Danail Dochev Istvan Simonics

Radoslav Pavlov (Eds.)

Generic Issues of Knowledge Technologies

HUBUSKA Second Open Workshop
Budapest, Hungary, 14 September 2005

Proceedings

HUBUSKA
Networking Centres of
High Quality Research on
Knowledge Technologies
and Applications

MTA SZTAKI
IMI BAS, IIT BAS
NOVITECH
UNIVERSITY
KLAGENFURT

Danail Dochev István Simonics
Radoslav Pavlov (Eds.)

Generic Issues of Knowledge Technologies

HUBUSKA Second Open Workshop
Budapest, Hungary, 14 September, 2005

Proceedings

The workshop presents the results of the project
INCO-CT-2003-003401 HUBUSKA
Specific Support Action under EU FP6 programme

Institute of Information Technologies – BAS

2005

Workshop programme and organising committee

Ottó Hutter (*Co-Chairman*), MTA SZTAKI eLearning Department, Hungary
László Böszörményi (*Co-Chairman*), Institute for Information Technology (ITEC), University Klagenfurt, Austria
Danail Dochev *(Editor),* Institute of Information Technologies at the Bulgarian Academy of Sciences – IIT BAS, Bulgaria
István Simonics *(Editor),* MTA SZTAKI eLearning Department, Hungary
Radoslav Pavlov *(Editor)*, Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences – IMI BAS, Bulgaria

ISBN 954-91700-2-0

PREFACE

The present volume contains contributions, presented to the HUBUSKA Second Open Workshop "Generic Issues of Knowledge Technologies", held in Budapest, Hungary on 14 September, 2005. The workshop is organised by the FP6 SSA international project INCO-CT-2003-003401 HUBUSKA. The papers discuss mainly some studies and results, developed by the project partners.

The first three contributions are dedicated to problems of Web Technologies – Web Mining and Semantic Web issues. The paper of Hr. Daskalova and T. Atanasova discusses some research tendencies in the field of the Semantic Web services. The paper of A. Benczur et al presents an architecture for Web Mining of large-scale Internet sites and result of experiments with its implementation. The paper of K. Staykova regards possible benefits from linguistically motivated presentation of semantic relations for the needs of ontology construction.

The paper of A. Velikov presents a review of modern approaches to information retrieval, including attempts to consider semantic relations in the process.

The next contributions deal with different problems in the design and development of distributed systems for specific problem domains. The paper of Cs. Domokos et al. describes a metrics for client behaviour prediction in distributed multimedia systems. The paper of D. Paneva, L. Pavlova-Draganova and L. Draganov discusses applications of digital libraries for presentation and preservation of cultural heritage. The paper of Sv. Braynov and R. Pavlov discusses problems of trust in electronic markets.

The last two contributions in the volume are dedicated to eLearning field and stand a bit further from the approaches of knowledge technologies. The paper of J. Bandakova et al. describes a distributed eLearning system for acquisition of specific programming skills. The paper of T. Urbanova and D. Siskovicova compares the practical problems of use of commercial Learning Management Systems /LMS/ versus Open Source LMS.

The editors hope that the presentations and the discussions on the Workshop will contribute to the multi-facetted picture of the modern trends and applications of generic knowledge technologies.

Sofia, 03 November 2005 D. Dochev, I. Simonics, R. Pavlov

Table of Contents

SEMANTIC WEB SERVICES – WHERE WE ARE AND WHERE WE ARE GOING

Hristina Daskalova, Tatiana Atanasova
Institute of Information Technologies - BAS
Acad. G. Bonchev 2, Sofia 1113, Bulgaria
daskalovahg@abv.bg
atanasova@iinf.bas.bg

Abstract

The paper tries to represent some research tendencies and priorities in the area of the semantic web services that follows the development of the Semantic Web. Semantics has been recognized as the key to next generation of more powerful information systems for better search, integration, question/answering as well as analysis/discovery. The basic technological components of the SWS are briefly considered. The obstructions and tendencies for SWS are discussed. The requirements for new tools for ensuring of SWS life-cycle are determined.

