 4. Calculate the power of the test at the given significance level α from the equation

$$\pi_\alpha = 1-\Phi\left(\frac{\overline{u}-\overline{v}}{s_V}\sqrt{n} + z_{1-\alpha}\frac{s_U}{s_V}\right). \qquad (13)$$

where $\Phi(x)$ is the c.d.f. of the standard normal distribution.

Step 5. Enter the application's data: For a given user's sample $(x_1,\dots,x_n)$, calculate the test statistic

$$w = w\,(x_1,\dots,x_n) = \sum_{1\le i\le n}(\ln f_1(x_i) - \ln f_0(x_i))\,.$$

Calculate the *p*-value for testing the null hypothesis H_0 : *f(x)* = f_0 *(x)* versus the alternative H_1 : *f(x)* = f_1 *(x)* by making us of the Likelihood Ratio Test from the equation

$$p\text{-value}= P_{H_0}\{w(X_1,\dots,X_n) > w\} \approx 1-\Phi\left(\frac{w-n\cdot\overline{u}}{s_u\sqrt{n}}\right),$$

where w is the calculated statistic from the sample. Make a decision by comparing the calculated *p*-value and α. Alternatively, reject the hypothesis H_0 if the inequality $w > t_\alpha$ holds, where t_α is calculated by (11).

Calculate the *probability of committing an error of type II* (when testing the null hypothesis H_0 : *f(x)* = f_0 *(x)* versus the alternative H_1 : *f(x)* = f_1 *(x)* by making use of the LRT by the simulation method) from the equation β = $1-\pi_\alpha$ with the

π_α calculated in Step 4.

Step 6. Print results:

- The chosen null hypothesis H_0: *f(x)* = f_0 *(x)* and alternative hypothesis H_1 : *f(x)* = f_1 *(x),* the selected significance level α, and the sample size *n* ;
- The *p*-value of the test;
- The power of the test, $\pi_\alpha = 1 - \beta$;
- The calculated test statistic w, and the calculated by simulation critical value t_α ;
- The graphs of the tails of the CDFs

$$\overline{F}_{0,N}(t) = 1 - \Phi\left(\frac{t - n \cdot \overline{u}}{s_u\sqrt{n}}\right), \text{ and } \overline{F}_{1,N}(t) = 1 - \Phi\left(\frac{t - n \cdot \overline{v}}{s_v\sqrt{n}}\right).$$

End.

4. The Software

For practical application of the above algorithms an appropriate software should be elaborated. The software should have a friendly interface, which allows work in two different regimes: *individual* (customized), and *automatic.*

In the individual regime only particular observations are tested for any pair of given null and alternative hypotheses. Automatic regime allows one to calculate and show the significance level and power functions as functions of the test's statistic, and also as functions of some parameters of the model. In this way it would allow one to investigate some parametric models.

The interface includes the main menu, which allows the users to choose:

- the regime for investigation;
- the p.d.f. for hull and alternative hypotheses from a given list of distributions, which include almost all standard discrete and continuous distributions, or
- propose an option to the user for selecting probability distribution's formulae or tables of his/her own choice.

The submenu allows:

- one to choose the parameter values for hypothesis testing for individual regime; or
- one to choose the intervals and steps of increment for parameters varying for the problem investigated in an automatic regime.

The software allows also different type of presentation of the results: numerical, graphical, comparison with respect to various variables, or with respect to family of functions. These and other appropriate possibilities make the content of the design menu.

The design will be based on the new technologies presented in [8].

5. An Example

Below we consider one example on which the work of algorithms in the previous section will be illustrated.

Example. An ALM distribution versus other ALM distribution with uniform distribution.

It is known that when in the ALM distribution

$$f(x) = (1-a) \cdot a^{\left[\frac{x}{c}\right]} f_Y(x - \left[\frac{x}{c}\right]c), \qquad (14)$$

where a is a parameter of distribution, c is the length of a period, and $f_Y(x)$ is an arbitrary distribution on the interval [0,c). More details about ALM distributions can be found in [5].

Here for the ALM distribution in the null hypothesis H_0 we choose $f_0(x)$, presented by (14) with parameters chosen in the following way

$$c=1, \quad a_0 = .5, \quad f_{Y,0}(x) = 1 \text{ for } 0 \leq x \leq 1. \qquad (15)$$

This means that the r.v. *X* with distribution (14) is based on the uniform distribution of Y_0 on [0,1] (cycle of length 1), and probability for jump over a cycle without success is a_0 =.5. Any other choice of the parameter $a_0 \neq .5$ will produce an ALM distribution $f_1(x)$, different from the chosen $f_0(x)$. And this p.d.f. $f_1(x)$ will appear in our considerations as an alternative hypothesis H_1 .

Thus, we study the likelihood ratio test according to Algorithms 1 and 2 above with the choice for the p.d.f. $f_0(x)$, with a_0 =.5, and choosing various other values for parameter $a_1 \neq .5$. In studying the power function dependence on significance level α we select a_1 =.05,.1,.15, . . . ,.9,.95; *N* = 10000, *n* = 10, c = 1.

The results for the power function in this case of significance level α=.05 are shown on Fig. 1.

$F^{(0)}$: a_0 =.5 $F^{(1)}$, a_1 = .1 Power f-n π_α

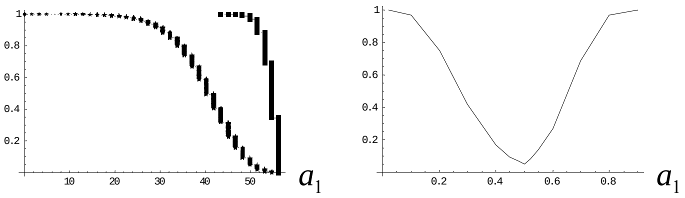


Fig. 1. Cumulative distribution functions for the test statistic and the power function of the test

6. Conclusions

The problem of hypotheses testing arises in many statistical applications. In analytical form its solution can be done for a very limited number of cases. The method proposed in this paper gives the solution for practically all cases. Nevertheless, for its practical realization special computer tools with friendly interface are needed. This work is now in the progress, and we show here some examples of the approach used for some special case of distributions, - so called almost lack of memory distributions.

REFERENCES

[1] **S.Chukova, and B.Dimitrov** (1992) On distributions having the almost lack of memory property. *Journal of Applied Probability*, **29** (3), 691-698.

[2] **S.Chukova, B.Dimitrov, and Z. Khalil** (1993) A characterization of probability distributions similar to the exponential. *The Canadian Journal of Statistics*, **21** (3), 269-276.

[3] **B.Dimitrov, Z.Khalil, M.Ghitany, and V.Rykov** (2001) Likelihood Ratio Test for Almost Lack of Memory Distributions. *Technical Report No. 1/01*, November 2001, Concordia University, 2001.

[4] **B.Dimitrov, and Z.Khalil** (1992) A class of new probability distributions for modeling environmental evolution with periodic behavior. *Environmetrics*, **3** (4), 447-464.

[5] **B.Dimitrov, S.Chukova, and D.Green Jr.**(1997) Distributions in Periodic Random Environment and Their Applications. *SIAM J.on Appl. Math*., **57** (2), 501-517.

[6] **B.Dimitrov, Z.Khalil, and L.Le Normand** (1993) A statistical study of periodic environmental processes, *Proceedings of the International Congress on Modelling and Simulation*, Dec. 6-10, 1993, The Univ. of Western Australia, Perth, **v. 41**, pp. 169 - 174.

[7] **R.A.Johnson** (1994) *Miller & Freund's Probability and Statistics for Engineers,5th edition,* Prentice Hall, New Jersey 07632.

[8] **J.Carroll** (2002) *Human - Computer Interaction in the New Millennium*, ACM Press, NY.

[9] **A,M.Law and W.D.Kelton** (2000) *Simulation Modeling and Analysis,* 3d edition, McGraw-Hill, N.Y.

THE MANDELBROT SET FOR JULIA SETS OF ARBITRARY ORDER. A REMARK ON THE SHAPE OF CUBIC MANDELBROT AND JULIA SETS. MANDELBROT AND JULIA SETS FOR THE POLYNOMIALS OF ARBITRARY ORDER.

Anna V. Tomova

Bul."Slivnica" No 8, Varna,9000,Bulgaria
Email: anna_bg_2000@yahoo.com

Abstract: The Mandelbrot set for Julia sets, associated with $f_c(z)=z^2+c$, $c \in C$ is in detail very good studied. This family is of special importance because it provides a model for the onset of chaotic behaviour in physical and biological systems.Moreover it was the first family of dynamical systems for which a useful computergraphical map was constructed by Mandelbrot. In this paper we restrict the attention to the families: $f_c(z)=z^n+c$, $f_{p,q}(z)=z^3+pz+q$; $c,p,q\in C$. We proof any theorems for the limits of Mandelbrot set for Julia sets of arbitrary order and for cubic Mandelbrot and Julia sets. We proof any theorems for the limits of the Mandelbrot set for Julia sets, associated with the polynomials of arbitrary order: $f(z)=z^n+a_1z^{n-1}+...+a_n$ and consider any examples.

Keywords: Complex dynamical systems, fractals, Mandelbrot set for Julia sets of arbitrary order, Mandelbrot and Julia cubic sets, Mandelbrot and Julia sets for the polynomials of arbitrary order.

2000 Mathematics Subject Classification Codes: 37Fxx, 37F50

1. Introduction

In Barnsley (1988) we have found the following remark: "…The Random Iteration Algorithm can also be applied to compute Julia sets of cubic and quatric polynomials, and of special polynomials of higher degree such as z^n+c when n=5,6,7,…, and $c \in C$." In this paper we restrict attention to the families:

$$\{\hat{C}:p(z) = z^n + c\},\ c \in \mathbf{C}, n \in \mathbf{N}, n \geq 3 \qquad (1)$$

2. Main results.

2.1. Mandelbrot Set for Julia Sets of Arbitrary Order.

In this paper we will follow the same *Definition* as in [1],[5] for the *Julia set* J_p and the *filled Julia set* F_p. We will give the similar definition as *Definition* for the *Mandelbrot set* as in [1],[5].

Theorem 1. The Julia set for a member of the family of dynamical systems

$\{\hat{C}:p_c(z) = z^n + c\}, c \in C$ $n \in N, n \geq 3$ is connected if and only if the orbit of the origin does not escape to infinity;that is

$M = \{c \in C: \left| p_c^{0k}(0) \right| < \infty$ as $k \to \infty\}$.

Proof. This is essentially the same as the sketch of the proof of Theorem 1,p.316 in Barnsley (1988). This theorems follows from Brolin,Theorem 11.2, which says that the Julia set of a polynomial,of degree greater than one,is connected if and only if none of the finite critical points lie in the basin of attraction of the Point at Infinity. We find $p_c'(z) = nz^{n-1}$,$p_c(z)$ possesses one finit critical point,*O*. Hence J(c) is connected if and only if $\left|p_c^{0k}(0)\right| < \infty$ as $k \to \infty$. The proof is completed.

Let us now consider the equation:

$$\mathbf{z^n - z - |c| = 0.} \qquad \mathbf{(2)}$$

We will proof the following

Theorem 2. The equation (2) has an unique positive root R_c>1. If n = 2m it has an unique negative root too. If n = 2m + 1 the number of the negative roots is two or 0.

Proof. *The first method*:see *Decart'sTheorem* for the number of the posirtive roots of the polynomials equations on the page 245 in (Kurosh,1968). *The second method***:**we can use the graphical method for proof the fact that (2) has an unique positive root R_c for every n and if =2m (2) has an unique negative root too and if n = 2m + 1 the number of the negative roots of (2) is two or 0. We have: $\mathbf{p_c(z) = z^n - z - |c| = (z - R_c)(z^{n-1} + R_c z^{n-2} + R_c^2 z^{n-3} + \ldots + R_c^{n-2} z + R_c^{n-1} - 1)}$,

$\mathbf{R_c(R_c^{n-1} - 1) = |c|}$,Also the unique positive root of (2) is $R_c > 1$.

Theorem 3. The Julia set for a member of the dynamical system(1) lies in the circle $|z| \leq R_c$ where R_c>1 is the unique positive root of the equation (2).

Proof. The method of the proof is borrowed from Beardon and Rippon (1994). We have:

$$|P_c(z)| \geq \left\| z^n \right| - |c| \,| = \left| \frac{|z^n|}{R_c^n} R_c^n - |c| \right| = \left| \frac{|z^n|}{R_c^n} (R_c + |c|) - |c| \right| =$$

$$\left| \frac{|z^n|}{R_c^n} R_c + |c| \left(\frac{|z^n|}{R_c^n} - 1 \right) \right| = \left(\frac{|z|}{R_c} \right)^n R_c + |c| \left(\frac{|z^n|}{R_c^n} - 1 \right) > \frac{|z^n|}{R_c^n} R_c > R_c \text{ if } |z| > R_c$$

Then if $|z| > R_c$ we have the estimate:

$$\left|p_c^{0k}(z)\right| > \left(\frac{|z|}{R_c} \right)^{n^k} R_c \to \infty \text{ as } k \to \infty \qquad (3)$$

The proof is completed.

Corollary. All others roots of (2) lie in the circle $|z| \leq R_c$.

Let us consider the case (1) in detail:$p_c(0)=c$,$p_c^{02}(0)=c^n+ c$. Let c is the solution of the equation $c^n + c = 0$;that is $c_1=0$,$c_{2,\ldots,n} = e^{\frac{i(\pi+2k\pi)}{n-1}}$,k = 0,…,n-2. Then c_j,j=1,2,…,n belong to M. Let n = 2m and we solve the equation:$p_c^{02}(0) = - c$; which solution are $c_1 = 0$, $c_{2,\ldots,n} = \sqrt[n-1]{2}\, e^{\frac{i(\pi+2k\pi)}{n-1}}$,k = 0,1,…,2m-2. Then c_j,j = 1,2,…,2m belong to M. This way we can proof own main result in the following

Theorem 4. The Mandelbrot set for the family of dynamical systems (1) satisfies: $M \subset \{c \in C: |c| \leq \sqrt[n-1]{2}\}$. $c_k = e^{\frac{i(\pi+2k\pi)}{n-1}} \sqrt[n-1]{2}$,k = 0,…,n-2 belong to M if n=2m.

Proof. We have proved that c_k and 0 belong to M if n=2m. Now we must proof that if $|c| > 2^{\frac{1}{n-1}}$ then $\left|p_c^{0k}(0)\right| \to \infty$ as $k \to \infty$. We consider the dynamical system $\{\hat{C}: f(p) = p^n + c, c \in C\}$.

Then we can prove the estimate (see the proof of the theorem 3):

$$\left|f^{0k}(p)\right| \geq \left(\frac{|p|}{R_c} \right)^{n^k} R_c \to \infty \quad \text{as } k \to \infty \text{ and } |p| > R_c$$

where R_c is the positive root of the equation $\mathbf{p^n - p - |c| = 0.}$

We denote :

$f(p_c(0))=[p_c(0)]^n + c =c^n + c = p_c^{02}(0)$;

$f^{02}(p_c(0)) = [f(p_c(0))]^n + c = [p_c^{02}(0)]^n + c = (c^n + c)^n + c = p_c^{03}(0)$…

$f^{0k}(p_c(0)) = p_c^{0(k+1)}(0)$.

The first proof's method

Let now we substitute p for $p_c(0)$ in (3): we have the estimate:

$$\left|f^{0k}(p_c(0))\right| > \left(\frac{|p_c(0)|}{R_c}\right)^{n^k} R_c \to \infty \text{ as } k \to \infty \text{ and } |p_c(0)| > R_c \qquad (4)$$

where R_c is the unique positive root of the equation $[p_c(0)]^n - p_c(0) - |c| = 0$. But $p_c(0) = c$. This way we can substitute for c $2^{\frac{1}{n-1}}$ and verify that is R = $2^{\frac{1}{n-1}}$ is the unique positive root of the equation **$p^n - p - |c| = 0$.** Then we can write (3) in the form

$$\left|p_c^{0(k+1)}(0)\right| > \left(\frac{|c|}{2^{\frac{1}{n-1}}}\right)^{n^k} 2^{\frac{1}{n-1}} \to \infty \text{ as } k \to \infty \text{ and } |c| > 2^{\frac{1}{n-1}} \qquad (5)$$

The proof is complete.

The second proof's method:

$$|f(p_c(0))| \geq \left(\frac{|p_c(0)|}{2^{\frac{1}{n-1}}}\right)^n \left(2^{\frac{1}{n-1}} + 2^{\frac{1}{n-1}}\right) - |c| = \left(\frac{|p_c(0)|}{2^{\frac{1}{n-1}}}\right)^n 2^{\frac{1}{n-1}} +$$

$$\frac{|p_c(0)|^n}{2} - |c| > \left(\frac{|p_c(0)|}{2^{\frac{1}{n-1}}}\right)^n 2^{\frac{1}{n-1}} \quad \text{if} \quad |p_c(0)| = |c| > 2^{\frac{1}{n-1}} \;;$$

$$\left|f^{02}(p_c(0))\right| \geq \left(\frac{|f(p_c(0))|}{2^{\frac{1}{n-1}}}\right)^n \left(2^{\frac{1}{n-1}} + 2^{\frac{1}{n-1}}\right) - |c| =$$

$$\left(\frac{|f(p_c(0)|}{2^{\frac{1}{n-1}}}\right)^n 2^{\frac{1}{n-1}} + \left(\frac{|p_c(0)|}{2^{\frac{1}{n-1}}}\right)^{n^2} 2^{\frac{1}{n-1}} - |c| > \left(\frac{|p_c(0)|}{2^{\frac{1}{n-1}}}\right)^{n^2} 2^{\frac{1}{n-1}} \quad \text{if}$$

$$|p_c(0)| = |c| > 2^{\frac{1}{n-1}} \;\ldots$$

$$\left|f^{0k}(p_c(0))\right| > \left(\frac{|p_c(0)|}{2^{\frac{1}{n-1}}}\right)^{n^k} 2^{\frac{1}{n-1}} \quad \text{if} \quad |p_c(0)| = |c| > 2^{\frac{1}{n-1}}$$

Then we can write (3) in the form (4). The proof is completed.

2.2. Mandelbrot Set for Cubic Julia Sets

Let us now restrict the attention to the families:

$\{C: f_{p,q}(z) = z^3 + pz + q; p,q \in C\}$ (6)

and consider the shape of cubic Julia sets. We will follow the same *Definition 1* for the *Julia set* J_p and the *filled Julia set* F_p and the same definition as *Definition 2* for the *Mandelbrot set* for (6). Let us now consider the equation:

$$z^3 - (|p| + 1)z - |q| = 0 \qquad (7)$$

We can proof the following theorems as *Theorem 4* and *Theorem 5.*

Theorem 5. If q = 0, the roots of (8) are 0 and $\pm\sqrt{1+|p|}$. Let $q \neq 0$. Then the equation (7) has an unique positive root $R_{p.q} > |p| + 1$. It has or two negative roots $r_1, r_2 : |r_1| \leq |r_2| < R_{p,q}$ too or two complex roots z_1, $z_2 = \overline{z_1}, |z_1| = |z_2| < R_{p,q}$.

Theorem 6. The Julia set for a member of the dynamical system(6) lies in the circle $|z| \leq R_{p,q}$ where $R_{p,q}>1$ is the unique positive root of the equation (7).

2.3. The Application of Mandelbrot and Julia Sets Theory to the Cubic Polynomials.

2.3.1. The Layers for the Cubic Mandelbrot Set.

Here we consider the following layers of the cubic Mandelbrot set:

- (Rep,Imp) – layer,where p = Rep + iImp is a constant, (Req,Imq) is a pixel;
- (Req,Imq) – layer, where q = Req + iImq is a constant, (Rep,Imp) is a pixel;
- (Rep,Imq) – layer, where Rep,Imp are constants,(Imp,Req) is a pixel;
- (Imp,Req) – layer, where Imp,Req are constants,(Rep,Imp) is a pixel;
- (Rep,Req) – layer, where Rep,Req are constants,(Imp,Imq) is a pixel;
- (Imp,Imq) – layer, where Imp,Imq are constants,(Rep,Req) is a pixel;

We can make the following remark with a great importance: we find $f'_{p,q}(z) = 3z^2 + p$, then the critical points are:$z_{1,2}$ = $\sqrt{\frac{-p}{3}}$ and $|z_{1,2}|^2 = \frac{|p|}{3}$,but $R^2_{p,q} \geq |p| + 1$,also we have the following

Remark. The layers of the cubic Mandelbrot set for (6) lie in the domain

$|z| \leq \max\limits_{|p|\leq|p_0|,|q|\leq|q_0|} R_{p,q}$, where $R_{p,q}$ is the *unique positive root* $R_{p.q} \geq |p|$ *+1* of (7) (see **Theorem 7**).

We know the formulae for the roots of the cubic polynomial equations(see [8],p.223). Then we can proof the following inequality:

$$\max_{|p|\leq|p_0|,|q|\leq|q_0|} R_{p,q} \leq \max(\sqrt[3]{4|q_0|},2\sqrt{\frac{|p_0|+1}{3}})$$

2.4. The general case:the Mandelbrot set for Julia sets,associated with the polynomials of arbitrary order.

We consider the general case,when f(z) is a polynomial of degree n:

$\{C:f(z) = z^n + a_1z^{n-1} + \ldots a_{n-1}z + a_n;a_i \in C,i = 1,\ldots,n\}$ (8)

and proof any theorems for the limits of *Mandelbrot* and *Julia sets* for (10) . We will follow the same *Definition 1* for the *Julia set* J_p and the *filled Julia set* F_p and the same definition as *Definition 2* for the *Mandelbrot set* for (8) as for (6). Let us now consider the equation:

$$z^n - |a_1|z^{n-1} - \ldots - (|a_{n-1}| + 1)z - |a_n| = 0,$$
$n \in N$ (9)

Theorem 7. The equation (9) has one unique positive root: $R > \sqrt[n-1]{1+|a_{n-1}|}$ *. If all coefficients* a_i*,i = 2,…,n of the equation (11) are different of 0 ,then the number of the negative roots of (11) is n-1 or n-3,or n-5,…or 1 (if n is odd) or 0 (if n is even).*

Proof. See the proof of *Decart'sTheorem* for the number of the posirtive and negative roots of the polynomials equations on the page 245 in (Kurosh ,1968). The fact that $R > \sqrt[n-1]{1+|a_{n-1}|}$ follows from the formula:

$$R(R^{n-1} - (1 + |a_{n-1}|)) = |a_1|R^{n-1} + \ldots + |a_n|$$

Now we will proof the following

Theorem 8. The Julia set for a member of the dynamical system(10) lies in the circle $|z|$ $\leq$ *R where* $R > \sqrt[n-1]{1+|a_{n-1}|}$ *is the unique positive root of the equation (9).*

Proof. The method of the proof is borrowed from Beardon and Rippon (1994) and it is similar to the proof of the *Theorem 3 and Theorem 4* .

Corollary. All other roots of the equation (9) lie in the domain: $|z| \leq R$,where R is the unique positive root of the equation (9).