Keywords

Semantic Web, Semantic Web Services, Ontologies, SWS life-cycle.

1. INTRODUCTION

The idea of the global information net on the base of World Wide Web arisen in the 80th of 20 century has been realized as modern Internet and its scales are not possible to be measured or limited now. The expanding and development of the global net is growing with indescribable rate; new and various forms are added as inexhaustible information that concerns to all aspects and areas of human life. Intended for the practical use Web-resources step-by-step came out of observation and their processing and investigation became complicated and even insoluble task.

2. WHAT IS THE SEMANTIC WEB AND WHAT IS ITS INFLUENCE ON THE SEMANTIC WEB SERVICES

The difficulties arising during searching of relevant information in the Internet are determined mainly from the circumstance that web-technologies are oriented to human information service following human's point of view and logic. Computer gets a static role of arithmetic logic device which selects the answer on the base of formal indications depending on the user request. The approach of computer search does not allow to receive the answer to such questions as how priorities to be arranged during the resource searching in the net; how the selection of indexed objects to be done; on which criteria the final selection of the collected materials to be made and transferred to the user. There is a need to develop new methods and tools for dialog between the user and the computer according to the presentation of information materials and resources in a format appropriate to machine-semantic analysis. The idea of Semantic Web as extension of the existing Net has appeared aiming at presenting of the whole information in its sense and essentially.

The conception of Berners-Lee and his colleagues from the *World Wide Web Consortium* (W3C) is quite simple but revolutionary: it directs to making steps to intellectualising of web-systems and creating of universal language structure. This conception is now been realizing in a number of projects where leading researchers are working in special settled groups within W3C.

Ongoing research in Semantic web service engineering can be seen in a number of European projects: DIP, SECT, Knowledge Web, SeCSE, ASG, Sodium, Infrawebs, WS2 and many others.

The complementing of web-pages descriptions with computer-oriented data by which the computers can define key terms and inference rules would make the Net to be Semantic.

Requirements for knowledge presentation languages in the Semantic Web are:

- Universal possibilities for visualization;
- Syntactical interoperability (the programs can read the data and receive idea how to work. The syntactical interoperability is sufficient condition for possibility to construct the syntactical analyser and application programs interface – API, needed to manipulate with the data);
- Semantic interoperability (the main requirement for interchange format is the data to be understandable and the terms used in the data to be correlated. This is necessary to make an analysis of the content).

2.1 Technological components of the Semantic Web

The Semantic Web should be based on a series of standards. These standards have to be organised into a certain structure that is an expression of their interrelationships. A suitable structure is a hierarchical and layered one. It may be represented as a pyramid of web standards – from SOAP and XML/RDF to OWL-S and WSMO.

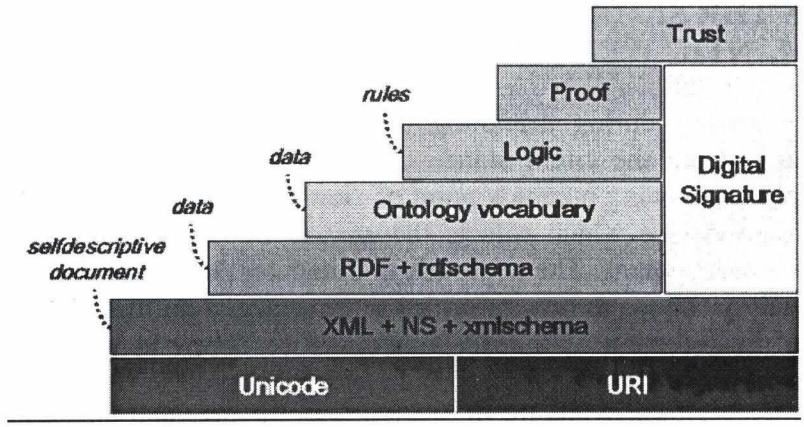


Fig. 2.1 The layers of the Semantic Web [Berners-Lee, 2001]

XML – Extensible Markup Language - is a specification that allows to define syntaxes and structure of the documents (they can have arbitrary structure). It replies to the requirement

for universality of visualization, for syntactical interoperability but can not ensure the semantic interoperability.

OWL – Web Ontology Language - is a specification that is designed for determination of terms and their relations. It includes the description of classes, properties and their instances. OWL is a language for defining and representation of Web-ontologies, the formal semantic of OWL describes how to receive logic consequences that not presented in ontologies literally, but follow from its semantically. These consequences can be based on one document or on a set of distributed documents which are combined through certain mechanisms in OWL. The ontologies are used for maintenance of automatic data interchange and for program integration. Mechanisms for searching use ontologies also for selecting pages with different syntaxes but with the same semantic.

WSMO – Web Services Modelling Ontology - is a specification of new environment for remote work and electronic commerce, presented by a group of European researches of related committee of W3C. This is an ontology used for semantic web services description. Development of WSMO is performed by working group from DERI (Digital Enterprise Research Institute). The project also consists of language specification WSML and Web Services Modelling Execution Environment (WSMX).

RDF – Resource Description Framework – is a specification for resource description that ensures the model for coding of instances, defined in ontologies. RDF determines and uses metadata which describes resources in the Web. It connects XML documents and high level tools for searching and navigation on the base of logic inference.

In general Semantic Technologies cover taxonomies, categorization, classification that should provide significant movement for Content, Document, Information Management, and other "intellectual assets".

Ontology is a headstone of the Semantic Web and may be become a standard technology for marking up electronic resources.

3. SEMANTIC WEB SERVICES

By definition a **Web service** is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Current Web service technologies describe the syntactical aspects of a Web service providing only a set of rigid services that cannot be adapted to a changing environment without human intervention. For the need for ensuring good cooperation in heterogeneous environment some additional information has to be added to the service description. Semantic is seen as a possible solution to resolve the problems of heterogeneity.

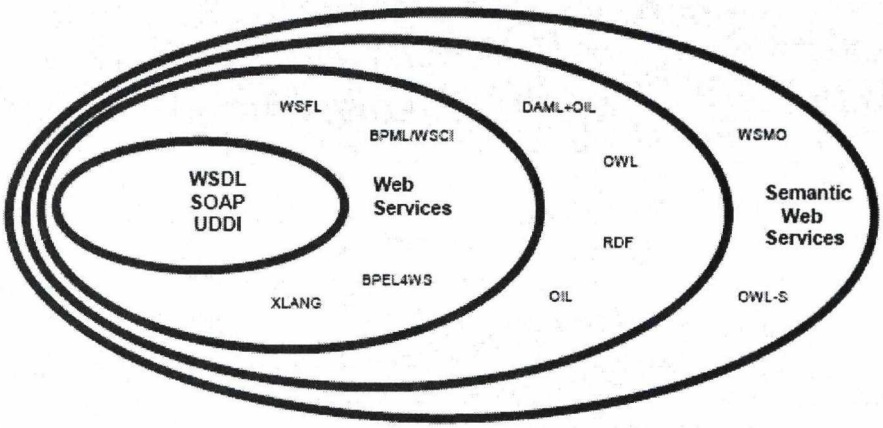


Fig.3.1. Semantic Web Services Technology [Arroyo, et al, 2004]

As the Semantic Web is extension of general WWW-net so the Semantic Web Services (SWS) extend the idea of traditional Web-services. Now programs can search ports and registries; for example UDDI-servers give accessible web-services. Apart from the fact that the program can find the needed web-service without user participation it does not know how to use the service and for what it is designed. The language for web-services description – WSDL – determines the interoperation between services while semantic realization gives information what the service does and what it proposes; it has to be sufficiently informative.

The requirements to the semantic Web-services description language are:

- Possibility to make discovery – SWS should give the description of its properties on the way that programs can recognize its purposes;
- Invocation – the program should initiate the needed services to fulfill given task. If the interoperation with other services is necessary then SWS should provide full list of activities to the program to involve and execute given service. The input and output data should be described too.
- Composition – using of several services by combining and selecting needed services to fulfill the goal. The services can interoperate seamlessly in such a way that the result of their combination can represent the solution to the given task. The software agents realize new services on the base of already existing services.
- Monitoring – software agent determines characteristics of the given service and verifies its execution.