We can make the following very important remark:we find:

$$f'(z) = nz^{n-1} + (n-1)a_1z^{n-2} + \ldots + a_{n-1} \tag{10}$$

and try to find the critical points of (10): that is to say it must solve the equation:

$$z^{n-1} + (1 - \frac{1}{n})a_1z^{n-2} + \ldots + \frac{a_{n-1}}{n} = 0 \tag{11}$$

But the equation (13) has the same type as the equation f(z) = 0 where f(z) is a polynomial of degree n. Then we must consider the equation:

$$z^{n-1} - (1 - \frac{1}{n})|a_1|z^{n-2} - \ldots \frac{|a_{n-1}|}{n} = 0 \tag{12}$$

It is clear that (12) is one equation of the same type as the equation (9) and it has an unique positive root R_1 too. We know the estimates for the positive roots of the equation (9) R and (12) R_1 from (Kurosh,p.234):

$$R \leq 1 + \max(|a_1|,\ldots,1 + |a_{n-1}|,|a_n|) = B(a_1,\ldots,a_n) \tag{13}$$

$$R_1 \leq 1+\max(|a_1|(1-\frac{1}{n}),\ldots,\frac{|a_{n-1}|}{n}) = B_1(a_1,\ldots,a_{n-1}) \tag{14}$$

It is clear that from (13) and (14) that $B_1(a_1,\ldots,a_{n-1}) \leq B(a_1,\ldots,a_n)$. We can show that the zeros of the polynomial (12) lie in the domain $|z| \leq R_1 \leq R$. Now we can consider the 2 dimencional- layers for the Mandelbrot set for the polynomials of degree n. The number of these layers is $\binom{2n}{2n-2}$ and they are: when a_1= Rea_1+$iIma_1$ is a pixel and the other a_i,i = 2,…n are constants; when a_2= Rea_2+$iIma_2$ is a pixel and the other a_i i =1,3,…,n are constants etc.

Also we have the following very important

Remark 1. The all 2D layers of the Mandelbrot set for (10) lie in the domain $|z| \leq B_0$,where B_0 satisfies :

$$B_0 = \max_{|a_i|\leq|a_i^0|,i=1,\ldots,n} B(a_1^0,\ldots,a_n^0)$$

It is very good known,that we can't resolve the equation f'(z) = 0 ,where f(z) is a polynomial of degree n>5 in the general case. The main idea is to go the opposite way. Instead of deriving the critical points from a polynomial,the great fellows start with the critical points and from these,they construct the parameters. Then there is to need for solving obscure equations and to use the formulas of Viete. Let we have: $f'(z) = (z - z_1)\ldots(z - z_{n-1})$, where $z_1,\ldots,z_{n-1}$ are the critical points. Then we can use the formulae of Viete,that is to say to calculate the coefficients of f'(z) as functions of the zeros of the polynomial f'(z).

$$f'(z) = (z - z_1)...(z - z_{n-1}) = z^{n-1} + a_1 z^{n-1} + \ldots + a_{n-2} z + \ldots a_{n-1}$$

where (Kurosh,p.151):

$$a_1 = -(z_1 + z_2 + \ldots + z_{n-1}),$$
$$a_2 = z_1 z_2 + z_1 z_3 + \ldots + z_{n-2} z_{n-1,\ldots,}$$
$$\ldots \qquad (15)$$
$$a_{n-1} = (-1)^{n-1} z_1 z_2 z_3 \ldots z_{n-1}$$

Now we can integrate:

$$f(z) = \frac{z^n}{n} + a_1 \frac{z^{n-1}}{n-1} + \ldots + a_{n-1} z + d$$

where d is the value f(0). Let us consider the equation:

$$z^{n-1} + a_1 z^{n-2} + \ldots + a_{n-2} z + \ldots a_{n-1} = 0,$$

where:$a_1,\ldots,a_{n-1}$ are done from the formulas (18). This equation is of the same type as the equation (13),also the roots $z_1,\ldots,z_{n-1}$ lie in the domain $|z| \le R_2$,where R_2 is the unique positive root of the equation: $z^{n-1} - |a_1| z^{n-1} - \ldots |a_{n-1}| = 0$. But we have the following estimate too:

$$R_2 \le 1 + \max(|a_1|,\ldots|a_{n-1}|).$$

Now we must consider the equation :

$$\frac{z^n}{n} - |a_1| \frac{z^{n-1}}{n-1} \ldots - (|a_{n-1}| + 1)z - |d| = 0 \qquad (16)$$

The equation (16) has one unique positive root (see *Decart'sTheorem* for the number of the positive roots of the polynomials equations on the page 245 in (Kurosh ,1968)). Now we can see where lie the Julia sets for the dynamical system:

$$\{C:f(z) = \frac{z^n}{n} + a_1 \frac{z^{n-1}}{n-1} + \ldots + a_{n-1} z + d\} \qquad (17)$$

Theorem 9. The Julia set for a member of the dynamical system(17) lies in the circle $|z| \le R$ where

$$|z| \le R_3 \le 1 + n \max \left(\frac{|a_1|}{n-1},\ldots |a_{n-1}| + 1, |d|\right)$$

is the unique positive root of the equation (9).

We can write the following very important

Remark 2. All 2D-layers of Mandelbrot set for the dynamical system (20) lie in the domain

$$|z| \le B_0 = 1 + n \max_{\substack{|a_i| \le |a_i^0|, i=1,\ldots,n-1, \\ |d| \le |d_0|}} \left(\frac{|a_1^0|}{n-1},\ldots |a_{n-1}^0| + 1, |d_0|\right)$$

2.5. The application of Mandelbrot and Julia sets theory for dynamical systems (17).

Now we consider the dynamical systems,where the critical points are:z_1=0.1, z_2=-0.1, $z_{3,4}$=0. Then the formulae for $f_q(z)$ are: **{C:$f_q(z)=z^5/5-0.01z^3/3+q$; q$\in$C}**. On Fig.1 we show the layer of the Mandelbrot set in this case,where (Req,Imq) is the pixel. On Fig.3 and Fig.4 we show the Julia sets in the cases: q=-1.328+0.88i and q=-1.264+0.88i. Let us now consider one other example,where the critical points are:z_1 = 1, z_2 = - 1,z_3 = i,z_4 = - i,this means that the dynamical systems are:**{C:$f_q(z)$ = $z^5/5$ - z + q;q$\in$C}.** On Fig.2 we show the other layer of the Mandelbrot set in this case,where (Req,Imq) is the pixel. On Fig.5 and Fig.6 we show the Julia sets for this systems in the cases:q=0 and q=1.2–0.8i.

Gratitudes. The author thanks of the heart to M. Ingvar Kullberg www.come.to/kullberg, who has made the computergraphical map of Fig.2,3 and 4.

3. Conclusions.

We consider and solve complementary the question about the limits of Mandelbrot sets for Julia sets for the polynomials of arbitrary order. This way it will be possible to make *computergraphical maps* for show the trajectories of the dynamical polynomial systems, using ESCAPE TIME ALGORITHM.

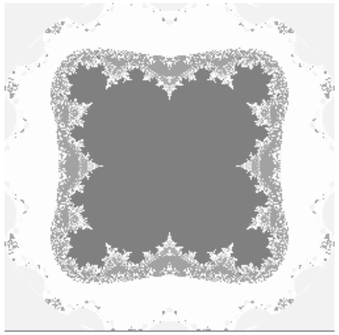


Figure 1

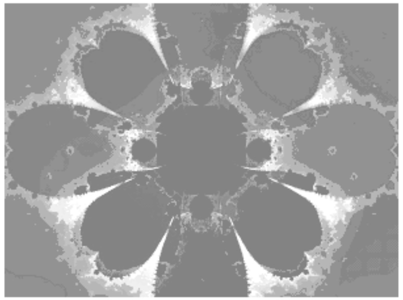


Figure 2

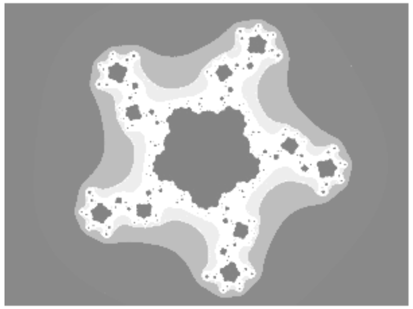


Figure 3

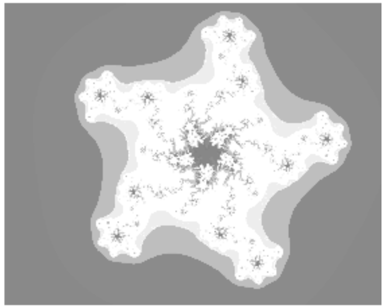


Figure 4

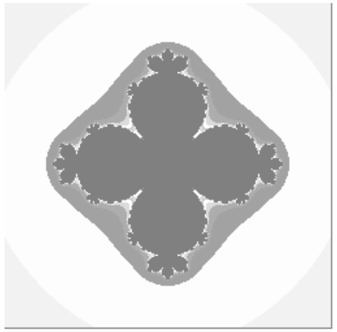


Figure 5

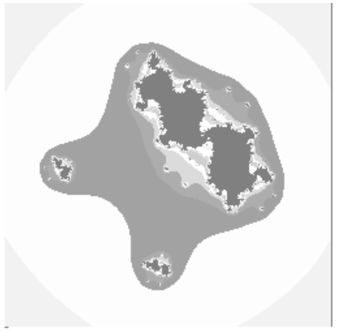


Figure 6

REFERENCES

1. Barnsley,M.F., Fractals Everywhere,Academic Press,Inc.,USA,1988.
2. Barnsley M. F., Lyman P.H., Fractal Image compression, AK Peters,Ltd.,1993.
3. Beardon A.F., Rippon P.J., A Remark on the shape of quadratic Julia sets, Nonlinearity 7 (1994), 1277-1280.
4. Brolin H., Invariant Sets Under Iteration of Rational Function, Arkiv For Matematik, Band 6 nr 6:103 to 144.
5. H.-O Peitgen, P.H.Richter, The Beauty of Fractals, Sptinger-Verlag, 1986.
6. Hastings H.M., Sugihara G., Fractals: A User's Guide for the Natural Sciences, Oxford University Press, 1993,1994.
7. Mandelbrot B., Les objets fractals: Forme, hassard et dimension, Paris:Flamarion,

1975,1984,1989.

8. Kurosh A., Cours on High Algebra, 1965, Sofia, (in Bulgarian).
9. Sy-Sang Liaw, Find the Mandelbrot-like Sets in any Mapping, Fractals, Vol.10, No.2 (2002), 137-146, © World Scientific Publishing Company.
10. R. L. Devaney and M. Morno Rocha, Geometry of the Antennas in the Mandelbrot Set, Fractals, Vol.10, No.1 (2002), 39-46, © World Scientific Publishing Company.
11. Prof. Peckham , Spring 2003, Math 1234, Chaos, Fractals, and Dynamics.

ON BULGARIAN TEXT-TO-SPEECH SYSTEM

G. Totkov, V. Angelova

Plovdiv University "Paisii Hilendarski", 24 Tsar Asen Str., Plovdiv 4000, Bulgaria,
totkov@pu.acad.bg, sinosuida@yahoo.com

Abstract: The paper solves a problem about the transformation of an input Bulgarian text into computerized sound output. For this purpose the necessary lexical resources and software tools are created. Different approaches to the transformation of Bulgarian texts to computerised speech are analyzed and corresponding algorithms are proposed.

Keywords: Computational Phonology, Text-to-Speech Synthesis, Allophone Database

1. Introduction

The foundations of the articulatory phonetics were laid down in ancient India in the year 800 B.C. when the mechanism in which the larynx produce sounds assimilation and the place and way of articulation are studied. Later – about 5th century B.C. the Stoics in Ancient Greece made some elementary observations on the speech and introduce the concept *vowel*, *consonant*, *syllable*, etc.

The interest in computational systems for the text-to-speech (TTS) transformation practically appears with the invention of the computers [3]. There are many market TTS systems: MITalk and Sproat (Bell Laboratories), Microsoft® .NET Speech SDK 1.0, QuickSig [5], etc.

Attempts for generating sound from computer-based Bulgarian text (CBT) were made more than 20 years ago (for PC Pravetc-8), and later – in single graduation and demonstrational works. These works remain on an experimental level because of the lack of solid lexical resources and software aids for computer-based processing of text in Bulgarian Language (BL). In 1995 the BABEL project [2] started (it's finished in 1998) with the main task – to make sound files corresponding to texts in many Eastern-European languages (including Bulgarian, Estonian, Hungarian, Polish and Romanian).

2. Main results

In order to convert Bulgarian texts to computerized speech we had to solve the following tasks:

T1. Creating software tools for performing *analysis and annotation* of CBT (for the purposes of the TTS transformation);

T2. Systematization of the rules for *phonetic transcription* of graphemes into phonemes (including pauses and stresses) and creation of a proper computer model;

T3. *Separation* of the transcribed text (the row of phonemes) to pronounceable elementary sound segments;

T4. Determining the elementary sound segments in the Bulgarian speech and *creating the corresponding sound database* (DB).

T5. Design and creation of *Bulgarian TTS system*.

Software tools (incl. segmentation, normalization, identification and morphological analysis of CBT) created in the Department of Computer Science of the Plovdiv University [6-9] are used for the solving of *Task 1*. At this preliminary stage it's not necessary to make a full morphological analysis. The size of the dictionary may be reduced by writing in it only the information necessary for the transcription (*annotation*), namely:

A. verb (some of the pronouns change from enclitics or accent words into clitics, depending on their position toward to the verb);

B. clitic (enclitic or postclitic);

C. abbreviation (pronounced depending on its sort);

D. occurrence of *a*: in the short form of the definite article after hard consonants, inflections marking 1st person singular and 3rd person plural for present verbs in 1st and 2nd conjugation after hard consonants;

E. occurrence of "*я*" in article morphemes for singular nouns in masculine gender *–я*, (*-ят*), in inflections for 1st person singular and 3rd person plural for present verbs from 1st and 2nd conjugation;

F. occurrence of *я* in forms used with the article in nouns (*ден*, *кон*, *крал*, *лакът*, *нокът*, *огън*, *път*, *сън*, *цар*; with basic form ending in *й* or with suffix *–ар* / *–тел*); in the forms of the adjectives, ordinal numerals and possessive masculine pronouns using the definite article; in the pronoun feminine forms *нея* and *я*; in the inflections of present verbs in 1st person singular and 3rd person plural after palatal consonants and after *й*.

The common-purpose morphological analyzer [8] based on a dictionary containing 1 174 607 Bulgarian word forms with placed stresses) can be used for partial annotation. However in our case another approach is preferred – the list of all word forms is preprocessed as follows: a) corresponding annotation is added to each word form (with placed stress); b) the preprocessed list of word forms is sorted in ascending order; c) the sorted list is presented by the trie-structure containing annotation in its leafs. We will note that the trie-structure can be minimized[13] twofold. It's not necessary to "record" the whole word in the structure but just its shortest

[13] It's a subject of another publication.

prefix, including the stressed vowel of the word. The shortest prefix should be different from every other prefix[14] of word form with annotation different from that of the treated word. As a result the dictionary's size (as a text file) decreases more than 20 times. The annotation and the placing of the stress are done by a number of operations and the greatest number of operations is equal to the length of the word (linear complexity). In cases when the word is "unknown" the stress may be "placed" incorrectly or there could be no hypothesis for the stress as well as for the annotation of the word. In the specific situation, in the cases when the place of the stress is not defined correctly it's accepted that its most probable place is on the "middle" vowel.

In order to solve *Task 2.* a number of Bulgarian phonetics rules [1, 4] (regressive assimilation, final devocalization, etc.) are rendered systematically. As a result a computer model of the phonetic transcriptions of graphemes into phonemes suitable for software realization is created. The final devocalization is performed first (voiced consonants at the end of the word turn to voiceless) and only after that – the assimilation of consonants into consonant groups. Only *в* behaves "ambiguously" – it doesn't influence the previous sounds but is influenced by the following ones.

Examples for transcription:

трезв → трезф →тресф (sober)

смарагд →смарагт → смаракт (emerald)

In Table 1 is displayed part of the *207 rules* used in our system for phonetic transcription. The transformation of each word into a sequence of phonemes is achieved through successive substitution of each grapheme (or a sequence of graphemes) with one or more phonemes starting with the final grapheme of the word and finishing with the first one. The substitution is done according to the context (grammar characteristics of the word, the previous and following character). At every following step the phonemes received at the previous step are also used (assimilation is chain-like: извпръсквам (besprinkle)→ [исфпръсквам]). After[15] the phonetic transcription the vowels in the word are reduced according to their position in relation of the word stress [1, 4].

Grapheme	**Phoneme**			**Annotation**
	Current	**Previous**	**Next**	
а	а	$	PC	–
а	а	$	HC	–
а	а	HC	#	
а	ъ	HC	#	D
....	....	....	....	....
я	‘а	й	–	–
я	‘а	SC	–	–
я	йа	$	–	–
я	йъ	VL	#	E
я	‘ъ	PC	#	F
я	‘ъ	PC	HC	F
Punct. marks	|| [16]	–	–	–
Full stop	! [17]	–	–	–

Table 1. 15 (of 207) rules for transformation "graphemes-phonemes"

The following abbreviations are used:

$ – word beginning
# – word end
UC – unvoiced consonant [п|ф|к|т|ш|с|ц|ч|х|щ]
VL – vowel [а|и|е|о|у|ъ]
VC – voiced consonant [б|в|г|д|ж|з]
PC – palatal consonant, followed by [ю|я|ь, о]; is designated by stress (`): [б'|п'|в'|ф'|д'|т'|з'|с'|г'|к'|ц'|дз'|х'|р'|л'|м'|н']
PS – pause || or ![18]
PC – conditional pause (if a clitic or a sentence end is following)
NC – nasal consonant [р|л|м|н|й]
HC – hard consonant [б|п|в|ф|д|т|з|с|г|к|ч|ц|дж|дз|х|р|л|м|н|ж|ш|х]

There are different possible variants for choice of elementary sound segments in the solution of *Task 3* (division of words into pronounceable units). From the articulation point of view it's not possible to pronounce a following segment without the relevant preparation still during the pronunciation of the previous one. In this sense the speech can be understood as comprised by sound segments and transitional sections between them. The segments (the speech units) may have different size – from allophones, diphones, and triphones till syllables [10]. The natural solution of the problem how to create a sound database lies in the two extreme cases – allophones (phonemes for which we know the following phoneme) and the syllables in Bulgarian Language (BL). Diphones are in an intermediate position – sound units with length from the middle of a sound to the middle of the next sound. The distinct defining of the boundary between the two sounds however creates certain problems in realizing a diphone database. A natural question is how

[14] In comparing prefixes the stressed and unstressed variants of one and the same vowel are considered different.

[15] The reduction can be performed parallel with the transcription.

[16] Average-lasting pause

[17] Long-lasting pause

[18] In this case we are not interested in the characteristics of the next word

many the syllables and diphones in the BL are. In order to find the answer we can use the dictionary mentioned above (containing 1 174 607 Bulgarian word forms) as a source[19] providing all syllables and allophones in BL. Research has proved that the dictionary contains 9800 syllables[20] and 1500 allophones. The realization of each of the two variants has its advantages and shortcomings. In the text-to-speech transformation of the syllable database the pronunciation of the text is more distinct but the number of required sound files is five times bigger and this hampers the gathering of the sound library. Mixed variants of sound library are also possible; this library should contain all allophones and part of the syllables (for example – the most commonly used).

The decision of the *Task 4 (creation of sound DB)* depends on the choice of variant. For its solving it's necessary: a) to select a minimal number of Bulgarian words which after phonetic transcription and separation should contain all elementary sound segments necessary for the creating of the sound DB; b) these words should be pronounced and recorded in the form of sound files. c) from the sound files received in that way elementary sound segments should be selected and recorded into the sound database.

In the specific realization the option for creation of an allophone sound database is chosen. First all 1 174 607 words are phonetically transcribed and divided into allophones (for example the word "кол" (stake) is divided into 3 allophones: к^{о}, о^{л} and л$^{\#}$[21]. After that the entire set of the different allophones is defined. For the creation of a sound library containing all the allophones, they should be isolated and extracted from a relevant set of words (chosen, pronounced and recorded as sound files). An audio editor[22] can perform the segmentation of a sound file corresponding to a specific word. Since it's technically easier to separate the allophone at the beginning of the word it's natural to start with a set of one-letter and two-letter words with which the beginning of the creation of a sound allophone database should start. At the next step it's necessary to choose a word from the dictionary (as short as possible) that should serve for the forming of the corresponding sound file for each allophone that is not segmented in the directed way. If such word doesn't exist, a "remainder" of an earlier segmented sound file is used or ultimately a specially pronounced word containing this allophone. The sound files recorded in the database are in PCM [23] format (a common wav-file that can be listened to from different kind of media). The same approach can be used in creating sound library of the syllables in BL (a subject of following development).

[19] The rules for separating words into syllables are published in another work.

[20] The number of syllables in English is estimated to be about 10000.

[21] The sign # means end of the word.

[22] The experiments are performed with the sound editor Sound Forge.

[23] These conditions are not restrictive and they can be changed.

The main modules and elements of the Bulgarian TTS system solving *Task 5.* are presented at figure 1. The left part of the diagram has no pretences for originality – it describes the main phases of analysis and synthesis in the solution of the problem for the text-to-speech transformation of CBT. An important element is the module used for "concatenation" of sound files, corresponding to the successive syllables (or allophones).

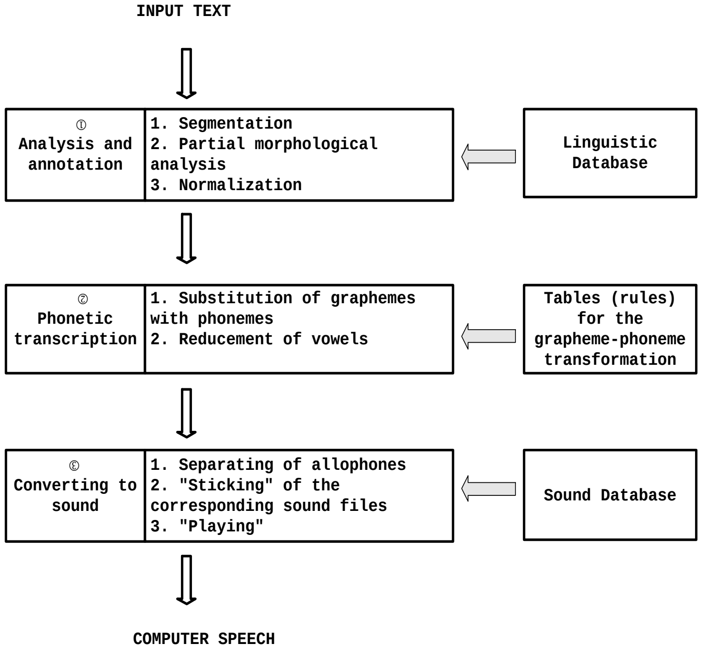

INPUT TEXT
①
Analysis and annotation
1. Segmentation
2. Partial morphological analysis
3. Normalization
Linguistic Database
②
Phonetic transcription
1. Substitution of graphemes with phonemes
2. Reducement of vowels
Tables (rules) for the grapheme-phoneme transformation
③
Converting to sound
1. Separating of allophones
2. "Sticking" of the corresponding sound files
3. "Playing"
Sound Database
COMPUTER SPEECH

Figure 1. The Bulgarian TTS system

3. Future development

The realization of an approach for creating sound DB based on "syllable" principle is possible. Part of the syllables is recorded (only 3500 out of 9800 syllables have frequency more than 0.000033) and the others are converted to sound files using the already created allophone sound DB. Another interesting problem is the automatic separation of a sound file into elementary pronounceable units. Solutions to this task are offered but it's still not clear if these models are suitable for BL. The questions for modeling the intonation and the pronunciation in text-to-speech transformation for the Bulgarian language are open just as the questions for the processing and converting into text form more complex (compound) text elements such as mathematical expressions and formulas.