When agents will have possibilities to discover, invoke, compose and monitor the service execution then the new environment can be realized which will give new possibilities to functional applications for the services exactly matched to the user requirements. The Semantic Web uses semantically described resources with static and dynamic content. By this way the SWS are undivided part of reformed Internet.

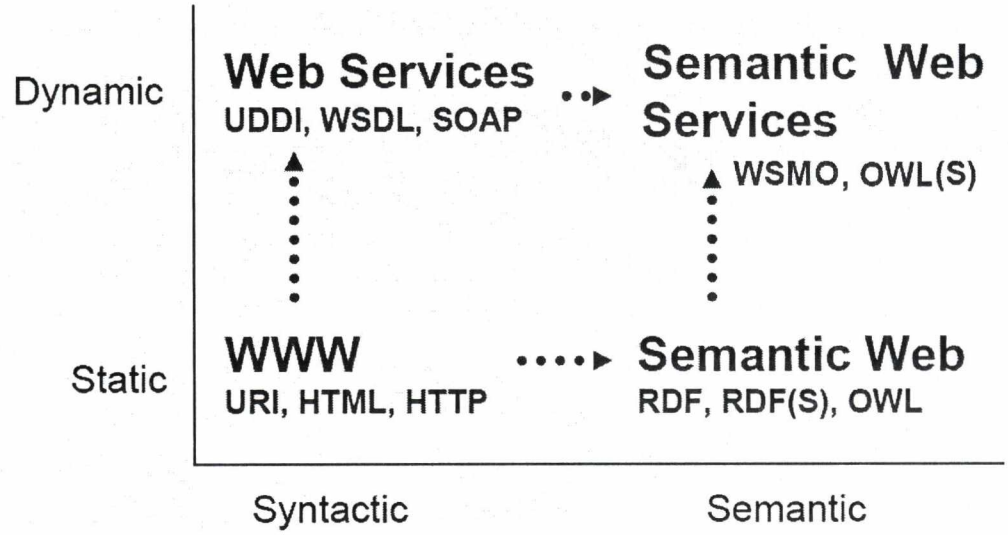


Fig. 3.2 Genesis of the Semantic Web Services [Bussler, et al, 2004]

Richer models need to be identified in order to characterize Web Services in terms of their exported dynamic behavior. Such richer semantics opens new challenging research directions and makes existing ones more overwhelming. New forms of Web Service composition and more complex forms of dataflow within Web Services need to be developed. This is connected closely with the discovery phase. An advanced Web Service registry has to provide support for semantic discovery by considering user specification involving Web Service capabilities and behavior.

The Web Service composition is an application domain where automated synthesis on the base of artificial intelligent methods may become successful.

Another problem concerned with SWS is retrieving of Web services with high Quality of Services (QoS).

A framework for providing high QoS Semantic Web services using intelligent methodology has to be developed. Since different application domains have different requirements for QoS, it is impractical to use classical mathematical modeling methods to evaluate the QoS of Semantic Web Services. Applying soft computing methods will allow handling fuzzy, uncertain and inconsistent QoS metrics effectively. The intelligent inference engine should be the core of the soft SWS agent that processes such fuzzy metrics.

The potential benefits in Service Reconfiguration / Semantic Adaptation will led to the establishment of an important class of research activities, both in industry and academia, aimed at the practical deployment of semantically rich service and process descriptions and their use across the Web service lifecycle:

- Architectures and Supporting tools for SWS Deployment;
- Applications of SWS to E-business and E-government;
- Supporting Enterprise Application Integration with SWS;
- Advertising, Discovery, Matchmaking;

- Ontologies and Languages for Service Description and Process Modeling;
- Foundations of Reasoning about Services and/or Processes;
- Composition of Semantic Web Services;
- Execution and Management of Semantic Web Services;
- Monitoring and Recovery for Semantic Web Services.