REFERENCES

1. Boayadjiev T., D. Tilkov, Phonetics of Bulgarian literary language. V. Tarnovo, 1999 (in Bulgarian).
2. Dimitrova, S., T. Kostadinova, E.Grogorova, G. Rouzhekov, The Bulgarian BABEL Database. RANLP'2001. Tzigov Chark, 2001, 264-266.
3. Dutoit T., An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers, Dordrecht, 1997.
4. Grammar of Contemporary Bulgarian Language, vol. I. (Phonetics), Printed by BAS, Sofia, 1983 (in Bulgarian).
5. Meister, E., A. Eek, T. Altosaar, M.Vainio, The Estonian Phonetic Database in the QuickSig Object-Oriented Environment. in A. Narin'iyani (eds.), Comp. Ling. and its Applications, Proc. of the Int. Workshop DIALOGUE'1999, 347-353.1999.
6. Totkov Г., Robust Analysis of Bulgarian Texts and Development of a Linguistic Processor, Mathematics and mathematical education, 19 th conference of UMB, Slanchev Briag, 6th-9th April 1990, 295- 302 (in Bulgarian).
7. Totkov G., Development of a Linguistic Processor: Problems, Results, Perspectives, Mathematics and Mathematical Education, 20th Conference of UMB, Sofia, 2th-5th April 1991, 43-50 (in Bulgarian).
8. Totkov G., Resources and Tools for Computerization of Bulgarian Language (1988-2000). Artificial Intelligence, No.3, 2000, 573-577.
9. Totкov G., Ch. Tanev, LINGUA – an Architecture for Robust Text Processing in Bulgarian. in A. Narin'iyani (eds.), Comp. Ling. and its Applications, Proc. of the Int. Workshop DIALOGUE'2002, Protvino, 6-11 June 2002, 582-589.
10. Zaharov L., Problems in Creating Allophone Database for Automatic Speech Synthesis, DIALOGUE'2000 (in Russian).

ON A NEW APPROACH FOR PROXIMITY AND CLASSIFICATION OF TEXT DOCUMENTS

Krastyu Gumnerov, Stefan Koynov

Institute of Information Technologies – BAS
gumnerov@iinf.bas.bg
baruch@iinf.bas.bg

This proceeding is realized with the support of NFSR-MOS contract No I-1005/2000

Abstract: A new context-oriented approach is introduced and grounded for proximity between text documents. It is based on the repeated co-occurrences of words in texts for some topic.

Keywords: text retrieval, intelligent searching, context, statistical techniques.

1. Introduction

A great amount of the contemporary searching systems are based on well known methods and algorithms which are developed before the origin of Internet. The general problem of the information search did not change: it is to assist the user to find the necessary information. During the last years statistical methods were developed. They introduce a proper metrics (a proximity criterion) in a set of texts based on the frequency of the comprising words and expressions. The simplified vector models traditionally are used for presenting any text as a set of the comprising words and word expressions [8, 13, 22]. This allows the search request to be formulated not only by keywords and logical conjunctions but also by an introduced or selected text. This was the way to bypass the complex problem of defining a short list of words; it was replaced by formulations of examples of interesting for the human texts instead. In addition the statistical methods amplified the adequate search by keywords and logical conjunctions. An abstract description of all these methods is given in [21].

The possibility to elevate the quality of the models is tied with the correlativity of meeting the words and their word expressions s in the text; it is determined by the existing semantic links between them. The offered in the paper approach for determining the proximity between documents is an alternative to the conventional methods for proximity between documents. It is based on the following two mutually correlated heuristics:

1. Existence of a link (correlation) between sets of words (or their word expressions) in texts on a given topic. I. e. their frequent appearance together in a window in the texts for the topic. For example if the topic is "production of wines" then it is very probable that in the documents on the topic besides the word "wine" its

environment will include the words “vineyard”, “working”, “sort”, etc.; if the topic is “stock exchange” then besides the word “exchange” it is very probable that its environment will include the words “money”, “markets”, “sales”, “shares”, etc.;

2. The processing is based not on single words or text word expressions, but on their meanings. This means that the proximity will be automatically extracted from the concrete presentable extract from the texts on the given topic and duly building a thesaurus on the topic based on 1.

Therefore, based on the context (the environment) for every word or a expression of words, associative nests (ANs) are created that determine the degree of associativeness of the concrete word (or word expression) with all the other words (or word expressions) from the subject area.

2. Grounds for The Suggested Approach

The introduced approach is a realization of the following axioms which have become a folklore for linguists, philologists and psychologists:

-1. The understanding of the sense of a concrete text is based on the determination of the aggregate of texts where the concrete text must be perceived;

-2. The text semantics is determined by the associative links of the text concepts;

-3. A single concept (meaning) can be expressed by different bearers.

We introduce the concept of a language unit (LU) according to the recognition of word expressions or only of separate words. The approach is realized according to the following scheme:

1) We use the context (the LU environment) for every LU to create an AN for the concrete LU. The AN are lists of the frequencies of meeting LU from the domain of the window of the concrete LU. They are obtained by measuring the frequency of meeting the every LU in every window of the fixed LU (for which we construct the AN) in all the texts of the domain. It is based on the research in psycholinguistics where there are evidences that the contextual similarity plays an important role in the human semantic categorization, see [10] where Miller & Charless find proofs in several experiments that humans determine the semantic proximity of words using the proximity of the contexts. The references contain some models of operation with the context of LUs presenting them as vectors in the space of LUs such as: based on glossary members [18]; manual coding of semantic features [5]; based on slight parsing [7]; comparisons of context vectors based on glossaries and corpses [11]; finding a discriminant indicator used to establish the meaning of words with multiple meanings [2], etc. All of them require external knowledge and a manual operation of a linguist, though. Our context work is automatic.

2) A thesaurus of the domain is built by an appropriate clusterization of the contexts – ANs based on a suggested metrics (proximity) in the set of ANs. In this way we reduce several decades the number of LUs which we shall use and consequently the dimensionality of the LUs space. This leads to a speed up of the algorithms that measure the proximity of the texts. The obtained clusters respond for the meanings of the comprising them LUs. A good review of the contemporary methods to clusterize text documents is given in [20]. Other works on clusterization of words are [1, 6] and clusterization of documents [19].

3) The processing is based not on the concrete LUs but on representatives of the clusters with their ANs (their meanings). We measure the frequency of meeting the representatives of the clusters for every document. Then we build ANs of the representatives on the document level. So we obtain the internal representation of every document as comprised of: a vector of the frequencies of meeting the representatives of the document clusters; an AN for every vector component (on the document level).

4) Convenient metrics (proximities) are defined in the set of documents. Now the documents are the vectors from 3). The proposed metrics are based on the frequency of meeting the meanings of LUs and their contexts – ANs of these meanings.

3. Description of the Methods Realizing The Approach

3.1. Using The Context to Create an AN

Step 0: We choose a representative set of documents that define the domain. Then we connect all texts together as a single document. After that we define the length of the context (of the LUs window) and it is a number of κ LUs before and a number of l LUs after the fixed LU (the one upon which we shall build AN). It is recommended that κ and l should be equal to 25 (or that the window width should be 50 LUs); no significant improvement of the operative quality is established from the tests cited in [16] for windows with bigger width.

Every LU is processed in the following sequence:

Step 1: We fix a LU and we call it the *head* of the AN. Then we measure the frequency of meeting every LU in all windows of the fixed head. We name the met LUs in some window *members* of AN.

Step 2: We determine the coefficient of representativeness (CR) for the head of AN being built. CR is the frequency of meeting the LU being processed, multiplied by the weight for the place in the document which can be the headline, the beginning of the document, a headline of an item, an abstract, a definition or some other important part of the document; all this is multiplied by a coefficient of importance for the domain (determined by the frequency of LU for the domain in a ratio with the LU frequency for the language). More details about CR are available in [21].

Step 3: We use the vector space model from [8, 13] to introduce AN. Along its dimensions we project the values from Step 1 and an additional dimension for the CR of the head. So our AN will be projected as a point in the LUs vector space. We can represent the AN in an alternative way as a set of its members arranged (rated) according to the value of their frequency (rating) of meeting in the windows of the head.

3.2. Determination of The Candidates for Stop Words.

Once having built ANs of all words in the domain, we select the ones with a big number of members but which are with a low rating (e.g. < 0,01 %). In other words these points are from the vector space of the words with a big value along the dimension of the head and with a low value along the dimensions of its members. I.e. if we restrict the vector space except the dimension of the head then we have the points in a *n*-dimensional sphere (to be more exact, an *n*-dimensional cone in the first *n*-dimensional quadrant) with some small radius *r* (*r* < 0,01%). The heads of these ANs shall be candidates for stop words or the words with a low "information value". Such are the words with a general usage like the prepositions, the pronominals, interjections, etc.

3.3. Automatic Building A Thesaurus of The Domain.

Step 1: We convert the set of ANs in a metric space by introducing (defining) a metrics in the set of ANs in the following way:

Let us have two ANs: I – first AN and II – second AN. Let also:

A is the proximity between I and II ANs;

$\alpha_i^{`}$ is the CR (the frequency) of the *i*-th same (in the I-II couple) member in the I AN or of the very head;

$\alpha_i^{``}$ is the CR (the frequency) of the *i*-th same (in the I-II couple) member in the II AN or of the very head;

m is the number of identical words in the both ANs. Then:

A = $(\sum_{i=1}^{m} (\alpha_i^{`} - \alpha_i^{``})^2)^{1/2}$ - is the Euclidian distance.

Step 2: We clusterize the set of ANs based on the metrics from Step 1 by a convenient clusterizing method [3, 20]. The choice of the method will depend on the nature of the subject domain.

Every cluster corresponds to a definite meaning of LUs (the heads of ANs) which belong to it. So the thesaurus of the domain will represent:

- A sequence of the representatives of the clusters (e.g. their centroids);
- A pointer to the number of the clusters with their distances between them;
- A pointer to the number and the names of LUs in every cluster.

3.4. Proximity of Documents.

Step 1: We fulfill the following algorithm for every document. We measure the frequency of meeting the representatives of the clusters multiplied by a weight for the place in the document (the headline, in the beginning of the document, in the beginning of an item, an abstract, a definition or some other important part of the document). All this we multiply by a coefficient of importance for the domain (it is determined by the LU frequency for the domain in a ratio with the LU frequency for the language). In this way we determine CR of the representatives of the clusters from the thesaurus in the document.

Step 2: We build the AN of the clusters representatives on the document level. With other words we do 3.1 on the document level.

So the internal representation of every document will be:

- A vector of CR for the representatives of the clusters in the document; and
- An AN for every component of the vector on the document level.

Step 3: We introduce the following definition for a proximity between two documents:

Let we have two documents: I and II. Let also:

B is the proximity between I and II documents;

$\beta_i^{`}$ is CR for the *i*-th same (in the I-II couple) representative in the I document;

$\beta_i^{``}$ is CR for the *i*-th same (in the I-II couple) representative in the II document;

μ_i is proximity defined in 3.3 between ANs of the *i*-th same (in the I-II couple) representative of I and II documents;

n is the number of identical LUs in the both documents; then:

$$B = \sum_{i=1}^{n} \mu_i \cdot \left(\beta_i^{'} + \beta_i^{''}\right) / 2 .$$

The following lemma holds:

<u>*Lemma*</u>: Let Q is the field of the rational numbers, x = (x_1, x_2, ..., x_n) $\in$ Q^n and y = (y_1, y_2, ..., y_n) $\in$ Q^n. If μ is metrics in Q^1 and c_1, ..., c_n are arbitrary fixed positive numbers $\in Q$ then the function ζ(x,y) = $c_1 \mu$ (x_1, y_1) + c_2 μ (x_2, y_2) + … + c_n μ (x_n , y_n) is metrics in Q^n.

Proof: ζ (x,y) will be metrics if we prove that it satisfies the following conditions:

1). ζ (x,y) $\geq$ 0, $\forall$ x, y $\in$ Q^n, ζ (x,y) = 0 $\Leftrightarrow$ x = y ;

2). ζ (x,y) = ζ (y,x) – symmetry;

3). ζ (x,y) + ζ (y,z) $\geq$ ζ (x,z) – inequality of the triangle.

Checking 1):

By definition $\forall i$ μ is metrics $\Rightarrow$ $\mu(x_i, y_i) \geq 0$, besides $c_i > 0 \Rightarrow$ $\zeta(x,y) = c_1 \mu(x_1, y_1) + c_2 \mu(x_2, y_2) + \ldots + c_n \mu(x_n, y_n) \geq 0$.

Let $\zeta(x,y) = 0$, i.e. $c_1 \mu(x_1, y_1) + c_2 \mu(x_2, y_2) + \ldots + c_n \mu(x_n, y_n) = 0 \Rightarrow \mu(x_i, y_i) = 0 \Rightarrow x_i = y_i$, $\forall$ i. It is equivalent to x = y.

Checking 2):

$\zeta(x,y) = c_1 \mu(x_1, y_1) + c_2 \mu(x_2, y_2) + \ldots + c_n \mu(x_n, y_n) =$
$c_1 \mu(y_1, x_1) + c_2 \mu(y_2, x_2) + \ldots + c_n \mu(y_n, x_n) = \zeta(y,x)$

Checking 3):

$\zeta(x,y) = c_1 \mu(x_1, y_1) + c_2 \mu(x_2, y_2) + \ldots + c_n \mu(x_n, y_n)$
$\zeta(y,z) = c_1 \mu(y_1, z_1) + c_2 \mu(y_2, z_2) + \ldots + c_n \mu(y_n, z_n)$

Consequently:

$\zeta(x,y) + \zeta(y,z) = c_1(\mu(x_1, y_1) + \mu(y_1, z_1)) + \ldots + c_n(\mu(x_n, y_n) + \mu(y_n, z_n))$, but $\forall$ i, $\mu(x_i, y_i) + \mu(y_i, z_i) \geq \mu(x_i, z_i)$, because μ is metrics $\Rightarrow \zeta(x, y) + \zeta(y, z) \geq c_1 \mu(x_1, z_1) + c_2 \mu(x_2, z_2) + \ldots + c_n \mu(x_n, z_n) = \zeta(x, z)$, $\Rightarrow \zeta(x, y) + \zeta(y, z) \geq \zeta(x, z)$

So we checked the three conditions for metrics and therefore we proved the lemma.

The following corollary holds:

Corollary: The defined in Step 3 in 3.4 proximity is metrics.

Proof: It follows directly from the lemma above.

4. Application of The Approach

Besides its main purpose this approach is capable of other applications: for example an amplification of the quality of operation in the searching algorithms by keywords and logical conjunctions.

4.1. Amplification of Algorithms for Searching by A Keyword (1st Version).

Step 0: We must have the clusters of ANs for the words in different domains.

Step 1: The user chooses some topic for searching by a keyword. After that he inputs the search request.

Step 2: The search starts using not only the input word but also all other words from the cluster where it resides. When a word from the cluster is met then the interpretation is that the word is found which was searched for.

According to this model as a response to the search request there can be found documents that do not contain the object word but which contain the words from its cluster, words with a near meaning.

4.2. Amplification of Algorithms for Searching by A Keyword (2nd Version).

Step 0: We must have the ANs of the words from different domains (topics). Besides every document must be represented by a sequence of ANs of its words on the document level.

Step 1: The user chooses some topic to search by a keyword. Then he input the search request.

Step 2: A comparison is started between the ANs of the domain with the ANs of the documents the heads of which are the object word. The priority to output documents is based on the ANs of documents which are nearest to the ANs of the domain for the word being searched.

An illustration of a similar approach is the *inter*Media Text as an ingredient of the Oracle8i database, one of the most powerful products allowing the realization and the support of full-text databases via an Internet access. The thematic classification in *inter*Media Text is combined with the searching possibilities available in the process of operation with relational data bases and in particular in the process of writing applications for text processing it is possible to use SQL with a developed language for requests for full-text information.

The basis of the *inter*Media Text technology is designed upon the usage of a semantic vocabulary of the English language classified on topical categories. The usage of the thesaurus may be helpful during searching relevant documents without any requests to expand the list of words. The thesaurus allows *inter*Media Text to fulfill a thematic analysis of texts in the English language. The reference of every word in the text to the corresponding partitions in the thesaurus and the monitoring the frequency of their meeting allows *inter*Media Text to fix several main topics in the document. The topical classification of the documents may be very helpful during the searching process, for example in the case when the user is hampered to select the exact keywords for the description of his interests.

5. Operation of The System That Realized the Described Approach

The operation of the system goes through the following steps:

5.1. Initial Training of The System

The purpose of this phase is that the system must be supplied with an initial knowledge and information which shall help it to start functioning in a real environment. The initial training is carried out assisted by an expert according to the following scheme:

Step 1: The expert suggests a representative set of documents that define the domain. Such set corresponds to the definition of a point of view which shall be used to assess all other texts.

Step 2: A model of the domain is constructed in the following way:

1) ANs of LUs for the domain are built using 3.1.

2) The expert defines the stop words using 3.2 and they are excluded from ANs and the texts. Then the ratings of the ANs members are rebalanced.

3) A thesaurus of LUs from the domain is created using 3.3. In this way the internal representation of every document shall be:

- A vector of CR for the representatives of the clusters in the document; and

- ANs for every component of the vector (on the document level).

4) The set of documents is transformed to a metric space based on the metrics from 3.4.

5) This metric space is clusterized in an appropriate way. In this way the expert defines the subdomains.

The general structure of the model for the domain comprises:

- A vector of CR for the representatives of the clusters in the domain;
- A sequence of ANs for the representatives of the clusters for every component of the vector;
- A pointer of the number of the subdomains with their distances to the domain (this is the distance according to 3.4 between the domain and the subdomains as distances between "generalized" documents) and also the distance between them;
- A pointer of the number and the names (the numbers) of the documents in every subdomain;

Step 3: The previous Step 2, fulfilled for every subdomain.

5.2. Operation of The System in A Real Operational Environment

When a new document is received then it is processed by the system, i.e. it is modeled (an internal representation of the document is made). Based on this model it is attached to the nearest subdomain (based on the metrics from 3.4).

In the case of proximity of documents a search in the nearest subdomain is started which is followed by searches in the other closer subdomains. If a fixed number of documents must be extracted from the subdomains according to their proximity with the request then the more remote the subdomain is from the request, the smaller number of documents will be extracted from it. This is the way to optimize the speed of the searches.

If the number of new documents in the system exceeds some percentage then the previous Steps 2 and 3 are done recursively.

6. Conclusion

Bearing in mind the research carried out by Schutze & Pedersen in [14] we can

expect an improvement in the operation of the algorithms based on the presented in this paper approach in comparison with the models operating just with LUs but without the values of LUs. In [14] the text corpse is a collection of TREC-1 (about 170 000 documents from *Wall Street Journal*). Responding to its inquiries (see pp. 51-75 in [8]), the extraction based on meanings improves the average preciseness (the definitions of the criteria preciseness and completeness are given in [21]) with about 7,4 %, compared with the extraction based on words (LUs). The combination based on words and meanings improves the quality with 14,4 %.

An expected drawback of the approach is the case when it will not adequately recognize homonyms. Consequently it must be applied to subject domains with almost no homonyms. Still it is expected that homonyms will not be a frequent phenomenon but more likely exceptions. This problem can be overcome by models that count for the unique representations of LUs. Methods for an automatic recognition of meanings of polyvalent words see [9, 15, 23].

REFERENCES

1. Brown, Peter F., Vincent Della Pietra, Peter deSouza, Jenifer Lai & Robert Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics, 18(4), 467-479*
2. Church, Kenneth & William Gale. 1995. Poisson mixtures. *Journal of Natural Language Engineering, 1(2): 163-190.*
3. Cutting, D., J. Pedersen, D. Karger & J. Tukey. 1992. Scater/Gater: A cluster-based approach to browsing large document collections. In proceedings of SIGIR`92, 318-329, Copenhagen, Denmark.
4. Finch S.P. 1993. Finding Structure in Language. PhD thesis. University of Edinburgh.
5. Gallant, Stephen I. 1991. A practical approach for representing context and for performing word sense disambiguation using neural networks. *Neural Computation, 3(3): 293-309.*
6. Grefenstette, Gregory. 1994. Corpus-derived first, second and third-order word affinities. *In proceedings of the Sixth Euralex International Congress, Amsterdam.*
7. Grefenstette, Gregory. 1994. Explorations in Automatic Thesaurus Discovery. Cluwer Academeic Press, Boston.
8. Harman, D. K. editor "TREC-1", Rockville, MD, USA, 1993.
9. Hearst, M. & Ch. Plant.1993. Subtopic structuring for full-length document access. In *Proceedings of SIGIR'93, 59-68.*
10. Miller, George & Walter Charless. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6(1): 1-28.*
11. Niwa, Yoshiki & Yoshiko Nitta. 1994. Co-occurrence vectors from dictionaries. *In Proceedings of COLING94, 304-309.*
12. Ruge, Gerda. 1992. Experiments on linguistically-based term associations. Information Processing & Management, 28(3), 288-297.
13. Salton, Gerard & Michael J. Mc Giil, "An introduction to modern information retrieval ",

New York: McGraw-Hill, 1983.

14. Schutze, H. & J. Pedersen. 1995. Information Retrieval based on word senses. *In Procedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval. 161-175, Las Vegas, NV.*
15. Schutze, Hinrich. 1992. Dimensions of meaning. In *Procedings of Supercomputing'92, 787-796, Mineapolis, MN.*
16. Schutze, Hinrich. 1997. Ambiguity Resolution in Language Learning. CSLI Publikations, Stanford, CA.
17. Van Rijsbergen, C.J. 1979. Information Retrieval. Second edition. Butterworths, London.
18. Wilks, Yorick A., Dan C.Fass, Cheng Ming Guo, James E.McDonald, Tony Plate & Brian Slator. 1990. Providing machine tractable dictionary tools. *Journal of Computers and Translation, 2.*
19. Willet, P. 1988. Recent trends in hierarhich document clustering: A critical review. Information Processing & Management, 24(5), 577-597.
20. Cyrichenko, K.M., Gerassimov, M.B., 2001, A Review of The Methods for Clusterization of Text Information, Dialogue 21, V. 2 (in Russian)
21. Stoyanov, St. & Kr. Gumnerov, 2002, Models for Intelligent Searching with Applications in the E-Commerce, "27th International Conference Information and Communication Technologies and Programming (ICT&P'02)"*, Proceedings, 157-169* (in Bulgarian).
22. Bellot, E., M. Crestan, L. El-Buze, L. Gillard, E. Crestan, C. de Loupy, 2002. "Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC 11 Question Answering Track", *TREC 2002.*
23. Stokoe, C.M., J. Tait, 2002. "Automated Word Sense Disambiguation for Internet Information Retrieval", *TREC 2002 Web Track.*

METADATA TAGGING AND INTERACTIVE MULTIMEDIA CONTENT REUSABILITY IN WEB-BASED LEARNING SYSTEMS

Maria Zheleva[1], Radoslav Pavlov[2]

1 - Bourgas Free University, 101 Alexandrovska str., Bourgas 8000, Bulgaria
E-mail: mariaj@abv.bg
2 - IMI, BAS, Acad. Bonchev str., Bl. 8, Sofia 1113, Bulgaria
e-mail: radko@cc.bas.bg

Abstract: There exists 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. There is growing consensus around object-based approach to constructing content for online delivery. Learning Objects(LO) refer to self-contained chunks of training content that can be assembled with other Learning Objects to create courses and curricula. Learning Objects are designed to be reused in multiple learning contexts, aim to increase the flexibility of training, and make updating courses much easier to manage. Reusing of interactive multimedia LOs is hindered by the lack of mechanisms for describing and publishing data sources, and for discovering the objects relevant to a specific problem.

Keywords: Learning Object (LO), Learning Object Reusability, LOM, XML, Metadata Tagging.

Introduction

The World Wide Web has become the preferred medium for the dissemination of information in every domain of activity.

The delivery of interactive multimedia educational content via Web-browser over the public Internet or a private intranet is a widespread e-learning solution.

The WWW is a system for sharing documents published and accessed via HTTP protocol and HTML document format standards. Recently XML [1] and XML Schema [2] have been widely adopted as a data exchange format and a model for exchanging structured data. WWW sources may cover different domains and they may considerably differ with respect to the domain metadata about their contents. The quality of data metadata that describe contents of the sources may also be different. Criteria for assessment of the quality of data in a source have been proposed in the literature [3-5].

To solve the problem of selection of appropriate LOs using quality parameters, some additional metadata have to be maintained.

The article is organized as follows: In Section 2 the main aspects of the object-based development of training content are briefly discussed. Some questions about LOs reusability are also presented. The advantages of using metadata are given in Section 3. Section 4 describes some problems of traditional way of tagging by standardized metadata scheme (such as the IEEE-LOM scheme) with respect to LOs selection and reusing in different learning contexts. A conceptual model of IEEE-LOM metadata set extension for effective reusing is presented in Section 5. Section 6 concludes.

Object-based development of educational content

Described variously as learning objects, reusable learning objects, or content objects, the concept is based on chunking the learning content into reusable components and building a flexible hierarchy so that various instructional sequences to be created [6-8]. Update a part of a learning object and the change should appear in any course using that Learning Object.

Learning objects are most effective when organized by a metadata classification system, stored in a data repository and managed by a learning content management system (LCMS). LCMS is an environment where developers can create, store, reuse, manage and deliver learning content from a central object repository, usually a database. Developers can store the learning objects metadata in repository, separate but linked from the learning content or together with the resource. The purpose of metadata is to provide common means of describing objects characteristics and objectives. Using the metadata facilitates searching, retrieving and managing of the LOs.

The learning objects originate from various sources and include various data types. These objects could also include interactive activities, test questions, role games. These objects can process input parameters, generate output parameters, and work internally with data which cannot be described by traditional metadata sets. This is the reason for which the achievement of high level of reusability with regard to such kind of dynamic learning objects is a real challenge.

The question about the object reusability has two aspects [7, 8]:

- *Inter-contextual reuse* – Addresses the object potential for reuse in different content areas or domains, that is, the number of different learning contexts in which the learning object might be used.
- *Intra-contextual reuse* – Addresses the number of times the learning object could be reused within the same content area or domain, that is, the number of different behaviors or situations (scenarios) the object could be tagged.

Online collections of structured data often comply with interoperability standards and even may conform to a common semantics – each component learning object is precisely defined. However, the sharing of learning content is still a very difficult process. Often educational software is not used because what is available is irrelevant to concrete curricular goals. The reason is the lack of mechanisms for LOs customization and adaptation to the users' needs and requirements.

Metadata

Metadata are descriptive (machine readable and machine understandable) information about resources which facilitates the resource finding, managing, and effective using. Metadata allow the learning resource to be tagged with searchable attributes or properties (such as author, publisher, keywords, version, language, learning objectives, etc.) Using of standardized metadata allows organizations to tag, store, and retrieve learning content resources in their own repositories and those of their external, third-party content suppliers over the Internet. The Dublin Core Metadata Element Set [9], Educom's Instructional Management System [10], the Alliance of Remote Instructional Authoring and Distribution Network for Europe [11], and IEEE's Learning Object Metadata Working Group [12] are the most important initiatives dealing with metadata. These initiatives closely relate to the Resource Description Framework (RDF) and otherworld Wide Web Consortium activities.

According IEEE-LTSC the main advantages of using the metadata classification scheme are as follow [12]:

- Possibility of the development of learning objects in units (which are also learning objects) that can be searched, combined, and decomposed in meaningful ways.
- Make easier dynamically personalized lessons to be composed.
- Provide consistent way for users/learners to locate the information they want
- Organizations are enabled to express educational content and performance standards in a standardized format that is independent of the content itself.
- Standard that is simple, yet extensible to multiple domains and jurisdictions so as to be most easily and broadly adopted and applied.
- Support necessary security and authentication for the distribution and use of learning objects.

Some problems in metadata tagging and objects reusing

Metadata provide proper descriptions for the static resources but the dynamic resources are described to some extent. This is one of the big disadvantages of the metadata standards. Reusing of learning objects is hindered by the lack of mechanisms for describing and publishing data sources, and for discovering the objects relevant to a specific problem. Another key problem in developing LOs is planing for change [13] in terms of flexibility and adaptability to the new, changing user requirements.

The traditional metadata schemes cannot describe the behavior of the dynamic objects. The reason is that the dynamic multimedia objects can process input parameters, work internally with data, and produce output parameters.

Metadata descriptors are fixed because their granularity remains as the original author defined it. Such metadata can't adequately describe interactive multimedia content because they usually contain universal object description and cannot influence the content itself.

The way of tagging a learning object, according IEEE LOM scheme, assume that for the specific object many descriptors can exist. Therefore with respect to the object selection the size of the search space is dominated by the size of the content descriptors, which may be very large. When the learning object depends of input parameters, we can reuse this object in a different learning context by appropriate parameters configuration. Another important problem is how several scenarios are to be tagged to a specific object.

The number of different scenarios depends of the number of object parameters (attributes) and of the number of allowed values of each attribute. For instance, if the object contains *m* attributes and each attribute has *n* different allowed values, so at least *m.n* different scenarios exist. Following the standard tagging scheme *m.n* different object descriptions have to be produced. Bearing in mind that IEEE-LOM has approximately 60 fields, even when using templates, it is obvious that this is unacceptable solution.

Another problem is the integration of learning objects into dynamically generated instructional sequences. During the process of integration, the standard metadata schemes don't provide necessary granularity. For example, in order to integrate some visual objects into dynamically generated object, the size of the elements should be in pixels or dots, while the resource size is specified in bytes.

Further, interactive multimedia learning content may be parameterized offline as well as online. Using the static IEEE-LOM can parameterize objects only offline.

Metadata extension

The standard static metadata set has to be extended to provide online object customization. One interesting solution is the static metadata object description to be extended with dynamic metadata description. Using dynamic metadata, learners (educators) can convert an algorithm implemented by the developer to a dynamic object by mapping algorithm variables and associating the execution points of the algorithm to produce the desired behavior.

The scheme of the additional (dynamic) metadata follows the format: <Attribute, ValueType, Values>.

The metadata about the collection of objects in some application domain are stored in a specialized repository (catalog). For each object, catalog maintains standard (LOM) metadata, and additional metadata, such as the set of types exported by the object, the domains for a subset of attributes of some types.

Assume that WWW will be used as a medium for describing and publishing sources. From the other hand XML and XML-Schema will be used as a data exchange format and a model for exchanging structured data on the WWW. Consequently, The catalog has also to maintain the URL of the XML document that published the data source. This XML document describes the exported types and domain metadata according to the XML-Schema

A common schema and associated semantics for describing the various types are necessary for extending the static metadata set. The schema consists of a set of types and attributes. In order to make it available to everyone, the schema has to be published on the Web in theXML file, shown in figure 1.

```
<xs:schema xmlns:xs="http://www.w3.org/XMLSchema" version="1.0">

    <xs:element name="Type">
        <xs:complexType content="empty">
            <xs:attribute name="Attribute1" type="Type1"/>
            ...
            <xs:attribute name="Attributen" type="Typen"/>
        </xs:complexType>
    </xs:element>

    <xs:element name="AnotherType">
    ...

    </xs:element>
</xs:schema>
```

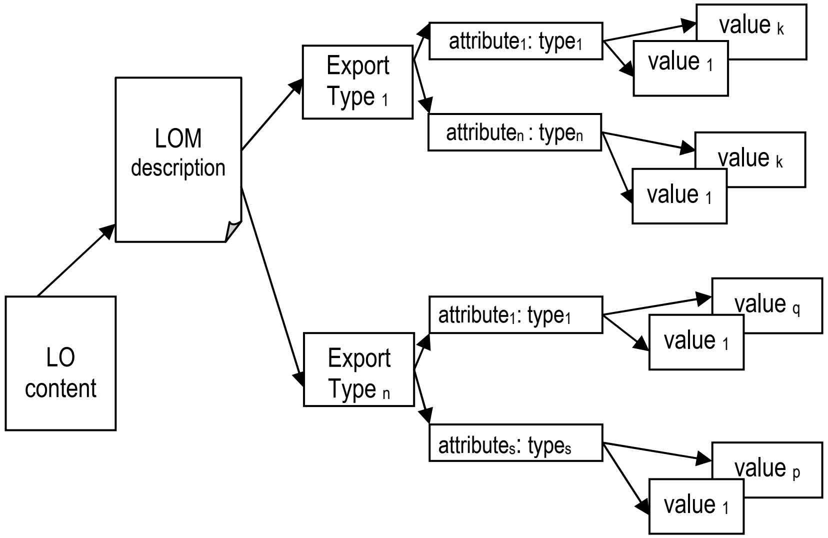
Fig. 1. XML file conforms to the XML-Schema conventions.

When a new object is registered during the repository construction, the catalog downloads the corresponding XML document and extracts types and domain metadata exported by each object. Under object selection, the set of available exported types is displayed in a listbox. The choice of a type from the listbox is actually the choice of the context the object would be used in.

Under selection of a type, the set of its attributes and their data types is displayed in a "type definition" listbox. Finally, the active domain of any attribute in the current object can be computed by selecting the attribute value. Thus, in fact a concrete scenario is tagged to a learning object. Figure 2. illustrates that tagging process.

Fig.2. A concept model of extended metadata tagging

Conclusions



An appropriate description of dynamic resources is necessary to match users' learning goals and to reuse dynamic multimedia content in different contexts.

In contrast to existing approaches and context strategies, we propose metadata set extension supporting the tagging process in respect to online object parameterization and dynamic customization. Further work is related with a query language for discovering relevant data description. This language must combine features for searching relevant XML documents that publish data sources with features for searching over the metadata describing these resources.

REFERENCES

[1] Extensible Markup Language (XML). http://www.w3.org/XML

[2] XML Schema (W3C working draft). http://www.w3.org/TR

[3] Huang, K. T., Lee, Y. W. and Wang, R. Y. (1998). *Quality Information and Knowledge.* Prentice Hall, Englewood Cliffs, NJ

[4] Mihaila, G., Rashid, L. and Vidal, M. E. (1999). "Quering 'quality of data' metadata" *Proceedings of the Third IEEE Meta-Data Conference,* Bethesda, MD. http://computer.org/proceedings/meta/

[5] Anzlic quidelines: Core metadata elements. http://anzlic.org..au/metaelem.html

[6] Longmire, W. (2000) A Primer on Learning Objects. *Learning Circuits.* http://www.learningcirquits.org/mar2000/primer.html

[7] Wiley, D. A. (2001). 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.* http://works.opencontent.org

[8] Wiley, D. A. (2002). The instructional Use of Learning Objects. http://reusability.org/read/

[9] Dublin Core Metadata Initiative (2000). http://purl.org/dc/

[10] IMS Global Learning Consortium, Inc. IMS Learning Resource Meta-data Best Practices and Implementation Guide. http://www.imsproject.org/metadata/mdbest01.html

[11] ARIADNE, Educational metadata. http://ariadne.unil.ch/Metadata/

[12] LOM, Learning Object Metadata working draft v4.1. http://ltsc.ieee.org/doc/wg12/LOMv4.1.htm

[13] Hansen, S., Schrimpsher, D., and Narayanan, N. (1998). Learning Algorithms by Visualization: A Novel Approach Using Animation-Embedded Hypermedia*, Proc. Third International Conference Learning Sciences, Association for the Advancement of Computing in Education*, pp.125-130

[14] Berners-Lee, T. (1998). Semantic Web road map. http://www.w3.0rg/DesignIssues/Semantic.html

Discussions

E-SERVICES AND E-INFORMATION DELIVERED TO BULGARIAN CITIZENS

Milena Staneva

Institute of Mathematics and Informatics, BAS
Association for the Development of the Information Society
e-mail: mstaneva@math.bas.bg

The progress in the information society, the fast Internet development and the broad popularity in Bulgaria, raise the question about possibility all Bulgarian citizens to be able to use new information and communication services.

The access to e-services and e-information will change people life at great extend. It gives new opportunity for education. The interactive electronic services make easier the interaction between citizens and state administration.

According to the goals of e-Europe initiative of EU, all citizens should be given equal possibilities to use modern effective and quality telecommunication and information services, as well as equal possibilities for training how to use them.

In Bulgaria number of people using Internet grows steadily, but as a whole the difference with high-tech countries is considerable. According to Vitosha research (2000-2002) in the end of 2002 only 9.9% of the population above 18 uses Internet.

This is due to:

- Low level of Bulgarian people income;
- Lack of enough e-information and interactive e-services for the Bulgarian citizens;
- Lack of necessary skills for using computer, finding information;
- Low security of information.

A MODEL OF AUTOMATED SCIENTIFIC LIBRARY SYSTEMS IN THE CONTEXT OF THE NEW LIBRARIES FUNCTIONS

Kristina Varbanova-Dencheva

PhD, Assist. Prof. - Central Library of BAS

Summary

The tendency of the development of science and technology library is towards establishing it as a major component of the virtual information space which will ensure the necessary basic conditions for the functioning and development of science and technology and actualizing the library’s new relation with society. These conditions determine the importance of the development of strategies for the automation of library services. The main aim will be to ensure dynamic functioning of science libraries, to meet the new requirements and to build-up its new identity.

Every new stage in the development of the technologies and the information transfer, which determine the development of the science library as a communications institution, imposes the need to redefine the aims and objectives of the automation in accordance with the changing science and society. At the present stage – the process of building up of information society and knowledge-based economy – the global aim of the automation is to be defined as the establishment of a virtual library.

The problem situations stimulate the “movement ahead” effect resulting from the automation. They are connected with the areas of interaction, respectively with the changes in the scientific, technological and social spheres having direct and strong effect on its development. The areas of interaction discussed in the dissertation are a reflection of the accumulation of the main changes in the development of science and technology and social practices, which entail changes in the functions, structure and system organization of the science library. The libraries will have to redefine their priority functions and internal structure with view to these areas so that they obtain their place in the new system environment.

The priorities areas may be defined are as follows:

1. The classification of the documents in the library stock should fully correspond to the classification of sciences in order to reflect the dynamic changes in the state of sciences and knowledge.
2. Library stock structure should be altered depending on the tendencies in the manufacture, circulation and utilization of new information carriers.
3. Manual operations in libraries as well as the corresponding equipment control should be done by robot information-searching and transportation

systems.

4. The information and structural aspect of the library control should be constantly adapted to the main tendencies in the development of the contemporary automated control systems.

The present problems situation indicated the new stage for the automation of library services. The following factors have been influenced:

- The potential for automation of science and technology libraries ensues from the contemporary technological achievements;
- The deficiencies of the existing approaches to the automation limits its application in the library services;
- Deployment of system analysis methods and modelling that will make it possible to outline new functional and structural tendencies in the automation of library services.

MOBILE INFORMATICS AND GLOBAL POSITIONING SYSTEMS

P. Barnev

People strive for making portable the information devices that they use: watches, typing machines, radio, calculators and etc. It does not pass on the computers and telephones.

Contemporary mobile devices and especially computers and telephones are connected in common systems, which include stationary devices too. Creating and using such systems is matter of mobile informatics.

The mobile informatics can be defined like science and technology for doing information activity with use of mobile information devices, connected among them and represent united system.

Among the problems of mobile informatics are: development of information and communication structure, creating variety services and specific software, helpful protection of the systems providing, social problems solving.

Global positioning systems (GPS) allow high precision specifying location in the space of static or movable objects. The quickly development is expected, because of their varied applications: navigations, watching over sick people and children, prevention and detection of thefts, disaster reactions and etc.

It is normal to expect integration between computer networks, telephone networks and global positioning systems, and their large scale of use.

Perspectives for producers and users of this systems and future problems are in interest.

Special Session
on
Multimedia Semantics

DETERMINING REGION SEMANTICS

Xin Li, W.I. Grosky**, and F.Fotouhi**

**Department of Computer Science, Wayne State University, Detroit, Michigan 48202*
***Department of Computer and Information Science, University of Michigan-Dearborn, Dearborn, Michigan 48128*
{lxiscas, fotouhi}@cs.wayne.edu, wgrosky@umich.edu

Abstract: Many methods have been proposed to extract local feature of regions for content-based image retrieval and annotation. In this paper, we introduce a Bayesian method for image segmentation and content-based annotation for satellite images based on wavelets.

1. INTRODUCTION

Content-based image annotation (CBA) refers to labeling the semantic content of images with a set of keywords. The need for image annotation has grown in parallel with that of content-based image retrieval [SWS00]. Typical methods for content-based image annotation first extract some features, such as color, texture, and shape from images, then cluster them, and finally associate keywords for each cluster.

2. RELATED WORK

Many methods for content-based image retrieval and annotation use histograms as features to compute similarity [NBE93]. However, due to its statistical nature, a histogram is only a coarse characterization of an image. Images with different appearances can have similar histograms; thus histograms can only index the images in limited way [RSZ99]. It seems that texture and spatial feature comparison of regions is a promising way to improve the accuracy of retrieval and annotation. In this project, texture is our primary feature, as we are working with grayscale satellite images.

Wavelets are widely used to analyze the image textures. In [JFS95], a method is proposed to integrate color, texture and shape into a unified framework, where users are not required to provide any parameter. A simple Haar wavelet transform is used to compute a feature vector for each image. The wavelet representation is then truncated and quantized, so that the magnitude of coefficients is discarded and only their presence or absence is recorded. This method was improved in [WWF98], using the Daubechies' wavelet transform, utilizing a better metric and a three-step process. However, their approach must decide the values of various weights, and does not support query by regions.

In the Walrus project [NRK99], features are extracted based on image regions using the Haar wavelet transform. The wavelet coefficient matrix is divided into many overlapping sliding windows, which are then clustered to generate regions. The similarity between two regions is based on the Euclidean distance between cluster centroids. Image retrieval is based on image region comparisons.

Each of these methods, however, could be further improved in three aspects. First, many of these methods extract spatial features from the global image [JFS95]. They do not consider local features of regions, which is necessary for region retrieval.

Image segmentation also presents problems. Some segmentation methods use fuzzy c-mean or k-mean clustering [NgB00], while others use neural nets [PMS95] or region growing approaches [Til98]. These methods, however, have some limitations. For example, data points of two clusters may overlap with each other, which cause difficulties in the fuzzy c-mean or k-mean algorithm. The number of clusters to use is also difficult to determine. Neural net approaches to clustering normally require tedious training, and it is difficult to set the threshold for region-growing algorithms.

Finally, a common problem is similarity computation. The Walrus project uses Euclidean distance to compare region similarity and to decide if two regions should be combined. This method, however, may cause problems in some situations, such as when the variance of region descriptors from different semantic groups are very different. Illustrating this, in Figure 1, we show sample data points from the 7 semantic groups in our project:

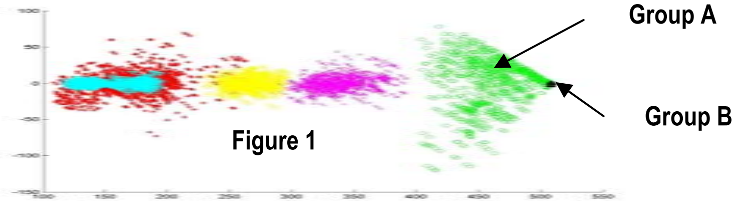


Figure 1

In this example, we can observe the two groups, A, and B, on the right side of the figure. The variance of region descriptors that belong to semantic group A is much larger than that of semantic group B. A region descriptor C_R of a region R that should belong to group A may be closer to region descriptors of group B due to the large variance of group A. If we label region R by comparing Euclidean distances of descriptors, we may assign region R to group B, which is not correct. This is exactly the error that can occur with the Walrus approach.

In this paper, we exhibit two clustering methods for image segmentation, both based on Bayesian statistics. Our second method is shown to avoid some of the above problems. We describe our method in Section 3, provide experimental results in Section 4, and conclude in Section 5.

3. IMAGE SEGMENTATION BASED ON FEATURE DISTRIBUTION

3.1. Overview

We provide a brief overview for the first version of our method. The difference between this and our final approach lies in the third step.

1. Generating features from wavelet coefficients.
2. Estimate the distribution parameter.
3. Cluster feature data from new images, using result from image segmentation and labeling.

3.2. Feature Extraction

We first perform a two-dimensional Haar wavelet transform over the image to get a new matrix of wavelet coefficients. This coefficient matrix is smaller than the original image matrix. Our next step is to extend this coefficient matrix to the original image size, which is called feature conditioning [NgB00]. Finally, we extract features from the resulting data, by using the local mean and deviation of the wavelet coefficients to reflect region properties.

For any wavelet coefficient at position of (x,y) in the matrix, we choose a block with fixed size n*n, whose center is (x,y). The local mean feature at position (x,y) is:

$$\text{Fm} = \frac{1}{n*n}\sum_{i=1}^{n*n} W(xi, yi)$$

$$\text{Fd} = \text{W(x,y)} - \text{Fm}$$

Thus we compute the local mean and deviation of wavelet coefficients in a small neighborhood area for each pixel, and take <Fm, Fd> as the resulting feature vector.

3.3. Estimate Distribution Parameter

For the second step, we manually labeled several sample regions for each semantic group, and observed the feature distribution. We observed that the features of sample regions from the same group roughly follows a normal distribution N(μ_i, Σ_i), Figure 2 shows the histogram of feature Fm from group 1 and its corresponding estimated normal distribution curve. We compute the distribution parameters μ_i and Σ_i for each sample group for clustering. We currently have 7 groups, and have used a total of 20 sample regions.

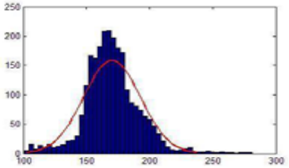


Figure 2

3.4. Image Segmentation

Finally, we use two methods for image segmentation. Our first method uses the normal distribution density to estimate the probability of each pixel in a new image belonging to each semantic group, based on Bayes rule,

$$P(W_i|x) = \frac{P(x \mid W_i)P(W_i)}{\sum_{j=1}^{n} P(x \mid W_j)P(W_j)}$$

Here, $P(W_i)$ refers to the probability of an image containing semantic group i, $P(x|W_i)$ refers to the probability of a pixel having value x given that it is in semantic group i, and $P(W_i|x)$ refers to probability of a pixel belonging to semantic group i given that it has value x. As a simplification, we assume that $p(W_1)$ = … $p(W_7)$. Thus, we need only compare the values of $P(x \mid W_i)$, for $1 \le i \le 7$. We label this pixel with the semantic group whose probability density is highest at that point. Below is an example of image segmented by our method.

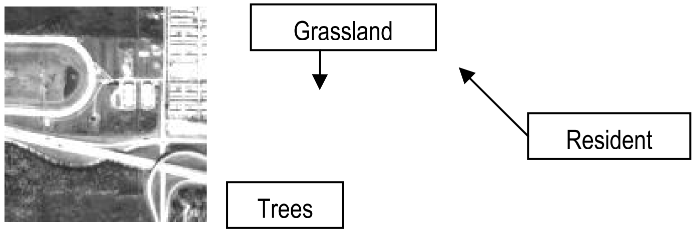


Figure 3

Due to the intersection of the various normal distributions, this technique results in type I and type II statistical errors. In order to improve the accuracy of this approach, we define the following feature of each block.