Efforts for the standardizing of the SWS follow the development of the Semantic Web and unfortunately SWS are on the more early stage of the development than the Semantic Web. Some specifications for designing and using of intelligent SWS have been proposed. As a base of Semantic Web the RDF is staying and there are some products based on RDF and RDF Scheme. There are real working servers as MusicBrainz (http://www.musicbrainz.org) which offer SWS-oriented interfaces for application programming (API) on the base of RDF.

The other line of investigations is concentrated on the systems for modeling of web services, for example WSMF that is partly based on the WSFL (Web Service Flow Language) of IBM. The researches are directed to practical using of SWS and making SWS public. But Web-services that use WSDL should support additional semantic languages for their interfaces. Combination of developments in the areas of Semantic Web and Web Services (fig.3.2) should make the information understandable for software agents and should accomplish the technologies for SWS.

4. OBSTRUCTIONS AND TENDENCIES

There are some reasons that delay the development of SWS technologies, for example the limited practical experience and not finished and approved yet standards and techniques. Beside that during transition to the SWS architecture some risks are expected and there are no enough arguments for investments. But by estimations of IDC it is assumed the growing of program products and services related to the Web and SWS up to 2008. At the moment the development concerns mainly integration on the "application-application" level, but from strategic point of view some organizations already consider integrated applications on higher level (composition). In this case using of SWS allows making adaptation dynamically for changing business processes. But there are not enough tools, technologies and standards developed yet.

Basic tendencies in the area of SWS may be summarized as:

- Complexity and concurrency between standards that are expected to be unified under the market pressure;
- The SWS development environment is changing – there are positive practical experience and knowledge that are necessary for introducing of new standards, but there are a lot of requirements and situations with various realizations leading to problems with compatibility. By these reasons the organization for interoperation between web-services is established – WS-I (Web Services Interoperability Organization). It is expected that approaches and architectures for software environment design based on Web and SWS that will ensure more simple operation

with business processes (e.g. REST – Representational State Transfer) will be developed;

- New market for Web and SWS products is expected to be consolidated (e.g. involving Trulogica and TalkingBlocks into HP corporation);
- Agreements between Sun and Microsoft are expected that will help resolving compatibility problems in the area of identification and control of business processes;
- New business models are arising that are realized by interactive portals as Google, Amazon and eBay which offer Web services that are not rigid but constructed according the user request.

The main difficulty for realization of the Semantic Web and following it SWS remains the necessity to provide all the information about Web content with semantic mark-up describing their meaning. The providers have to use RDF, OWL, OWL-S, WSMO and others new tools while creating their services.

It is expected that producers of web pages authors' development tools as Adobe, Macromedia and Microsoft will add into their products Semantic Web technologies. With full realization of the Semantic Web and SWS it seems to be possible to cardinal change the human life and to open totally new horizons.

5. CONCLUSION

The full semantic usage of Web services has many obstructions. At present the establishment of semantic based software components is reserved to specialists and programming experts. There is a lack of semantic based systems that use decision supported techniques, which allow automate process of Semantic web service designing, composing and maintaining.

Currently, the service oriented software development paradigm is mostly used for internal integration and not so many services are exposed to the outside world. This is partly caused by the fact that Semantic web service technology is still difficult to use, and there is a lack of trust in using someone else's services.

New tools and research are needed, that will deal with

- the requirements for creation of taxonomies and ontologies with high expressive representations;
- automatic taxonomy generation;
- semantic annotation of data in heterogeneous formats as well as machine generated data;
- semantic search;
- semantic integration of heterogeneous data;
- wrapping of legacy systems with semantically annotated Web Services;
- development of semantic web processes;
- more expressive representation that follow the ideas of intelligent technologies;
- probabilistic reasoning with imprecise probabilities;

- extension to the ontology languages that allows probabilistic and fuzzy reasoning;
- intelligent techniques in classifying web services automatically;
- improving of deduction capability of search engines — the capability to synthesize an answer to a query from distributed bodies of information in the knowledge bases;
- developing intelligent computing techniques that address the problem of multi-criteria decision making dealing with subjective and imprecise data;
- developing of measures for semantic services similarity;
- automatic composition synthesis of SWS.