If each block has m pixels, we use $\sum_{k=1}^{m}(x_k-\mu_i)'*\Sigma_i^{-1}*(x_k-\mu_i)$ as the elements of feature vector for each block. Here, x_k is the two-dimensional feature vector of pixel k in the given block, and μ_i , Σ_i are the two-dimensional mean vector and 2 × 2 covariance matrix of the ith semantic group, respectively.

Assuming that each pixel's feature value is independent and that they all follow the same normal distribution of $N(\mu_i, \Sigma_i)$, $\sum_{k=1}^{m}(x_k-\mu_i)'*\Sigma_i^{-1}*(x_k-\mu_i)$

should follow the centered chi-square distribution with free degree of m*d, where d = 2 is the dimensionality of the pixel feature vector. In Figure 4 we show a histogram from our actual data, and notice that it follows the chi-square curve.

Finally, we find the group i that minimizes $\sum_{k=1}^{m}(x_k-\mu_i)'*\Sigma_i^{-1}*(x_k-\mu_i)$, and assign this group as the label for the given block.

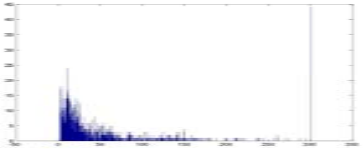


Figure 4

4. EXPERIMENTS

Below are some experimental results. After finding the appropriate statistics using 20 sample regions, Table 1 lists the accuracy of our initial approach for data points from 16 different regions. Each row is labeled with the given block size used.

	R1	R2	R3	R4	R5	R6	R7	Overall
2	0.469	0.754	0.828	0.832	0.994	0.701	0.999	0.797
3	0.495	0.758	0.816	0.872	0.988	0.748	0.999	0.811
4	0.573	0.781	0.821	0.916	0.991	0.713	0.999	0.828
5	0.601	0.783	0.810	0.931	0.982	0.743	0.999	0.836
6	0.639	0.796	0.816	0.964	0.985	0.701	0.998	0.843

We notice that the overall accuracy grows when we increase the block size. This is reasonable, since when the block size grows, the feature stability also increases, which improves the overall accuracy. However, the effect is reduced when we continue to increase the block size, due to problems in recognizing region borders.

The table below lists the accuracy of our improved version for data points from sample regions that we have labeled.

	R1	R2	R3	R4	R5	R6	R7	Overall
2	0.975	0.681	0.834	0.974	0.929	0.620	0.985	0.857
3	0.999	0.702	0.821	0.982	0.909	0.699	0.990	0.871
4	0.999	0.775	0.835	0.995	0.841	0.623	0.996	0.866
5	0.995	0.762	0.819	1.000	0.829	0.708	0.994	0.872
6	0.997	0.811	0.835	1.000	0.637	0.630	0.994	0.844

Table 2 illustrates that our improved approach is better in two aspects. First, our average accuracy is improved. Secondly, and more importantly, the accuracy variance for each semantic group is much reduced. This result shows that utilizing blocks results in more stability than would result from single pixels. Below, we compare the result of image segmentation with our initial version as well as with our improved version. This shows that the improved version reduces a lot of scattered error blocks in the trees region.

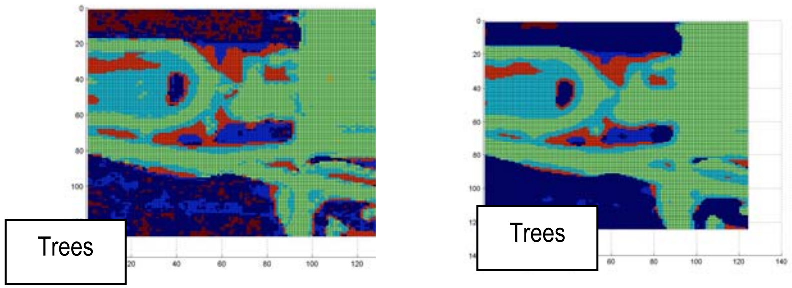


Figure 5

5. CONCLUSION AND FUTURE WORK

Our method employs a distribution parameter for supervised clustering and image segmentation, thus supporting query-by-regions. Our results are promising, and lend themselves to the development of more refined image segmentation approaches. We avoid several common difficulties of other non-parameter clustering methods, such as setting a unified threshold or radius in our application. We use the chi-square distribution to detect block patterns, which shows more accurate for content-based image retrieval and annotation than just using the normal distribution density.

Our future work will include further comparison of block patterns and developing a Bayesian approach for extracting region borders. The integration of multilevel wavelet coefficients is also a possible direction.

REFERENCES

[JFS95] C. Jacobs, A. Finkelstein, and D. Salesin, 'Fast Multiresolution Image Querying,' *Proceedings of the ACM SIGGRAPH Conference,* Los Angeles, California, 1995, pp. 277-286.

[NBE93] W. Niblack, R. Barber, W. Equitz, et. al., 'The QBIC Project: Querying Images by Content, Using Color, Texture and Shape,' *Proceedings of the Conference on Storage and Retrieval for Image and Video Databases (SPIE),* San Jose, California, 1993, pp. 173-187.

[NgB00] *B. Ng and A. Bouzerdoum, '*Supervised Texture Segmentation using DWT and a Modified K-NN Classifier,' *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR'00),* Barcelona, Spain, 2000, pp. 2545-2548.

[NRK99] A. Natsev, R. Rastogi, and K. Shim, 'WALRUS: A Similarity Retrieval Algorithm for Image Databases,' *Proceedings of the ACM SIGMOD Conference*, Philadelphia, Pennsylvania, 1999, pp. 395-406.

[PMS95] *S. Pemmaraju, S. Mitra, Y.-Y. Shieh,* and *G.H. Roberson, '*Multi-resolution Wavelet Decomposition and Neuro-fuzzy Clustering for Segmentation of Radiographic Images,' *Proceedings of the Eighth Annual IEEE Symposium on Computer-Based Medical Systems (CBMS'95),* Lubbock, Texas, 1995, pp. 142-149.

[RSZ99] A. Rao, R. Srihari and Z. Zhang, **'**Spatial Color Histograms for Content-Based Image Retrieval,' Proc. of the 11th IEEE International Conference on Tools with Artificial Intelligence, Chicago, Illinois, 1999, pp. 183-186.

[SWS00] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, 'Content-Based Image Retrieval at the End of the Early Years,' *IEEE Trans. Pattern Analysis and Machine Intelligence,* Volume 22, Number 12 (2000), pp. 1349-1380.

[Til98] J. Tilton, 'Image Segmentation by Region Growing and Spectral Clustering with a Natural Convergence Criterion,' *Proceedings of the 1998 International Geoscience and Remote Sensing Symposium,* Seattle, Washington, 1998.

[WWF98] J. Wang, G. Wiederhold, O. Firschein, and S. Wei, 'Content-Based Image Indexing and Searching Using Daubechies' Wavelets,' *Intl. Journal of Digital Libraries (IJODL)*, Volume 1, Number 4 (1998), pp. 311-328.

CONTENT-BASED NAVIGATIONAL QUERYING

D.V. Sreenath, W.I. Grosky**, and F. Fotouhi**

**Department of Computer Science,*
Wayne State University, Detroit, Michigan 48202
***Department of Computer and Information Science,*
University of Michigan-Dearborn, Dearborn, Michigan 48128
{sdv, fotouhi}@cs.wayne.edu, wgrosky@umich.edu

Abstract: Image semantics are context sensitive. User browsing paths over an image collection provide, we believe, the necessary context to derive image semantics. Such semantics emerge over a period of time, through user interaction. We have attempted to use this emergent semantics for content-based image retrieval. Based on experiments which retrieve images from a diverse image collection using a sequence of images as a query against a database of semantically coherent browsing paths through the image collection, our results are encouraging.

1. INTRODUCTION

In traditional web mining efforts, the emphasis has been on keywords or links. In either case, it is predominantly text-based. With respect to navigation on the web, when requesting pages similar to a current page, the emphasis is again on text. We have used a variation of latent semantic analysis to derive the emergent semantics of web pages using the users' browsing paths [1]. In recent years there has been a lot of work in the areas of content-based image retrieval, where low level features of the images are used to derive high-level concepts or semantics.

Latent semantic indexing [2] and principal component analysis are two popular techniques used for large text collections to derive latent features. There are numerous publications in the areas of content-based image retrieval and techniques for retrieval of images from large collections [3-7]. In [5], the singular value decomposition is used as a dimension reduction technique for a text-image matrix. The terms closer to an image are weighted more than the terms farther away. This is used for image retrieval and not to derive the image features that co-occur with the same set of textual keywords. The experiments in [6] combine textual and image features for image retrieval using latent semantic indexing.

In [8], it is argued that that the images by themselves do not have any semantics, but they do when placed in the context of other images and through user interaction. Thus, image semantics are context sensitive. From the image database perspective, the semantics of an image can be extracted by interpreting the sequence of queries posed by the user. In their image database system, the semantics is not an intrinsic property captured during the image filtering process, but an emergent property of the interaction of the user and the database.

In our research, images have multiple semantics that vary over time. Each element of an image's multiple semantics corresponds to a group of users who visit this particular image through semantically similar browsing paths[24][9]. Simply put, the semantics of any image on a semantically coherent browsing path is defined in terms of the extracted features from each image on this path. As different users visit a given image, its semantics can change, since the same image can be part of many different semantically coherent browsing paths.

2. OUR APPROACH

This research effort is an application of our earlier research in deriving emergent semantics as detailed in [1]. We use the technique of latent semantic analysis [2], to derive the multimedia semantics. We represent user browsing paths (the sequence of images) in a reduced dimensional space, where each dimension corresponds to a concept, each concept representing a set of visual-keywords.

In our approach, we extract features from the JPEG images in the compressed domain. The existing de-facto JPEG file format standards specify YCbCr, since this permits greater compression. The luminance component (Y) contains the grayscale and the chrominance components (C_{red} and C_{blue}) contain the color information. The DC (0,0)th coefficient of the cosine transforms of the 8x8 block represents the average intensity value. Depending on whether we have an 8-bit or 12-bit implementation, there could be 255 or 4095 DC coefficient values. We use the DC coefficients of the all the three: Y, Cb and Cr components as the basis for our features. Each image is represented by 860 components composed of three histograms of the absolute DC coefficient values, three histograms of the deviations of these values from their neighbors, the maximum and minimum absolute values, and the maximum deviation.

Now, deviation = $\left(\sqrt{\sum_{i=1}^{N} (x - x_i)^2} \right) / N$, where N represents the neighbors whose values range from 3 to 8 depending on the location of the block. This approach takes care of the blocks around the edges where there may not be 8 neighbors. The number of neighbors considered can be expanded to simulate the variable-size sliding window as in the WALRUS system[10]. The benefit of our approach is that we are able to extract the features directly by reading the compressed domain image file without any additional wavelet computational overhead. Also, the use of the deviation helps in identifying a uniform region or an edge.

[24] We assume that there is some external mechanism, such as the approach taken in content-based hypermedia [9] allowing the user to navigate from one image to another image of his interest. In this paper, we ignore this external mechanism.

There are relevance feedback-based systems aimed to use the user's input to aid in image retrieval. In our research, we try to derive the semantics using the users' browsing paths without any direct feedback from the user. The image-browsing path is defined as a sequence (n_1, n_2…, n_q). We extract features from each image in the browsing path to build a term-path feature matrix. The actual value used in the matrix represents the number of times that a particular feature appears in that path. An element $x_{i,j}$, the (i,j)th element of the term-path matrix, M, is determined by the strength of the presence of the ith keyword, t_i, along the jth browsing path, ($n_{1,j}$, $n_{2,j}$, …, $n_{q,j}$). By the singular value decomposition, any rectangular matrix M can be decomposed into a product of three other matrices U, Σ and V. $M=U\Sigma V^T$. A rank k approximation is performed choosing an appropriate value of k < n to reduce the dimensionality. Defining $M_k = U_k \Sigma_k V^T_k$, M_k provides the best rank k approximation of M. M_k reveals the latent structure of M, using linear combinations of keywords as new concepts that can be used to derive semantics.

A single image is a special case of a browsing path whose length is 0. Each image or path represented as a vector in the term-path matrix can be visualized as a point in the reduced dimensional space. The semantics of a query image, w, can then be defined as the subset of the points in the reduced dimensional space that are within the threshold distance from w. The semantics of a user's browsing path is the collection of concepts represented by the semantics of the images traversed by him. We accomplish this by placing the query (also represented as a point) in the vector space and computing the distance between the query point and the set of all the points in the vector space.

We use the cosine distance measure in the reduced-rank vector space model as in [11]. We compare a query vector q to the columns of the approximation M_k to the term-by-path matrix M. If we define e_j to be the jth canonical vector of dimension d (the jth column of the d x d identity matrix), the jth column of M_k is given by $M_k e_j$. The cosines of the angles between the query vector q and the approximate path vectors can be computed by $\cos \theta_j = (s^T_j (U^T_k q)) / (\| s_j \|_2 \| q \|_2)$ for j = 1,2,…d, where $s_j = \Sigma_k V^T_k e_j$ for j = 1,2, …d.

3. PRELIMINARY RESULTS

We experimented with 720 images with 860 dimensions each. We reduced the dimensions to 200 using rank-K approximation. We tested our prototype with simple queries with single images to successfully retrieve images of snow, sunsets, flowers, buildings, sceneries, people, homes, etc. It is difficult to derive the semantics of just one image. Images have more than one meaning and the emotions evoked are very subjective. Instead of using just an image as a query, we have used multiple images as a query to simulate a user browsing a sequence of images. Similarly, the database was also populated with sets of images, each set exhibiting one predominant concept.

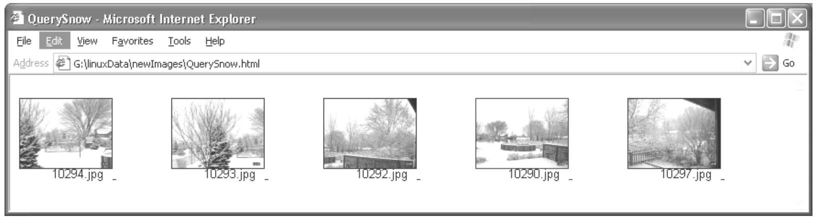

QuerySnow - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address G:\linuxData\newImages\QuerySnow.html
Go
10294.jpg
10293.jpg
10292.jpg
10290.jpg
10297.jpg

Figure 1 – Composite image query composed of several images snow

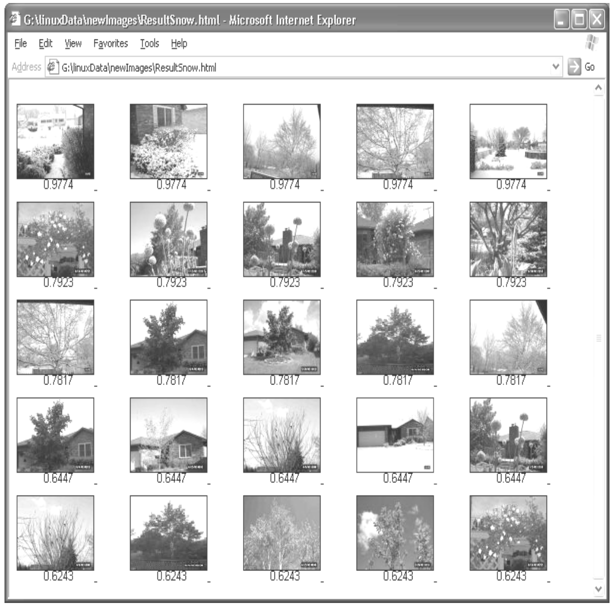


Figure 2 – The result set for the query in Figure 1 retrieving images of snow.
The cosine distance function values are displayed below the icons

We conducted 4 classes of experiments to study the behavior of stand-alone images and image sets. When a query composed of images of snow-covered trees in Figure 1 was used to simulate the viewer's desire, the results in Figure 2 extracted the path that contained similar images. It is to be noted that our approach retrieves images that are similar in concept and work beyond the color, shape and texture limitations. In the second class of experiments, we used the same image path to query the database of individual images. This approach did not make any significant improvement to the rank of the retrieved images. In the third class of experiments, we used individual images to query the database of individual images. The results were impressive from the content-based image retrieval perspective. But the results did not include additional images (semantically similar) as retrieved in the first set of experiments.

For example, when the context of the query image was *snow*, the retrieval puts similar weight on matching every object, like trees or buildings, in the query image and not just on the high-level concept. It is interesting to compare the results in Figure 2 with that obtained when each image in Figure 1 is used as a query. Each image in each of the result sets is ranked based on the distance measure. The ranks for each of the retrieved images are tabulated in Table 1.

For example, the column labeled *10290 rank* represents the rank of each of the images in Figure 2 based on the result set obtained using 10290.jpg as a query for a database of individual images.

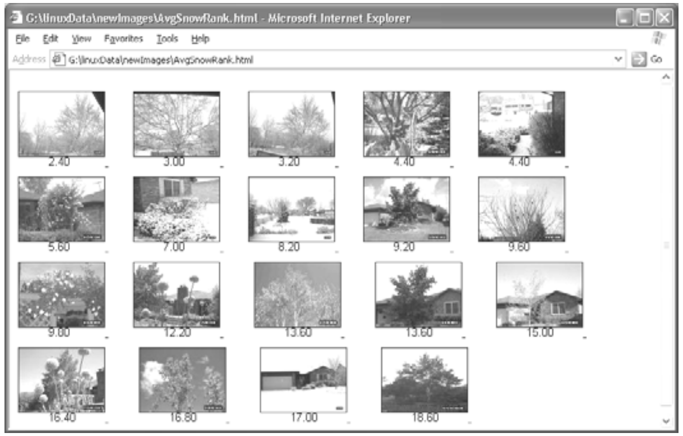

G:\linuxData\newImages\AvgSnowRank.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address G:\linuxData\newImages\AvgSnowRank.html
Go
2.40
3.00
3.20
4.40
4.40
5.60
7.00
8.20
9.20
9.60
9.80
12.20
13.60
13.60
15.00
16.40
16.80
17.00
18.60

Figure 3 – The computed result set using the average rank of images.
The average rank is computed as in Table 1.

Comparing Figure 2 and Figure 3 it can be inferred that using a single query comprised of a path of images, retrieves semantically better results than when the average rank is computed after performing several queries one at a time. It is also computationally efficient. The final class of experiments used each one of the images in the query-set as a query against the database of image-sets. The actual images and additional queries used for these experiments can be viewed at

http://www.cs.wayne.edu/~sdv/images/imageSemantics/results.html.

Table 1 – Rank comparison: using a path versus individual images in the path. For each of the images in Figure 2 the rank of the images in the result set when queried using each one of the images in Figure 1

Image in Figure 2	10290 rank	10292 rank	10293 rank	10294 rank	10297 rank	Average rank
10296.jpg	4	1	2	3	2	2.40
10298.jpg	3	4	1	2	5	3.00
10291.jpg	2	2	3	5	4	3.20
10163.jpg	7	5	5	4	1	4.40
10288.jpg	8	3	4	1	6	4.40
10651.jpg	5	6	7	7	3	5.60
10302.jpg	9	7	6	6	7	7.00
10289.jpg	1	12	8	8	12	8.20
10477.jpg	11	8	10	9	8	9.20
10378.jpg	6	11	9	11	11	9.60
10602.jpg	10	9	11	10	9	9.80
10223.jpg	17	10	12	12	10	12.20
10280.jpg	12	14	14	14	14	13.60
10532.jpg	16	13	13	13	13	13.60
10271.jpg	15	15	15	15	15	15.00
10246.jpg	14	16	18	18	16	16.40
10470.jpg	18	17	16	16	17	16.80
10304.jpg	13	19	17	17	19	17.00
10548.jpg	19	18	19	19	18	18.60

REFERENCES

[1] W. I. Grosky, D. V. Sreenath, and F. Fotouhi, 'Emergent Semantics and the Multimedia Semantic Web,' *SIGMOD Record*, vol. 31, pp. 54-58, 2002.

[2] S. Deerwester, S. T. Dumais, and G. W. Furnas, 'Indexing by Latent Semantic Analysis,' *Journal of the American Society for Information Science*, vol. 41, pp. 391-407, 1990.

[3] J. Bigun, "Unsupervised Feature Reduction in Image Segmentation by Local Transforms," *Pattern Recognition Letters*, vol. 14, pp. 573-583, 1993.

[4] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih, "Spatial Color Indexing and Applications," *International Journal of Computer Vision*, vol. 35, pp. 245-268, 1999.

[5] S. Sclaroff, M. La Cascia, S. Sethi, and L. Taycher, "Unifying Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web," *Computer Vision and Image Understanding*, vol. 75, pp. 86-98, 1999.

[6] Z. Pecenovic, M. N. Do, and M. Vetterli, "Integrated Browsing and Searching of Large Image Collections," *Advances in Visual Information Systems, Proceedings*, vol. 1929, pp. 279-289, 2000.

[7] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, pp. 233-254, 1996.

[8] S. Santini, A. Gupta, and R. Jain, "Emergent Semantics Through Interaction in Image Databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, pp. 337-351, 2001.

[9] W. I. Grosky and D. V. Sreenath, "Metadata Mediated Browsing and Retrieval in Semantically-Rich Cultural Image Collections," presented at Proceedings of the 2001 Tokyo Symposium for Digital Silk Roads, Tokyo, Japan,, 2001.

[10] A. Natsev, R. Rastogi, and K. Shim, "WALRUS: A Similarity Retrieval Algorithm for Image Databases," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 395-406, 1999.

[11] M. W. Berry, Z. Drmac, and E. R. Jessup, "Matrices, Vector Spaces, and Information Retrieval," *Siam Review*, vol. 41, pp. 335-362, 1999.

UTILIZING RELATIONS IN MULTIMEDIA DOCUMENT MODELS FOR MULTIMEDIA INFORMATION RETRIEVAL

Temenushka Ignatova & Ilvio Bruder

Computer Science Department, University of Rostock
E-Mail: {ti005, ilr}@informatik.uni-rostock.de

Abstract: This article describes the first considerations during a work in a postgraduate program at the University of Rostock, Germany [3]. The insufficient support in multimedia information systems for semantic relations between multimedia document components is brought out. Besides encoded spatial, temporal and structural relations, interpreted semantic relations can also be used for multimedia document retrieval. A classification and approaches to integrate these relations in multimedia document models are proposed.

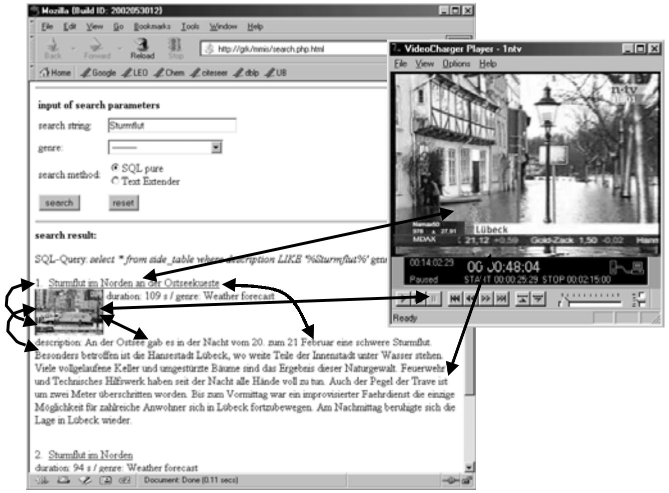
Keywords: multimedia document modeling, multimedia content-based relations, multimedia retrieval, multimedia semantics

1. Introduction

Multimedia (MM) presentation scenarios implement the semantics for the interaction and representation of the components of MM documents. These semantics can and should be incorporated in the retrieval of MM documents. In this article the representation semantics are defined as relations between the single components of the documents. An example of a MM document is given in figure 1. In the upper frame, part of a query form can be seen, with which the results displayed in the lower frame have been extracted. The first document from the result list is the video document, taken as an example. The whole view consists of MM components such as text for the title, description, genre, an image, representing a frame from the video and giving the user the opportunity to play back the video by clicking on the image. Since the video is a news report it consists also of an audio track. The scenario is encoded in the HTML file. The arrows in the image illustrate the semantic relations between the MM components. For example the thumbnail image generalizes the information represented in the video.

There are various models used to specify the structure and layout of MM documents (e.g., HTML, SMIL, MPEG7).

Basic elements of the models are the MM components (text, image, audio, video), their attributes and the relations between the components. The models pointed above provide the possibility to define presentation scenarios with temporal, spatial, structural and integrating relations. However, to enhance the possibilities to perform content-based retrieval on MM documents it is important to be able to identify also semantic relations between the MM document components. The presentation and logical models realize two of the basic system levels of MM information systems (MMIS) for storage, management and retrieval of MM data. A simple abstraction of an MMIS can be represented by a three-level model consisting of an external, a logical and a physical level.


Mozilla (Build ID: 2002053012)
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://gk/mmis/search.php.html
Home Google LEO Chem citeseer dblp UB
VideoCharger Player - 1ntv
File View Options Help
n-tv
Lübeck
MDAX 21,12 +0,59 Gold-Zack 1,50 -0,02
00:14:02:29
00:00:48:04
Paused
START 00:00:25:29 STOP 00:02:15:00
Ready
Document: Done (0.11 secs)

input of search parameters

search string: Sturmflut

genre: ———

search method: ● SQL pure ○ Text Extender

search reset

search result:

SQL-Query: *select * from side_table where description LIKE '%Sturmflut%' gen*

1. Sturmflut im Norden an der Ostseekueste

duration: 109 s / genre: Weather forecast

description: An der Ostsee gab es in der Nacht vom 20. zum 21 Februar eine schwere Sturmflut. Besonders betroffen ist die Hansestadt Lübeck, wo weite Teile der Innenstadt unter Wasser stehen. Viele vollgelaufene Keller und umgestürzte Bäume sind das Ergebnis dieser Naturgewalt. Feuerwehr und Technisches Hilfswerk haben seit der Nacht alle Hände voll zu tun. Auch der Pegel der Trave ist um zwei Meter überschritten worden. Bis zum Vormittag war ein improvisierter Faehrdienst die einzige Möglichkeit für zahlreiche Anwohner sich in Lübeck fortzubewegen. Am Nachmittag beruhigte sich die Lage in Lübeck wieder.

2. Sturmflut im Norden

duration: 94 s / genre: Weather forecast

Figure 1: MM document components and their semantic relations

2. Related Work

Document models implement different aspects, such as logical or presentation aspects and different research field points of view, for example information retrieval, databases or computer graphics.

In chapter 2 of [7] Wu, Kankanhalli, Lim and Hong have represented content-based definitions for MM documents and systems from the computer vision point of view. A MM object (MOB) is defined as a 6-tuple with a type, a set of features, a set of interpretations, attributes, structural relations and a status. Another logical model for MM documents is proposed by Chiaramella, Mulhem and Fourel in [4], which implements the information retrieval aspect. The authors describe a MM document with a logical structure, attributes and a navigation structure. The logical structure consists of objects which are arranged hierarchically or sequentially using relations. A logical model for MM documents based on the description logic ACL has been enhanced with fuzzy logic based functionality by Meghini, Sebastiani and Straccia [6]. Simple documents, for example components of MM documents such as text or images are defined in this model as a set of fuzzy assertions. A MM document is a complex object with a defined structure. The structure of the document is specified as a tree, the elements of which are arranged based on the specified intervals.

Boll, Klas and Westermann [2] have studied SMIL, MHEG, HyTime and HTML as external MM documents models. The authors have analyzed the proposed models for problems with reuse and user-specific adaptation of MM documents. The ZyX model, developed by the research group has been proposed to solve these problems. MPEG7 (an overview is given by Martinez in [5]) is developed by the MPEG-Group to describe different MM data, whereby mostly content describing data, represented as features are considered. The structure and semantics of documents is defined using a so called description scheme. It is possible to define user specific description schemes to describe complex relations between MM components for example.

To answer the question about how far the above models are suitable for the description of complex relations the comparison has to be made against the different levels of the MMIS. On the logical level the existing approaches offer a relatively good support. Partial changes by extending or specifying these logical models are also relatively easy to perform. The document models which are primarily used for the representation are less flexible and cannot be easily adapted because a change in the structure or semantics will lead directly to need of changes in the interpretation applications.

3. Relations in Multimedia Documents

Relations in MM documents are responsible for the description and control of the structure, layout, state and the behavior of MM documents. We distinguish two basic types of relations. At first "encoded" relations which are directly represented in the scenario, for example an image has to be displayed in a region at coordinates x,y. Then not encoded, respectively "interpreted" relations from the data of the MM documents, eventually using additional knowledge can be identified. For example one can interpret a spoken message (audio) as a describing element for a news program video. To the encoded relations belong the following subtypes as shown in figure 2: spatial, temporal, structural and interactional. These relations are supported by most MM document models. Describing relations reflect the fact that there exist semantic relations between two MM components. These relations can be represented either as additional information, the same information but in another form, or as a description referring to a single part of the information. To represent the same information can be also referred to as transformations, where the same content is encoded in another format, for example an image of a manuscript can be transformed into text using OCR functions.

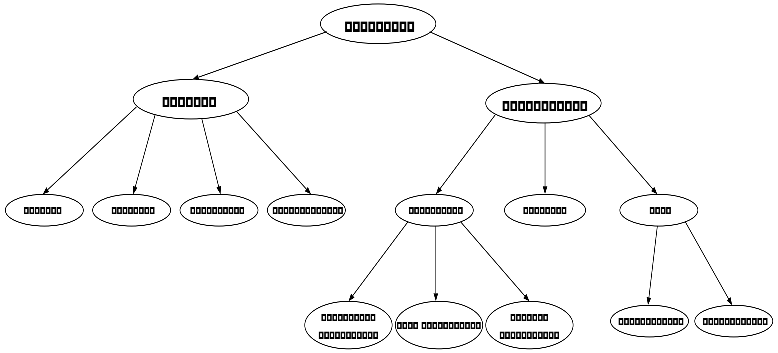


Figure 2: Classification of possible relations in MM documents

An objective overview of relations does not exist because exactly during the interpretation the ambiguity problems arise, which can be settled only by determining a concrete application and respectively a concrete perception. An important question for determining the perception is whether the decision basis is the functionality, respectively purpose of the relation or the modeling, encoding of a relation. For example a rendered headline of a web site has the function of an element to be read, but it can be also interpreted as an image with an interesting texture.

Semantic relations can help users of MMIS perform their search through introducing new query and presentation options of MM documents. An example of such a query could be the search for text and audio components, representing the same information for the example in figure 1.

The ranking of the results from the query can be also modified to be adapted to the enhanced query. For example a general information to a theme, related by a generalization relation will have a lower ranking value than information from a document with a specialized relation.

The relations may also be represented as a result of the document query. The user may use the described relations to draw additional information for his search and respectively adjust his information perspective. In Table 1 possible roles of relations are defined in the processes mentioned above. It is important to take account of the fact that these roles are in general application dependent.

Relation	Roles		
	Query	Presentation	Ranking
encoded relations			

spatial	searching on relative position to other components	components relative to other components	corresponding to relative position
temporal	searching for components with a certain time scenario	representing components in a certain time scenario	ranking depending on the relative time attributes
structural	searching for specific document structures	structure, or hierarchical presentation of content	local hits may have higher ranking weights
interactional	responding to a certain user behavior or events	corresponding to user behavior or other events	e.g. link-based ranking
interpreted relations			
describing additional information	searching for documents or components that explain or describe others	optional representation of additional information	different weighting for hits in abstract and special information
desc. the same information in another form	searching for exactly the same information, but in another media form	alternative presentation	finding relevant components following such relations
describing partial information	searching for components that give additional information	optional representation of additional information about document elements	corresponding ranking of documents with additional information
information encoded in another way	searching for information in components in different media formats	alternative representation	relevance by following relations; corresponding to the number of relations
is_a – specializing	searching for a more specific specialized content	optional representation	relevance by following relations; corresponding to the number of relations
is_a – generalizing	searching for related documents with more abstract content	optional or more abstract representation	ranking corresponding to the number of relations

Table 1: Possible roles of relations

4. Multimedia Document Models in the 3-Level Model

Each level of the abstract presentation of a MMIS can be represented by a model. For the external and logical level a detailed description and analysis of the corresponding models is given below. The internal model is not discussed in this paper.

External Model The external MM document model has the primary function to describe the attributes and relations in respect of content for the representation to the user and support the interaction with the user. The realization of such models determine the structure and type of the documents. Examples of such models are

SMIL, HyTime, MHEG, HTML or ZyX. The relations between the MM components which can be defined by these models are not sufficient for the growing needs of a MM retrieval system. The reuse of content and user-specific adaptation possibilities need to be extended. New requirements arise for a model incorporating the semantic relations between the MM components. Since interpreted relations are application dependent and it is likely that they often change, they cannot be implemented as static elements in the models. MHEG, HTML, HyTime and SMIL are models used by corresponding applications which makes changing these models difficult, since the changes will lead to format incompatibility. However, a model based on a conceptual extension of one of these models can be defined and implemented using XML with all needed components, representation rules and relevant relations from application and author point of view. ZyX is one of the most highly-developed presentation models with formal internal definitions. Boll and Klas [1] define in ZyX a couple of possibilities for representing temporal and spatial relations and also elements for referencing other objects. Using the concepts switch, decide, or query, presentation parameters could be switched or selected. Most representation aspects of our application examples can be implemented with ZyX. A further aim of the research is defining an XML extendable schema as basis for implementing such models.

Logical Model The logical level implements an abstract model for the management of the MM components and meta data which is responsible for providing the schematic and machine understandable presentation of the MM documents.

For representing all components of a MM document basic elements such as objects and their attributes have to be defined. Relations to represent the links between the basic elements have to be also abstracted.

From the analysis of the models common basic elements can be identified: Objects, representing simple data types and collections reflecting complex data types, such as sets, lists, multisets. Hence a hierarchical composition of simple objects can be defined to structure MM documents. For these simple objects (MM components) type specific characteristics (features) could be defined, for example the color distribution and color resolution of an image object. Besides the structural relations, references to other documents can also be supported. It should be possible to define elements for relation interpretation or generally specify complex relations. The MM data model in [7] represented from the computer graphics point of view can be extended using the specified relations, so that besides the structural relations, interpreted relations can be represented. Therefore a set of additional functions are needed to map new relations. In the model, described in [4] the structural relations come also on first place. Furthermore references for navigation between the documents are defined in a so called Hyperbase. The best approach here would be to define an additional construct explicitly for relations. This construct should define for each relation a name and a function which represents a relation between two data objects. It makes sense to define attributes for the relations so that descriptions respectively connections to other relations can be represented. In the description logic model from [6] generally relations are described using assertions. The structure of the model is built based on intervals. In this way it is not possible to represent all possible kinds of relations. A workaround would be to define a set of assertions for a MM document, which specify different kinds of object relations (referred to as layouts by the authors).

Each of these logical models can be extended to support additional relation types. The suitability of each model can be determined based on the particular application.

5. Conclusions and Future Work

In this article MM document models are analyzed with respect to their semantic and structural relations. The content or as referred above interpreted relations are exactly the ones, which are hard to identify and have often a subjective character. However they represent data, which can be used to define complex and combined queries. An evaluation of the concept after its implementation is planned to be carried out on a database of music manuscripts, which represent MM documents with diverse structure and MM components.

REFERENCES

[1] S. Boll and W. Klas. ZyX – A Multimedia Document Model for Reuse and Adaptation of Multimedia Content. *IEEE Transactions on Knowledge and Data Engineering*, 13(3): 361–382, 2001.

[2] S. Boll, W. Klas, and U. Westermann. Multimedia Document Models – Sealed Fate or Setting Out for New Shores? Kluwer Academic Publishers, 2000.

[3] I. Bruder and T. Ignatova. Ausnutzung von Beziehungen in Multimedia Dokumenten fuer Speicherung und Retrieval in multimedialen Informa-tionssystemen. In H. Höpfner and E. Schallehn, editors, *Tagungsband zum 15. GI-Workshop "Grundlagen von Datenbanken", Tangermünde*, 2003. (in german).

[4] Y. Chiaramella, P. Mulhem, and F. Fourel. A Model for Multimedia Information Retrieval. Technical report, CLIPS, IMAG-Campus - BP 53, Cedex, France, 1996.

[5] J. M. Martínez. Overview of the MPEG-7 Standard. Technical report, ISO/IEC, March 2001.

[6] C. Meghini, F. Sebastiani, and U. Straccia. A Model of Multimedia Information Retrieval. Technical report, B4-21-09-98 CNR-IEI, Pisa, Italy, 1998.

[7] J. Wu, M. Kankanhalli, J.-H. Lim, and D. Hong. *Perspectives on Content-based Multimedia Systems*. Kluwer Academic Publishers, 2000.

HIGH LEVEL COLOR SIMILARITY RETRIEVAL

Peter L. Stanchev, David Green Jr., Boyan Dimitrov

Kettering University, Flint, MI 48504
{pstanche, dgreen, bdimitro}@kettering.edu

Abstract: In this paper a new method for image retrieval using high level color semantic features is proposed. It is based on extraction of low level color characteristics and their conversion into high level semantic features using Johannes Itten theory of color, Dempster-Shafer theory of evidence and fuzzy production rules.

Keywords: Image Databases, Contents Retrieval, Color

1. INTRODUCTION

More and more audio-visual information is available in digital form in various places around the world. MPEG-7, formally called "Multimedia Content Description Interface", was created to describe multimedia documents. The most used features for image description is color. Many of the existing image databases allow users to formulate queries by submitting an example image. The system then identifies those stored images whose feature values match those of the query most closely, and displays them. Color features are usually represented as a histogram of intensity of the pixel colors. Some systems, such as Color-WISE [10], partition the image into blocks and each block is indexed by its dominant hue and saturation values. Color and spatial distribution can be also captured by an anglogram data structure [6].

High level image semantic representation techniques are based on the idea of developing a model of each object to be recognized and identifying image regions which might contain examples of the image objects. One early system aimed at tackling this problem is GRIM_DBMS [9]. The system analyzed object drawings, and use grammar structures to derive likely interpretations of the scene. The concept of the semantic visual template is introduced by Chang et al [3]. The user is asked to identify a possible range of color, texture, shape or motion parameters to express his or her query, which is then refined using relevant feedback techniques. When the user is satisfied, the query is given a semantic label (such as "sunset") and stored in a query database for later use. The use of the subjective characteristics of color (such as warm or cold) to allow retrieval of images is described in [4].

There is a ‘‘semantic gap’’ between information that can be derived automatically, at archiving time, and what is convenient for usability at querying time. How we can search by painting stiles? Some stiles descriptions follow. The work that distinguishes the Baroque period is stylistically complex, even contradictory. In general, the desire to evoke emotional states by appealing to the senses, often in dramatic ways, underlies its manifestations. Some of the qualities most frequently associated with the Baroque are grandeur, sensuous richness, drama, vitality, movement, tension, emotional exuberance, and a tendency to blur distinctions between the various arts. The Cubist style emphasized the flat, two-dimensional surface of the picture plane, rejecting the traditional techniques of perspective, foreshortening, modeling, and refuting time-honored theories of art as the imitation of nature. Cubist painters were not bound to copying form, texture, color, and space; instead, they presented a new reality in paintings that depicted radically fragmented objects, whose several sides were seen simultaneously.

In this paper a method for image retrieval, based on high level color image semantic features is discuss. It is a generalization of the method described in [11] and gives possibilities for retrieval painting stiles by color contrast. The layout of the paper is as follows. In section 2 we explain the image feature extraction mechanism. In section 3 we describe image retrieval based on high level color semantic features. In section 4 we detail our experiments, and finally in section 5 the conclusions of this paper are formulated.

2. COLOR FEATURE EXTRACTION MECHANISM

In this section the color characteristics extraction technique and transformation of low level color characteristics into high level color features are presented.

2.1. Color characteristics extraction mechanism

Color characteristics are usually represented as a histogram of intensity of the pixel colors. Based on a fixed partition of the image, an image could be indexed by the color of the whole image and a set of image sub regions. In our method the color feature extraction procedure includes color image segmentation. For this purpose ideas from the procedure described in [4] are adopted. Fist the standard RGB image is converted as L*u*v* (extended chromaticity) image, where L* is luminance, u* is redness–greenness, and v* is approximately blueness–yellowness [2]. Twelve hues are used as fundamental colors. There are yellow, red, blue, orange, green, purple, and six colors obtained as linear combinations of them. Five levels of luminance and three levels of saturation are identified. As a result every color is transferred into one of 180 references colors. After that, clustering in the three dimensional feature spaces is performed using the K-means algorithm [8]. After this step the image is segmented as N regions, and each is presented in extended chromaticity space.

2.2. Low level color characteristics translation into high level color semantic features

The purpose of this phase is to compose more complex image semantic interpretation from those derived through the low-level image analysis characteristics. It is accomplished by applying methods for extracting high level features and recursively applied production rules from a set defined for the correspondent application domain. The rules are defining also the degree of recognition (RD) of a high level semantic feature as a distance among characteristics implied in the rule and those found in the image. RD is calculated with the help of fuzzy measures. An interference mechanism based on backward chaining tries to derive from the low level characteristics more general features and to give a recognition degree to the features recognized.

In this phase a generalized inference mechanism is used. After this step a sequence in the form (1) is obtained:

(1) $$O_{1\,1}(m_{1\,1}, l_{1\,1}), \ldots, O_{1\,S1}(m_{1\,S\,1}, l_{1\,S\,1}), \ldots\ O_{n\,1}(m_{n\,1}, l_{n\,1}), \ldots, O_{n\,Sn}(m_{n\,S\,n}, l_{n\,S\,n}).$$

Such a sequence describes an image with n distinct high level semantic features. The unit $O_{i\,j}(m_{i\,j}, l_{i\,j})$ is a semantic representation of the image feature *i* *(i=1,2, … ,n)* in the *j-th* *(j=1,2, …,* s_j*)* recognition. $m_{i\,j}$ and $l_{i\,j}$ are respectively the RD and the list of attributes of the *i-th* semantic feature in the *j-th* recognition.

To reduce the sequence (1) a procedure similar to Barnett's scheme [1], based on the Dempster-Shafer theory of evidence [5] is applied. The results obtained from applying the production rules are converted into a list of new structures containing information for each semantic feature:

(2) $$O_{11}([Bel(O_{11}, 1\text{-}Bel(not\ O_{11})], l_{11}), \ldots, O_{1q1}([Bel(O_{1\,q1}, 1\text{-}Bel(not\ O_{1\,q1})], l_{1\,q1}), \ldots, O_{n1}([Bel(O_{n1}, 1\text{-}Bel(not\ O_{n1})], l_{n1}), \ldots, O_{\ q1}([Bel(O_{n\,qn}, 1\text{-}Bel(not\ O_{n\,qn})], l_{n\,qn}),$$

where $q_i \le s_i$ (i=1, 2, ..., n).

The function *Bel(Oij, 1-Bel(not* O_{ij}*)]* is a belief function. In such sequences, features interpretations with low belief, according to the user understanding are omitted. The belief function *Bel(*O_i*)* *(i=1,2, …,n)* gives the total amount of belief committed to the features O_i after all evidence bearing on O_i has been pooled. The function *Bel* provides additional information about O_i, namely *Bel(not* O_i*)*, the extent to which the evidence supports the negation of O_i, i.e. *not* O_{ij}.

3. RETRIEVAL BASED ON HIGH LEVEL COLOR SEMANTIC FEATURES

In this section we discuss image retrieval based on high level color properties. It uses the theory formulated by Johannes Itten in 1961 [7]. In this theory color aesthetics may be approached from impression (visually), expression (emotionally) and construction (symbolically). Six different types of contrasts are identified:

1. Contrast of hue. It presents undiluted colors in their most intense luminosity. Some color combinations are: yellow/red/blue, red/blue/green, blue/yellow/violet,

yellow/green/violet/red/, violet/green/blue/orange/black.

2. Light–dark contrast. It is based on comparison of day and night, light and darkness. Rembrandt paintings are often done with such contrast.
3. Warm–cold contrast. Colors or color combinations such as: yellow, yellow-orange, orange, red-orange and red-violet are referred as warm. Colors combinations like yellow-green, green, blue-green, blue, blue-violet are referred as cold.
4. Complementary contrast. Two colors are called complementary if their pigments mixed together yield in neutral grey. Examples are: yellow-violet, blue-orange, red-green. This contrast gives the effect of a stability fixed image.
5. Simultaneous contrast. It results from the fact that for any given color the eye simultaneous requires the complement color.
6. Contrast of saturation. Saturation relates to the degree of purity of the color.

The used 180 colors in our method correspond to the colors in the Runge-Itten sphere. The equators, horizontal and vertical views of the sphere are given in Figure 1.

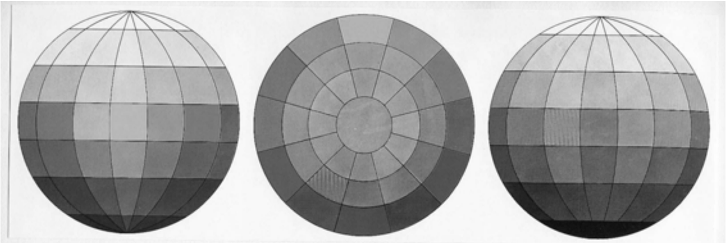


Figure 1. Three views of the Runge-Itten sphere

The color impression is connected with the color effects on our sense of vision. The main color expression properties are the following. Yellow color, the brightest and lightest color suggests power and luminous. It symbolizes knowledge. With combination with dark tones it presents cheerfulness. Red is the color of the cardinals and it is associated with blood and it is always warm. Blue color is always cold and shadowy, and it has retiring nature. Green is the color of the vegetable, the spirit of early summer morning and many different expressions can show by variations in green contrast. Orange color gives radiant feelings. Violet is the color of unconscious, mysterious, chaos, death. The color construction is connected with the juxtapose of two or more colors in such way that they jointly produce a distinct and distinctive expression. Production rules are generated for different painting stiles.

4. THE EXPERIMENTS

The proposed method is a process of realization in a system named "Flint". In our experiments we use an image database obtained with the help of Google image search engine.