ACKNOWLEDGEMENT

This work is carried out under EU Project INFRAWEBS-IST FP62003/IST/2.3.2.3 Research Project No. 511723.

REFERENCES

[1] Arroyo, S., Lara, R., Gomez, J. M., Bereka, D., Ding, Y., Fensel, D. 2004. "Semantic Aspects of Web Services". In: *Practical Handbook of Internet Computing*. Munindar P. (Ed.) Chapman Hall and CRC Press, Baton Rouge.

[2] Corcho, O., Fernández-López, M., Gómez-Pérez, A., and Lama, M. 2003. "An Environment for Development of Semantic Web Services". *Proc. IJCAI-2003 Workshop on Ontologies and Distributed Systems*, Acapulco, México, 13–20.

[3] Fensel, D., C. Bussler, Y. Ding, B. Omelayenko. 2002. "The Web Service Modelling framework WSMF". *Electronic Commerce Research and Application*, 1(2).

[4] Hendler, J. 2001. "Agents and the Semantic Web". *IEEE Intelligent Systems,* 16 2:30–37.

[5] Narayanan, S. and McIlraith, S. 2002. Simulation, Verification and Automated Composition of Web Services. *Proc. 11th Int. World Wide Web Conference WWW-2002*, Hawaii, USA, 77-88.

[6] Nern, H.-J., G. Agre, T. Atanasova, J. Saarela. 2004, "System Framework for Generating Open Development Platforms for Web-Service Applications Using Semantic Web Technologies, Distributed Decision Support Units and Multi-Agent-Systems - INFRAWEBS II. *WSEAS TRANS. on INFORMATION SCIENCE and APPLICATIONS*, 1, Vol. 1, 286-291.

[7] Bussler Christoph, Sinuhe Arroyo, Michael Stollberg, Matthew Moran, Michal Zaremba, John Domingue, Liliana Cabral and Jos de Bruijn, "The Web Service Modelling Ontology – WSMO", *Proceedings. Lecture Notes in Computer Science 3263 Springer* 2004, ISBN 3-540-23201-X, NetObjectDays 2004, 27-30-2004, Erfurt, Germany

[8] Berners-Lee T., J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, 2001, 279.

AN ARCHITECTURE FOR MINING MASSIVE WEB LOGS WITH EXPERIMENTS*

András A. Benczúr[12] Károly Csalogány[1] Kata Hum[1] András Lukács[12]
Balázs Rácz[31] Csaba Sidló[2] Máté Uher[1]

[1] Computer and Automation Research Institute, Hungarian Academy of Sciences
MTA SZTAKI, Kende u. 13-17., 1111 Budapest, Hungary
[2] Eötvös University, Pázmány P. stny. 1/c, 1117 Budapest, Hungary
[3] Budapest University of Technology and Economics, Egry J. u. 1., 1111 Budapest, Hungary
{benczur,cskaresz,khum,lukacs,bracz+origom,scsi,umate}@ilab.sztaki.hu
www.ilab.sztaki.hu/websearch

Abstract

We introduce an experimental web log mining architecture with advanced storage and data mining components. The aim of the system is to give a flexible base for web usage mining of large scale Internet sites. We present experiments over logs of the largest Hungarian Web portal [origo] (www.origo.hu) that among others provides online news and magazines, community pages, software downloads, free email as well as a search engine. The portal has a size of over 700,000 pages and receives 7,000,000 page hits on a typical workday, producing 35 GB of raw server logs that remains a size of 3.5 GB per day even after cleaning and preprocessing, thus overrunning the storage space capabilities of typical commercial systems.
As the results of our experiments over the [origo] site we present certain distributions related to the number of hits and sessions, some of which is somewhat surprising and different from an expected simple power law distribution. The problem of the too many and redundant frequent sequences of web log data is investigated. We describe our method to characterize user navigation patterns by comparing with choices of a memoryless Markov process. Finally we demonstrate the effectiveness of our clustering technique based on a mixture of singular value decomposition and refined heuristics.