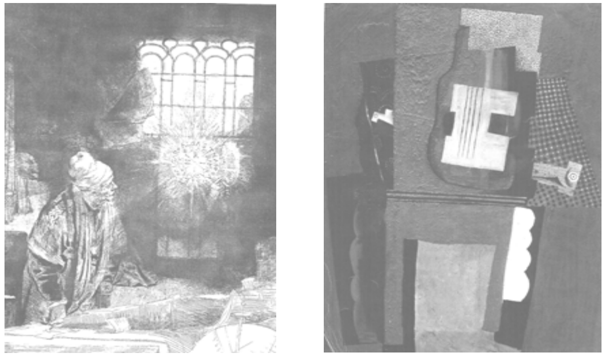


Figure 2. Light-dark contrast: "Dr. Faustus in his study room" sketching by Rembrandt and "Guitar on Mantelpiece" by Paulo Picasso

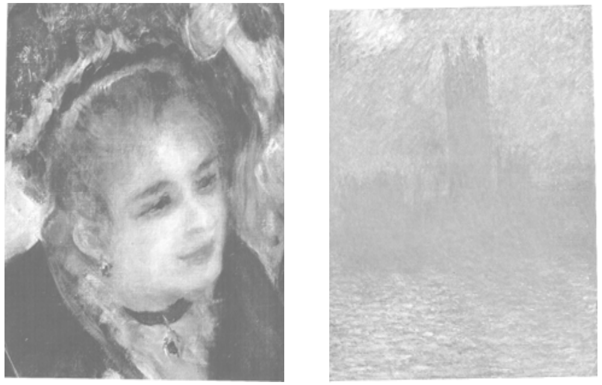


Figure 3. Cold-warm-contrast in "Le Moulin de la Galette" by Auguste Renoir, and "Houses of Parliament" by Clode Monet

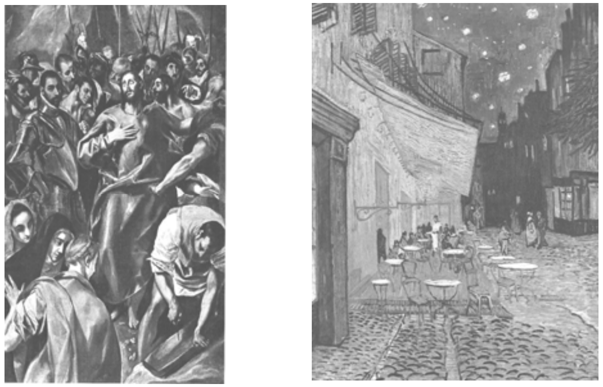


Figure 4. Simultaneous contrast in “Stripping of Christ” by El Greco, and “Café at Evening” by Vincent van Gogh

For retrieval based on color features possible query could be: “Find paintings with the following contrasts: Light-dark, cold-warm, and simultaneous”. Parts of the retrieve painting are given in figures 2, 3 and 4.

5. CONCLUSIONS

The main advantage of the proposed method is the possibility of retrieval using high level color semantic features. After the full system realization we will be able to obtain statistical characteristics about the usefulness of the suggested method.

REFERENCES

[1]. Barnett J., Computational Methods for a Mathematical Theory of Evidence, *Proc. 7-th Inter. Joint Conf. on Artificial Intelligence*, Vancouver, BC, 1982, 868-875.

[2]. Carter, R., Carter, E., CIELUV color difference equations for self-luminous displays, *Color Res. & Appl.*, 8(4), 1983, 252–553.

[3]. Chang, S., Chen W., Sundaram H., Semantic visual templates: linking visual features to semantics, in IEEE International Conference on Image Processing (ICIP’98), Chicago, Illinois, 1998, 531-535.

[4]. Corridoni, J, Bimbo A., Vicario E., Image retrieval by color semantics with incomplete knowledge, Journal of the American Society for Information Science 49(3), 1998, 267-282.

[5]. Gordon J., Shortliffe J., The Dempster-Shafer Theory of Evidence in Rule-Based Expert Systems, in B. Buchanan, E. Shortliffe (edt.), *Mycin Experiments of the Stanford Heuristic Programming Project*, Addison-Wesley Publishing Company, 1984, 272-292.

[6]. Grosky W., Stanchev P., An Image Data Model, in Advances in Visual Information

Systems, Laurini, R. (edt.), Lecture Notes in Computer Science 1929, 2000, 14-25. [m2]

[7]. Itten, J., *The art of colors*, Reinhold Publishing Corporation of New York, 1961.

[8]. Jain, A., *Algorithms for clustering data*. Englewood Cliffs, NJ, Prentice Hall, 1991.

[9]. Rabbitti, F and Stanchev, P., GRIM_DBMS: a graphical image database management system, in Visual Database Systems, Kunii, T. (edt.), Elsevier, Amsterdam, 1989, 415-430.

[10]. Sethi, I., Coman, I., Day, B., Jiang, F., Li, D., Segovia-Juarez, J., Wei, G., and You, B., Color-WISE: A System for Image Similarity Retrieval Using Color, Proceedings of SPIE Storage and Retrieval for Image and Video Databases, Volume 3312, February 1998, 140-149.

[11]. Stanchev P., Using Image Mining for Image Retrieval, IASTED International Conference “Computer Science and Technology”, May 19-21, 2003, Cancun, Mexico.

LEXICON DESIGN FOR SEMANTIC INDEXING IN MEDIA DATABASES

Apostol Natsev, John R. Smith, Milind R. Naphade

IBM T. J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532, US.A.
Email: {natsev, jsmith, naphade}@us.ibm.com

Abstract: Semantic understanding of multimedia content is important for describing, indexing, and accessing digital content effectively. Semantic descriptors of multimedia content enhance the value of media assets by enabling more effective search for re-purposing, licensing, or commerce purposes. Automatic content analysis at the semantic level is however a daunting task, and most media repositories resort to low-level feature descriptors or manually ascribed annotations. Alternatively, recent techniques have been developed for detecting simple generic concepts in video, such as indoor, outdoor, face, people, nature, and so on, but these labels directly support only a small set of queries.

In this paper we describe a novel framework for capturing existing semantic information - such as manual annotations, collaborative filters, or explicit statistical models - and leveraging such information for general-purpose automatic multimedia processing at the semantic level. In particular, we present an indexing method for describing objects succinctly in terms of their membership to existing known semantic classes. We study the model vector object representation as a low-dimensional and computationally inexpensive semantic descriptor, and we focus our attention on model vector construction issues, such as the design of semantic lexicons. We propose and evaluate several lexicon design methods, and compare them with low-level feature descriptors on a large video corpus. The empirical results demonstrate significant performance gains in retrieval effectiveness.

1. Introduction

Recent advances in content analysis and classification are improving capabilities for effectively searching and filtering of multimedia content. However, a significant gap remains between the low-level feature descriptions that can be automatically extracted, such as colors, textures, shapes, motions, etc., and the semantic descriptions of objects, events, scenes, people and concepts that are meaningful to users. Thus, the problem of semantic indexing of multimedia content is of prime importance for its effective utilization.

In this paper we present a *semantic model vector framework* as a novel, efficient, and effective solution for capturing and leveraging existing semantics for semantically enhanced processing of multimedia content. The basic idea is to describe multimedia content in terms of its membership to a few fixed semantic classes, and then to leverage the above semantic description in tasks such as retrieval, classification, visualization, and data mining.

We argue that in many applications, the size of the explicitly modeled semantic lexicon is limited due to high manual labor requirements, storage or computational costs, or the scarcity of training data. In such situations it may be more reasonable to model rare concepts or compare documents based on their relationship with few well-known concepts. We therefore propose and show that it is better to train and leverage a smaller number of robust detectors, than a large number of mediocre ones. In light of that, we investigate several methods for selecting a good *semantic basi*s, and we evaluate them empirically on a large video corpus. The experiments show significant retrieval effectiveness improvements of the model vector approach, as compared to traditional content based retrieval approaches.

1.1. Proposed approach

Given that multimedia indexing systems are improving capabilities for automatically detecting some concepts in multimedia documents, new techniques are needed that leverage these classifiers or other existing semantic information in order to extract additional semantic descriptors. Our proposed approach is very intuitive-given a fixed set of known semantic classes, we describe unknown content in terms of its containment relationship to the known classes. The system works by scoring each multimedia document with respect to a set of concept detectors. The resulting detection scores are mapped to a fixed multi-dimensional vector space. As an example application, semantic retrieval may then be carried out by searching the model vector space and identifying nearest neighbors to a given query model vector.

The above idea is illustrated in Figure 1, which shows the model vector representation of an unknown video in terms of its relationships to a fixed set of categories, or movie genres in this case. The advantage of the model vector representation is that it captures the concept labeling broadly across an entire lexicon. It also provides a compact low-dimensional and storage-efficient representation that captures the uncertainty of the labels and enables efficient indexing in a metric space. Once model vectors are extracted, complex similarity operations can be replaced by simple vector operations in the semantic space, which enhances the scalability of the database system. This allows development of efficient and effective systems for similarity searching, relevance feedback searching, classification, clustering, etc., that operate at a semantic level rather than at the audio-visual feature level. More details on the model vector representation can be found in [1].

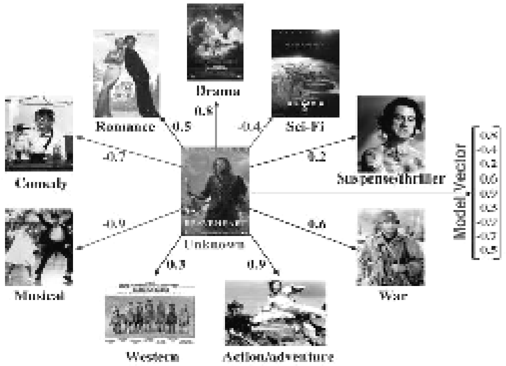


Figure 1: Model vector representation for semantic indexing.

2. Model vector construction

Construction of model vectors may be done in a variety of ways, including statistical modeling, collaborative filtering, or even manual annotations. Our specific approach involves two stages of processing [1]: (1) *a priori* learning of statistical models and (2) concept detection and score mapping to produce model vectors. These steps are only briefly summarized here.

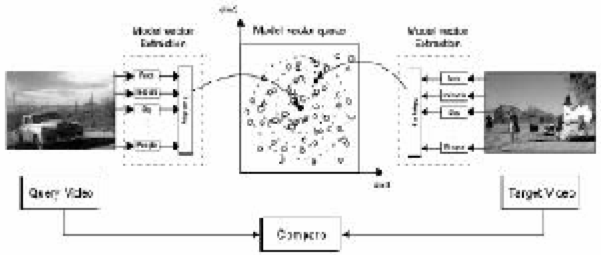


Figure 2: Semantic similarity searching through model vector-based retrieval.

The concept learning process uses labeled examples as training data for building statistical models of semantic concepts. We construct a set of *N* binary detectors, each corresponding to a unique concept in a fixed lexicon.

The detectors may take any number of forms including Support Vector Machines (SVM), GMM, HMM, Neural Nets, Bayesian Nets. For our purposes, we have

investigated SVMs for visual concept modeling, as described in [2].

Once the *N* detectors are constructed, multimedia documents are analyzed, classified and scored by each detector. For each of the *N* detectors, a confidence score is produced for each multimedia document that measures the degree of certainty of detection of the corresponding concept. Once the concept detectors are evaluated, the scores are mapped to produce model vectors, as illustrated in Figure 2. In the simplest case, a one-to-one mapping of *N* confidence scores is achieved by concatenating the confidence scores to build an *N*-dimensional vector*. Once mapped to this semantic vector space, documents can be compared, searched, classified, or mined by means of the corresponding vector-space techniques. For example, simple Euclidean distance between semantic model vectors can be used in place of expensive and complex semantic (dis)similarity measures.

3. Semantic lexicon design

A critical issue in the construction of model vectors is the design of the semantic lexicon to be used as a basis for semantic detection. The lexicon lists all semantic concepts which will have corresponding detectors. One approach includes the use of fixed lexicons such as MPEG-7 Classification Schemes (CS) or the Library of Congress Thesaurus of Graphical Material (TGM) which provides a set of categories for cataloging photographs and other types of graphical documents. Such fixed lexicons may contain thousands of entries, however, and an important consideration is that they should be modeled relatively accurately. With statistical learning approaches, this typically translates into a requirement for having a sufficient number of training examples for each concept in the lexicon. Other requirements pertaining to the number of lexicon entries include reasonable computational, storage, or manual labor costs. Another important criterion is that the semantic classes should be as distinct as possible and cover a significant portion of the semantic space so that they can characterize the content more broadly and effectively. The complexity of the above issues highlight the need for fully automatic or assisted manual methods for lexicon design.

While resolving all of the above issues in an automated way is a very challenging problem, we adopted the following practical approach. Our lexicon was designed for a 25-hour video corpus provided as part of the TREC-2002 Video Retrieval benchmark**. The entire set was semi-automatically annotated using an initial lexicon of concepts that were coarsely determined to have at least some minimal support in the data set. Based on the refined statistics gathered through the precise annotation process, as well as the complexity and the (subjective) importance of each semantic

* Alternatively, the scores may be normalized prior to forming model vectors or they can be processed using dimensionality reduction techniques such as Latent Semantic Indexing.

** http://www-nlpir.nist.gov/projects/t2002v/t2002v.html.

class label, we finally selected 32 semantic concepts to model explicitly with SVM detectors. The concepts were modeled using multiple low-level features, such as color, texture, and edge, sometimes resulting in multiple models for the same concept. We thus built a total of 74 SVM models, comprising our initial lexicon, L_{All} , for model vector construction purposes.

The lexicon design decisions up to this point - including the choice of initial labels for annotation purposes, initial set of concepts for modeling purposes, as well as the final set of models - were all made subjectively and with human intelligence. From here on, however, we considered several fully automatic and systematic ways of designing the final lexicon for model vectors.

The first method for lexicon design was based on the observation that the majority of the models were undertrained and exhibited poor performance as detectors. We observed a very strong correlation between the number of training examples and the detector classification performance on an independent validation set. We therefore decided to consider only the set of robust models, which were defined as models built from a minimum number of training examples (300 in our case). We therefore picked only frequently occurring concepts in the training set, resulting in a set of 12 robust models, L_{Robust}.

Lexicon L_{Robust}	Lexicon $L_{Distinct}$	Lexicon L_{Random}
Building	Transportation (edge)	Transportation (edge)
Face	Building	Car (edge)
Greenery	Train	Transportation (texture)
Indoors	Office Setting (edge)	Office Setting (edge)
Landscape	Tractor (edge)	Household Appliance (edge)
Outdoors	Tractor (texture)	Chicken (edge)
People	Household Appliance	Office Setting (texture)
Person	Boat	Airplane
Road	Beach	Sky
Sky	Airplane (edge)	Cloud (edge)
Transportation (texture)	Flower	Smoke
Tree (texture)	Road	Farm

Table 1: Three lexicons of reduced dimensionality constructed from the proposed lexicon design methods.

Since we wanted to investigate the importance of the basis selection method, we formed another 12-dimensional subset, $L_{Distinct}$, consisting of the 12 most distinct models, to compare against the performance of L_{Robust}. The motivation for this lexicon design method is based on the heuristic that more distinct semantic models will lead to a broader and more complete description of the multimedia documents, and will therefore lead to better object differentiation and retrieval performance. The set of most distinct models was formed by measuring model correlations from their detection outputs on a validation set, and defining an appropriate distance between models. In particular, each model was mapped to a high-dimensional vector space, where the i-th model coordinate consists of the detection score of the model on the i-th image in a validation dataset. The distance between two concept models was calculated based on the cosine of the angle between the corresponding normalized high-dimensional vectors. Having defined a proper distance metric for semantic models, we used a greedy algorithm to select the 12 most distinct - or spatially spread-out-models, forming lexicon $L_{Distinct}$.

The final method for basis selection that we considered was based on random sampling of the full lexicon L_{All}, producing a 12-dimensional random subset L_{Random}. To summarize, Table 1 lists the contents of the three reduced dimensionality lexicons. The complete set of models from lexicon L_{all} is not shown here due to space considerations.

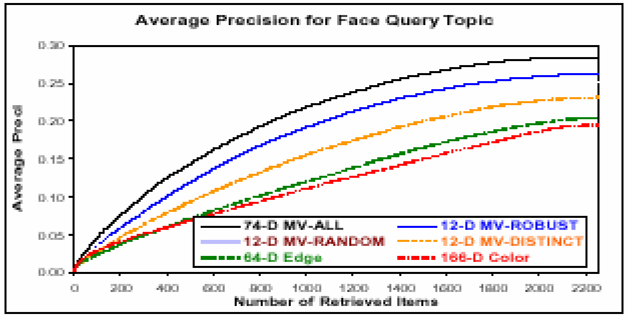


Figure 3: Retrieval performance comparison of several retrieval approaches: 166-D color histograms (“Color”), 64-D edge histograms (“Edge”), 74-D model vectors (“MV-ALL”), and three 12-D model vector-based features using different basis selection methods (“MV-ROBUST”, “MV-DISTINCT”, and “MV-RANDOM”).

4. Experiments

In this section we evaluate the performance of model vector-based retrieval against content-based retrieval methods based on purely visual features. We also evaluate and compare the different lexicon design methods.

For training and validation purposes, we annotated 24 hours of video content provided as part of the TREC 2002 Video Retrieval benchmark*. 19 hours (9,495 video clips) were reserved for training SVM models using low-level visual features (for details, see [2]) and constructing model vectors based on the above lexicons.

* http://www-nlpir.nist.gov/projects/t2002v/t2002v.html (TREC-2002 Video Retrieval Track).

The remaining 5 hours, or 2,249 video clips, were used for testing the retrieval effectiveness of the model vector-based and content-based retrieval approaches.

We conducted retrieval performance experiments for several query topics and averaged the results over 20 queries with a random positive example as a query for each topic. The average retrieval effectiveness is computed over all random searches using each method for each topic. Figure 3 plots the Mean Average Precision (MAP) *vs.* the retrieval scope, or cutoff point, over the sequence of searches for the "Face" topic. The MAP score corresponds to the weighted area under the Precision *vs.* Recall curve. Table 2 summarizes the MAP scores for all topics comparing the different search methods.

Topic	Color	Edge	ALL	ROBUST	DISTINCT	RANDOM
(#Relevant)	Hist.	Hist.	Models	Models	Models	Models
Car (206)	**0.16**	0.14	0.19	**0.18**	0.16	**0.18**
Face (427)	0.19	**0.20**	0.28	**0.26**	0.23	0.25
Indoors (416)	**0.24**	0.23	0.28	**0.26**	0.24	0.24
Landscape (117)	0.09	**0.10**	0.14	**0.14**	0.11	0.12
Road (341)	**0.23**	0.19	0.26	**0.25**	0.23	**0.25**
Sky (614)	**0.35**	0.30	0.36	0.37	0.30	**0.38**
Tree (319)	**0.19**	0.18	0.20	**0.20**	0.18	0.19

Table 2: MAP score comparison of different retrieval methods over several query topics. Bold numbers indicate best performance in same category.

Based on the experiments, we can make the following observations:

- The compact 12-dimensional model vectors outperform the higher-dimensional color and edge based descriptors by 5% to 40% in MAP.
- The 12-dimensional model vector descriptors are nearly as effective as the full 74-dimensional model vectors.
- The lexicon of robust models leads generally to best performance.

5. Conclusions

We investigated a novel framework for semantic indexing of multimedia content using model vectors, or vectors of confidences with respect to a known set of semantic classes. The model vector approach uses a fixed concept lexicon as a basis to provide semantic descriptors of multimedia content. The semantic descriptors can then be used in a variety of semantic indexing applications, including similarity retrieval, relevance feedback search and semantic concept classification. In this paper we considered the specific problem of lexicon design—or semantic basis selection—for model vector construction purposes, and investigated several lexicon design methods. The proposed methods were evaluated and validated empirically on a large video database.

REFERENCES

[1] J. R. Smith, M. R. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *IEEE Intl. Conf. on Multimedia and Expo (ICME '03*), Dec. 2003.

[2] IBM TREC-2002 Team. IBM Research TREC-2002 Video Retrieval System. In *Proc. Text Retrieval Conference (TREC '02*), Gaithersburg, MD, Nov. 2002.

FROM MULTIMEDIA DOCUMENT COLLECTIONS TO LEARNING MATERIAL

Eva Heinrich, Frederic Andres**, Kinji Ono***

** Massey University, Palmerston North, New Zealand, E.Heinrich@massey.ac.nz,*
*** National Institute of Informatics (NII), Tokyo, Japan, {andres,ono}@nii.ac.jp*

Abstract: This paper introduces an approach, called stories, to build a bridge between annotated document collections and the use of these documents to construct learning material. The main components of a story are its narrative, the links to specific sections within documents and metadata attached to stories that allow customization of stories according to users' needs and profiles. The story metadata are based on MPEG7 and its Semantic DS including additional attributes on instructional information and semantics of the story narratives. These generic, story metadata are complemented by domain specific metadata based on the domain of the document collection annotated.

The paper explains the story concept and sets it in context to research on learning objects, metadata and educational modeling languages. The structure of stories and their metadata are then discussed in more detail and the process of writing and using a story is outlined.

Keywords: Digital Multimedia document Collection, Annotation, Story

1. Introduction

The context for the research presented in this paper is the UNESCO project on Digital Silk Roads (Ono, 2002). This project aims at the preservation of cultural heritage utilising digital technologies. A wide variety of techniques are used to present the cultural heritage: ancient maps and drawings are digitised, 3D animations of artefacts are constructed, images and videos are recorded to show the present condition of historic sites, historic documents are analysed. The resulting document collection needs to be annotated using metadata to facilitate the searching for documents. This is partly done with standardised meta data schemes like Dublin Core (DC, 2002) or LOM (2002) and partly with specifically developed meta data schemes like the multidimensional meta data schemes developed in Ono et al. (2002) that include dimensions like economics, religion, architecture or geography.

From the perspective of eLearning a project like the Digital Silk Roads project delivers an enormous repository of potential learning objects. The challenge the individual teacher faces is to locate the appropriate information. This task is compounded by several factors: the shear number of documents of various types and content, the distribution of documents and their meta data across multiple sites, the limitations of standardised meta data, the lack of context for standardised meta data, the restrictions in time available to 'surf' and search for resources, the variety of languages connected with such a project, or possibly the lack of domain knowledge in highly specialised areas.

The research presented in this paper aims to address some of the issues mentioned to facilitate access to information in contexts of large collections of documents relating to a common subject area as given in the example of the Digital Silk Road project. The idea is to introduce an interpretative, semantically rich layer, informally called 'stories', between document collection and learning material that links documents according to themes. One motivation behind this approach is to add a more focused, semantic layer on top of the untargeted metadata that are commonly used to describe single documents. Speaking from an eLearning context the stories build on learning objects and become information sources and building blocks for learning material.

The next section outlines the background of choosing the term 'story' and through this discussion emphasises the thinking behind the concept that is explained in its main aspects thereafter. This is followed by a brief review of the grounding of the stories in the literature. Then, the structure of stories and their annotation metadata are introduced, followed by a section on authoring and use of stories. A summary and plans for future research conclude the paper.

2. Why the term stories was chosen

Stories are a form of narrative discourse. Scholars from fields like literature or psychology have provided a variety of definitions of what constitutes a narrative. Graesser et al. (1991, p174) have given the following general definition: 'Narratives are expressions of event-based experiences that (a) are either stored in memory or cognitively constructed, (b) are selected by the teller/writer to transmit to the audience/reader, and (c) are organized in knowledge structures that can be anticipated by the audience'. The authors continue to discuss among other topics 'characters', 'temporal and spatial placements', 'points, morals and themes', or 'points of view and perspectives'. While certainly not all common characteristics of a narrative apply to our view of the 'stories' we want to employ to support the use of document collections like in the Silk Road project, there are certain characteristic that made us choose this term:

- The writer of a story has a message for the audience.
- The story is set in a temporal and spatial context that helps readers to orientate themselves.
- A story is written from a certain point of view.
- Every person is intuitively familiar with the common structure and the

concept of a story.

These characteristics support our idea of providing an interpretive, semantic layer on top of document collections that sets these documents into context and gives a message to the reader. The term 'story' further matches well the call for 'subjective' meta data which we will outline in later sections of this paper.

3. The story concept

A story consists of three main elements: the narrative, links to learning object segments and metadata. The narrative is textual data that the author of the stories writes to tell the story. There are no restrictions on what the author can write. The author can use the narrative to tell facts, provide interpretations, make comparisons, draw attention or similar. The narratives of stories are structured into units and paragraphs. Within a unit a line of argument will be preserved. Units can be used to tell different aspects of a story.

The links to learning object segments are embedded into the narrative. The purpose of these links is to relate the narrative closely to the underlying documents. The documents can provide examples, illustration, proof, further explanation, additional material or similar. It is important to note here that the links refer to particular segments in the documents. As the documents, being learning objects can be of any size it is important to connect the narrative with the appropriate part of the document. Depending on document type, this specific part can be, for example, a segment of a video clip, a paragraph in a text document or a specific area in an image. Further it is important that, while the links refer to specific document sections, still the whole document is reachable via the link. This is done to preserve context and provide the reader with the flexibility to venture outside the boundaries of the story.

The role of the story metadata is to make it possible to search for a story and to customise a story according to personal needs. Associated with a story are two different types of metadata. Firstly, a story can be seen as a learning object itself and as such can be described by metadata of LOM or Dublin Core format. The purpose of these metadata is to allow users to locate suitable stories from repositories. A second form of metadata serves users for selecting individual pathways through a story. These metadata can be attached to story units or paragraphs. The type of information stored relates to issues like narrative structure of the story, subject area, dimension, type of argument, or language the story is written in. These metadata serve to include multiple perspectives in a story. The reader of a story can use these metadata to find different pathways through a story or to focus on specific aspects provided in the story.

4. The foundation for the story concept in the literature

As any new concept the story concept needs to be set in context to other

research. This is done in this section of the paper in abbreviated form. A fuller discussion can be found in a parallel paper (Heinrich and Andres, 2003).

In the areas of electronic document collections or eLearning a variety of metadata schemes have been defined, ranging from Dublin Core and LOM to educational modelling languages. Dublin Core is a very general standard and we can use Dublin Core metadata to describe stories seen as one document. LOM is used to annotate learning objects to support re-usability in eLearning. We want to use LOM similar as Dublin Core for the description of stories as learning objects but we as well want to address some of the criticisms voiced on LOM in the literature:

- The definition of a learning object is very wide and provides no restriction or guidelines on the size of a learning object (Wiley, 2000).
- LOM contains no reference to instructional models and theory (Allert et al., 2002).
- The definition of metadata without context is very difficult (Shabajee, 2002).
- LOM does not cater for subjective metadata that become increasingly important as personalisation is seen as key element of learning (Hodgins, 2001).
- Taxonomy for learning objects is required to apply instructional design theory (Wiley, 2001).

The story concept aims at addressing these criticisms. A story contains references to individual documents or learning objects. As these documents can be of unlimited size the reference in the story points to a segment within the document. This means that even for large documents the reader of a story can be referred to specific sections of relevance. The context to the overall document is preserved, as the reader of a story is free to move beyond the referenced segment. The story meta data include instructional attributes that can be used to label units of the story narratives and that can be placed with the links to referenced documents. A story is written with a theme or message to the reader in mind. The narrative of a story links documents that are related to this theme and therefore provides the context required specifying and successfully using metadata. The narratives of a story provide the space for subjective meta data. The author of the story can provide interpretations and points-of-view and is not restricted to a single set of keywords or short descriptions for each learning object. Using Wiley's taxonomy (Wiley, 2000) it can be suggested that stories can be classified as generative-presentation learning objects.

The idea behind learning objects is that these can be re-used and combined into learning material. To create learning material from learning objects the instructional goal has to be specified and a learner context that considers elements like learner preferences, pre-knowledge, course level, course duration or delivery mode has to be provided. Buendia et al. (2000) talks about the 'learning scenario' that includes learning modes, timing schedules, instructional modes and learning goals. Educational modelling languages model learning material and not just learning objects by addressing these properties of learning contexts.

Educational modelling languages address the learning context-specific properties of learning material. Stories are not seen as learning material but as a resource for teachers to assist in the creation of learning material. The intention behind the story concept is not to address a specific learning context but to create rich resources that link thematically related learning objects and provides additional information about subject area and content of these learning objects. As stories do not address the instructional context but stay on learning object level their purpose is different from educational modelling languages.

Figure 1 provides a picture indicating the place of stories in relationship to metadata and re-use of learning objects. It is based on the approach of Shabajee (2002) that shows the 're-purposing' of educational documents. Meta data are added to raw documents and these two components are stored together in a database. Through searching of the meta data suitable raw documents, called 'assets', are identified and then, in a step of re-purposing, combined into new composite objects. The idea of stories means to add one step into this process. The stories are placed between the assets with their meta data and the composite objects. The stories contain references to the assets and narratives that describe or interpret the assets (more exactly specific segments within the assets) or create semantic relationships between assets. The stories, that themselves carry meta data, are searched for and build, together with the assets they refer to, the building blocks for the composite objects.

5. The structure and annotation of a story

A story has to be structured and annotated by metadata to be searchable and customisable. We are currently working on a XML schema to define this structure. The schema is based on MPEG7 (2002) description schemes and here specifically on the Linguistic DS, which is a description tool for linguistic annotation to be included in MPEG-7 MDS Amendment 1 in 2003. The Linguistic DS is based on the Global Document Annotation, GDA; tag set that is described in detail in GDA (2002). GDA supports detailed linguistic tagging of textual documents that facilitates knowledge representation or applications like automated document summarisation. For our stories we initially will not make use of the full power of GDA but restrict our mark-up to annotate the structure and language of a story, capture links to multimedia document segments and denote special terms like places, times, dates. In particular we annotate for the following characteristics using elements of the Linguistic DS and MPEG7:

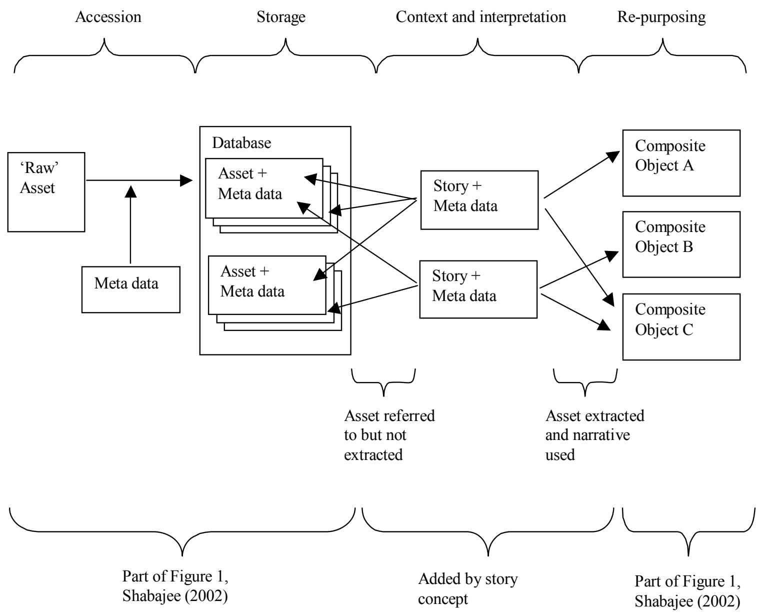


Figure 1: Extension of the 'Re-purposing' concept provided by Shabajee (2002)

- Document structure: part, chapter, section, subsection, headings, paragraphs, sequences of sentences, sentences, quotations;
- Language: definition of both the language of the narrative and the language of special terms;
- Special terms: dates, times, periods, names, personal-, geographic-, organisational-names, numbers, addresses, bibliographical references;
- Reference to multimedia document sections: reference to media locator, start and length of a segment; description of audio, visual, audio-visual and multimedia segments.

These annotation categories taken from the Linguistic DS and MPEG7 provide us with information about the structure of the story narratives and the links to the segments of the underlying documents. To achieve our goals we have to extend these description schemes in terms of narrative structure and instructional information. We want to provide instructional information that is sufficient to direct a teacher to appropriate parts within a story but we want to keep the instructional information simple as the authors (and probably to a large degree as well the readers) of our stories are content and not instructional theory experts, a thought that is supported for example by Sampson et al. (2002). Our current thinking is to follow Merrill's work on instructional transaction theory (Merrill et al., 1996) and annotate for instructional transactions and instructional types.

The Linguistic DS provides us with the facility to structure the story narratives into units like chapters, sections or paragraphs. We add some semantic to this structure by defining the type of a chapter or section. Depending on the preferences of the author of a story the narrative can be less formal, more like a story as commonly understood as a literary form, or more formal, like academic writing in a journal article. Accordingly we suggest different sets of attribute values:

- For a 'informal' type story: setting, beginning, goal, outcome, ending, …
- For a 'standard' document type story: introduction, main body, summary, …
- For a 'academic' type story: table-of-contents, abstract, introduction, main body, conclusion, references …

Given this type of semantic information an instructor can, for example, focus on the introductions of several stories that prior have been selected via their high-level, DC-type meta data, and can quickly find the most appropriate among these stories.

6. Authoring and using a story

Stories are to be authored by domain experts that are familiar with both specifics of the domain and of the document collections they are discussing in the stories. There are a number of reasons for using domain experts for writing the stories. As implied by the term, domain experts possess specialist knowledge. This is required to write stories of high quality and correctness. In the example of the Silk Road Project expert knowledge will be required to discuss highly specific topics like the caravanserais or guesthouses. Additionally, special language skills might be necessary to read original material written in one of the many languages related to the Silk Road countries. A teacher of history at a high school is unlikely to possess this type of specialist knowledge but would benefit from learning from the explanations and interpretations provided by the experts.

Despite the increasing use of meta data it is commonly accepted that the locating of suitable documents can be very difficult and is certainly time consuming. Domain experts, who might even be involved in document collection and annotation, will acquire intimate knowledge about data repositories and meta data structures of their specialist area. This knowledge will assist in finding suitable documents within a reasonable time frame.

As outlined earlier, various meta data are associated with a story. There are the meta data that describe a story seen 'from the outside' as one unit or learning object and there are the meta data the describe a story 'from the inside'. Formulating the latter type of meta data can be partly supported by a story-writing software environment and by domain specific dictionaries.

A story-writing software environment could provide an authoring interface that semi-automatically defines the structural units of a story and allows to 'labelling' these with attribute values for semantic type of unit, instructional transaction or instructional type. Meta data for the links to document segments can be generated once the story author has made the appropriate selections in, for example, a graphical user interface. As the Semantic DS contains support for multiple languages, sections of stories can be written in various languages and annotated accordingly.

The meta data discussed earlier in this section stem from the concept of stories and are generic in terms of specific content domains. The generic story meta data are complemented by domain specific meta data. The creation of these meta data can be supported by multi-dimensional meta data as suggested in Ono et a. (2002) and by dictionaries of domain specific terms, place names, periods specifiers or similar.

Once a story is written and annotated with meta data it has to be stored in a publicly accessible repository. A teacher or instructor locates a story via its story level metadata from a repository of stories. The arguments why it is easier to access relevant information via the intermediate step of stories as compared to searching for documents (or learning objects) directly are as follows. A story is an entry point to a large wealth of information. A story is annotated by meta data that are richer both in terms of instructional and content specific information than DC-type or LOM-type meta data. Once several possible stories are located these story internal meta data help to locate story sections and to quickly decide if and in which section the story provides the desired kind of information. The story narrative provides information and interpretation related to the referenced documents and sets the related documents in context. Additionally, by providing pointers to segments within documents the specific points of interest in a potentially large document are found easily. This entire means that once a relevant story is located access to a whole range of suitable information (and not to only one document) is provided.

7. Conclusions and future work

The 'story' concept introduces a new semantic layer between collections of learning objects and learning material. The story narratives link semantically related learning objects. As learning objects by definition are not restricted in size the links from the story narrative refer to specific segments within the learning objects to guide the reader precisely to relevant sections. Stories provide the context that is required for the specification and use of metadata. The story narratives can be seen as rich metadata that annotate the referred learning objects. Story specific metadata and domain specific metadata allow for the customisation of stories according to the needs of individual users. The stories themselves form new learning objects that provide annotation to thematically linked learning objects stemming from large document collections.

The story metadata are based on MPEG7 for the links to the multimedia document segments of the referenced learning objects and on the MPEG7 Semantic DS for the structure of the story narratives. These metadata are complemented by attributes for instructional information and story semantics. We envisage that the stories will be written by domain experts and used by teachers. The term “stories” was chosen to express the desire to communicate via a medium, the story, that is familiar to everyone and therefore can be easily authored and easily understood. The story narratives provide the mechanism required expressing knowledge and interpretations in a natural way, the story metadata deliver the precision for search and retrieval both of story sections and underlaying documents.

The story concept aims at application in large document collections like they are provided by the Silk Road project for the preservation of cultural heritage utilising digital technologies. Work is currently undertaken to specify multi-dimensional metadata definitions for the caravanserais or guesthouses of the Silk Road and to collect these data from experts. The next step will be for caravanserais experts to write stories that set the vast amount of single documents or learning objects in context to make them more easily accessible by teachers.

REFERENCES

Allert H., Dhraief, H., Nejdl, W. (2002). How are Learning Objects Used in Learning Process? Instructional Role of Learning Objects in LOM. Proceedings of EDMedia2002. P. Barker, S.Rebelsky (Eds.), AACE, Denver, Colorado, USA, 40-41.

Buendia, F., Diaz, P., Benlloch, J. V. (2002). A framework for the instructional design of multi-structured educational applications. Proceedings of EDMedia2002. P. Barker, S.Rebelsky (Eds.), AACE, Denver, Colorado, USA, 210-215,.

DC (2002). Dublin Core. http://www.dublincore.org/groups/education.

GDA (2002). Global Document Annotation. http://www.i-content.org/GDA/.

Graesser, A., Golding, J. M., & Long, D. L. (1991). Narrative representation and comprehension. In R. Barr, M. L. Kamil, P. Mosenthal, and P. D. Pearson (Eds.). Handbook of Reading Research (Volume II, 171-205). White Plains, NY: Longman.

Heinrich, E., Andres, F. (2003). Enriching Document Collections through the Writing of ‘Stories’. Submitted to WWW2003 Conference, Education Track.

Hodgins, H. W. (2001). The Future of Learning Objects. . In The Instructional Use of Learning Objects (online version). D.A. Wiley (Ed.). Available at http://reusability.org/read.

LOM (2002). IEEE Learning Technology Standards Committee (LTSC). http://ltsc.ieee.org/wg12.

Merrill M. D., & ID2 Research Team (1996). *Instructional Transaction Theory: Instructional Design based on Knowledge Objects*. Educational Technology, 36(3), 30-37. Available at http://www.id2.usu.edu/Papers/Contents.html.

MPEG7 (2002). MPEG7 Overview. http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

Ono, K. (2002). (Ed.). Proceeding of the Tokyo Symposium for Digital Silk Roads. National Institute of Informatics, Tokyo, Japan.

Ono, K., Satoh, S., Andres, F., Katayama, N. (2002). Creating a global Multimedia Repository Framework for Cultural Heritage with Advanced Information Technology. Proceeding of the Tokyo Symposium for Digital Silk Roads. Ono, K. (Ed.). National Institute of Informatics, Tokyo, Japan.

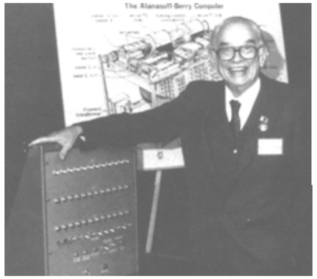
Sampson, D., Karagiannidis, C. (2002). From Content Objects to Learning Objects: Adding Instructional Information to Educational Meta-Data. Proceedings of IEEE International Conference on Advanced Learning Technologies. V. Petrushin, P. Kommers, Kinshuk, I. Galeev (Eds.), Kazan State Technological University, Tatarstan, Russia. pp 513 – 518.

Shabajee, P. (2002). Primary Multimedia Objects and ‘Educational Metadata’. D-Lib Magazine, Vol 8, No 6. ISSN 1082-9873.

Wiley, D. A. (2001). Connecting learning objects to instructional design theory: A definition, a metaphor and a taxonomy. In The Instructional Use of Learning Objects (online version). D.A. Wiley (Ed.). Available at http://reusability.org/read.

Information Matherials

About John Atanasoff



John Atanasoff was born in Hamilton, New York State, on October 4, 1903. He was the first of 10 children in the family of a Bulgarian immigrant from the village of Boyadjik, Yambol Region. He earned his Bachelor's Degree at Florida State University (1925); his Master's Degree at Iowa State University (1926) and his Ph.D. in Physics and Mathematics at Wisconsin University, Aims, as a professor of Mathematics at first and of Physics later.

In the late 1930s, he sought a way of facilitating linear algebraic calculations (solving large systemic linear algebraic equations) which he needed in his research work. He rejected the analogue type machines as too slow and inaccurate, and focused on the digital approach. In December 1939, he created the prototype of the first electronic calculating machine (finished in 1942) together with his student Clifford Berry.

Four entirely new principles of work were applied in it: the binary system, the regenerative memory, logical schemes as elements of software and electronic components for storing data.

After 1945, he devoted his professional career to various governmental and industrial projects, being at the head of two companies founded by him.

John Atanasoff met John Mauchly who was interested in his computer in 1941. What exactly happened then? 26 years later, that issue was looked into at the lawsuit which had to determine whether John Mauchly and J. Presper Eckert used without permission Atanasoff' s invention to build the first (officially recognized until then) electronic computer, ENIAC, in 1942-1946. John Atanasoff was the main witness at the trial. The court, headed by Federal Judge Earl R. Larson, pronounced categorically a verdict in favor of John Atanasoff who was declared the father of the computer.

"Eckert and Mauchly did not invent the first electronic digital computer but borrowed the basic idea for it from John Atanasoff." (Verdict of the Court of Minneapolis, 1973)

In 1970, John Atanasoff was invited to visit Bulgaria by the Bulgarian Academy of Sciences and was awarded the Cyril and Methodius Order First Class. This was his first award and public recognition. But the biggest prize in his life was received from the hands of President Bush – the National Medal of Science and technology.

Association for the Development of the Information Society

Acad. G. Bonchev Str., 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

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: Plovdiv University 'Paisii Hilendarski,' Shoumen University 'Konstantin Preslavski,' Technical University—Gabrovo, Southwestern University 'Neofit Rilski' (Blagoevgrad), National Sports Academy (Sofia), High Institute for Officers' Training and Scientific Activity of the Ministry of the Interior (Sofia), High Military School 'Panayot Volov'—Shoumen, the 'Informatics' Section of the Union of Scientists in Bulgaria, the Institute of Mathematics and Informatics, the Institute of Information Technologies, and the Central Laboratory of Computer Security of the Bulgarian Academy of Sciences (Sofia), and other organisations. Societies in the cities of Plovdiv (the second-biggest in Bulgaria), 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, fora, 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 seven years the main organiser of the international conference *Information and Communication Technologies and Programming.*

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 citisents 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.

Fortieth Anniversary of the **Department of Information Research**

The department originates from the group on Computer-Aided Programming formed in July 1963 within the Numerical Methods Department of the Mathematical Institute with Computer Center, Sofia. This group became the basis on which on November 1, 1968 the Department of Computer-Aided Programming and later the Computer Science Department were established. The department was renamed in 1995 to its present name: Department of Information Research.

The head of these units during their development has been and currently is prof. Petar Barnev.

As the first of its kind in Bulgaria, the department carried out research on a wide range of directions: programming methods and languages, compilation methods, operating systems, databases, information service systems, computer graphics, computer modelling, office automation, decision support systems, computational linguistics, telecommunication systems, issues of education in informatics, etc. For many of these directions, research teams were formed in the department which later became independent departments: Computer Graphics, Artificial Intelligence, Telecommunications, and the Laboratory of Applied Mathematics (the latter in Plovdiv).

Research in the field of Cognitive Science especially the reasoning by analogy was undertaken at the department which later allowed the establishment of the Department of Cognitive Science at the New Bulgarian University. At present, the major interests of the department staff are in the fields of information theory, general theory of information, multiagent systems, multimedia, automated analysis and synthesis of information materials, cognitive modelling, etc.

During the initial period of activity of the department it was the chief team in Bulgaria that produced software. The first Bulgarian software products were created at the department: compilers, operating systems, specialised software systems. The department also established the Minsk-32 Association of the Minsk-32 computer users in Bulgaria.

Besides research activities, the department has been carrying out a broad range of activities since the earliest years of its inception for the training of computer scientists and specialists in informatics.

Associates of the department were entirely responsible for computer science education at Sofia University until 1989 when part of the department staff formed a department bearing the same name at the university. Associates of the department teach at a number of other Bulgarian universities: Plovdiv University,Shoumen

University, the New Bulgarian University, the South-West University, Bourgas Free University. More than 100 different courses have been taught, many of them in a systematic way for many years.

Besides the thousands of trained students, associates of the department have been scientific advisors to about 300 graduate students and 40 PhD students from Bulgaria and abroad. The department is a specific kind of a school for specialisation and training of professionals and structural units in the field.

The number of staff working at the department has varied between 8 and 48 at different times and is now 18. About 100 former associates of the department now at many academic and industrial organisations in Bulgaria and also in many other countries.

The department was also part of the foundation for high school education in informatics in Bulgaria. Associates of the department produced the first educational programmes and textbooks.

The department maintains relations and works on joint projects with many academic organisations from France, Germany, Italy, USA, Russia and other countries.

The department has organised a number of international and national conferences and workshops among which is the series of conferences Information Technologies and Programming.

Tenth Anniversary of **IJ ITA**

International Journal
"Information Theories and Applications"
ISSN 1310-0513 www.foibg.com e-mail: foi@nlcv.net

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**) is the successor of the scientific co-operation organised within 1986-1992 by international workgroups (IWG) researching the problems of data bases and artificial intelligence. As a result of tight relation between these problems in 1990 in Budapest appeared the scientific group of Data Base Intellectualisation **(IWGDBI**) integrating the possibilities of databases with the creative process support tools. Heads of the IWGDBI are **R.Kirkova** (Bulgaria) and **V.Gladun** (Ukraine).

IJ ITA has been established in 1993 as independent scientific media for publishing original and non-standard ideas. For ten years IJ ITA became as well-known international journal. Till now, more than 200 papers from more than 400 authors have been published in nine volumes, which are separated in 61 numbers. IJ ITA authors are widespread in 26 countries all over the world: *Bulgaria, Canada, Czech Republic, Egypt, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Japan, Lithuania, Netherlands, Poland, Portugal, Romania, Russia, Scotland, Senegal, Spain, Sultanate of Oman, Turkey, UK, Ukraine, USA.*

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

Founder and Editor in chief of IJ ITA is **Krassimir Markov**.

During the years main co-editors of IJ ITA have been *R.Kirkova, V.Gladun, P.Barnev, Kr.Ivanova*. The IJ ITA Editorial Board also includes the members of the IJ ITA International Conferences Program Committees.

IJ ITA Publisher is **FOI-COMMERCE Co**., Sofia, Bulgaria.

IJ ITA official language is English.

Subscription for one year:
- for libraries and organisations: EURO 60.
- individual subscription: EURO 20.

Papers accepted by the editorial board of the IJ ITA are published in the following order and publishing fees:
- invited papers - free of charge;
- papers submitted by IJ ITA individual subscribers - EURO 3 per page A4;
- papers submitted from other sources - EURO 7 per page A4.

International Journal on INFORMATION THEORIES & APPLICATIONS
ISSN 1310-0513
Eduted by Institute for Information Theories and Applications "FOI ITHEA"
Editor in chief: Krassimir Markov
Publisher: FOI-COMMERCE: Sofia, 1090, P.O.B. 775, Bulgaria; e-mail: foi@nlcv.net
® "Information Theories and Applications" is a trademark of Krassimir Markov

International Prize "ITHEA"

International Prize "ITHEA" is aimed to mark achievements in the field of the information theories and applications.

Prize "ITHEA" is established by FOI Institute of 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 till now:

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. Vashchenko
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