Keywords

Data Mining, Knowledge Discovery, Web Server Log, Frequent Patterns, Clickstream Analysis, Clustering

1. INTRODUCTION

We observe an emerging demand of the telecommunication industry, the service and content providers to collect usage data and analyze it using data mining methods to answer questions regarding security, service improvement, marketing or business policy issues. Even medium sized companies can however easily produce log files of extreme sizes (up to hundreds of gigabytes per month). Collecting, keeping these log data sets in a storage-efficient and easily accessible way suitable for direct processing by several types of data mining algorithms is a challenging problem.

* Support from NKFP-2/0017/2002 project Data Riddle and OTKA and AKP grants

We present an architecture for web log mining designed for the largest Hungarian Web portal [origo] (www.origo.hu). The site [origo] that among others provides online news and magazines, community pages, software downloads, free email as well as a search engine. The portal has a size of over 700,000 pages and receives 7,000,000 page hits on a typical workday, producing 35 GB of raw server logs that remains a size of 4.5 GB per day even after cleansing and preprocessing, thus overrunning the storage space and analysis capabilities of typical commercial systems.

The system includes several modules integrated into a complete service for processing and analyzing web logs. After collection and filtering data may enter advanced analysis, optimized high density compression for long-term storage in a form appropriate for off-line data mining, as well as an OLAP-based statistical unit that provides fast on-line service for basic aggregation and detailed short-term queries.

To demonstrate the usability of our architecture for processing web logs we introduce the results of our first experiments over the [origo] site. The experiments include overall statistics, clickstream patterns as well as user behavior models. Distributions related to the number of hits and sessions are presented with somewhat surprising examples that different from an expected simple *power law* distribution.

For clickstream analysis first we apply *frequent sequence* mining where the problems of setting the minimum support and interpreting the output are investigated in detail. We leverage the usefulness of the *closed frequent itemset* concept and *association rules*. We also model navigational behavior by assuming that the user has a limited finite memory of past pages and the choice of the next hyperlink is more or less determined by the last few pages visited. We consider the support of frequent (strict referrer) sequences. We compare their support to the postfix shorter by one element, we measure the *bias* from a memoryless *Markov process*.

We demonstrate the use of spectral clustering for massive Web logs for forming groups both of users and of Web documents. Cluster qualities are measured in the following way. In case of document clusters we compared the results with the existing topics hierarchy of the site in question. A cluster of users, on the other hand, can be represented by the documents downloaded by the members of the cluster.

This paper is organized as follows: In Section 2 we describe the main properties and tools of the system for collecting, storing and analyzing web log data. Section 3 contains our experiments on the web log data collected from the [origo] site. Section 3.1 contains overall statistics about the number of hits and sessions. Our observations on click-stream analysis of web logs can be found in section 3.2 (finding frequent sets of downloaded pages) and in section 3.4 (Markov and non-Markov properties of the clicking process). In section 3.5 we describe how spectral clustering is used over the log data.

2. GOALS AND THE ARCHITECTURE

Handling very large volomes of data forms our primary design goal. The leading Hungarian portal, [origo] records over 6.5 millions of page hits per day, resulting in text file logs over 35 Gigabytes per day and over a Terabyte in a month.

In the current solution, front end server http logs are rotated over the night and stored gzip'ped on tapes. The compressed size of 140 GB of a one-month log makes even reading the logs back infeasible. Prior to our work the only means of analyzing weblogs for [origo] portal editors consisted of standalone fixed report generators and Perl scripts, with the drawbacks of limited flexibility and usability. These tools can answer only a limited number of questions by pre-generated reports, and customization typically results in rewriting the utility.

We design and build a data warehouse-like solution offering a flexible base for ad-hoc analysis, OLAP-like queries, data mining components and long term storage of log files. Notice that an OLAP architecture design for weblog mining is a challenging task; the statement given in 1998 [22] remains valid: "due to the important size and the ever exploding nature of web log files, the construction of multidimensional data cubes necessary for on-line analytical processing and knowledge discovery, is very demanding and time consuming".

The logical architecture of the system is depicted on Figure 1.

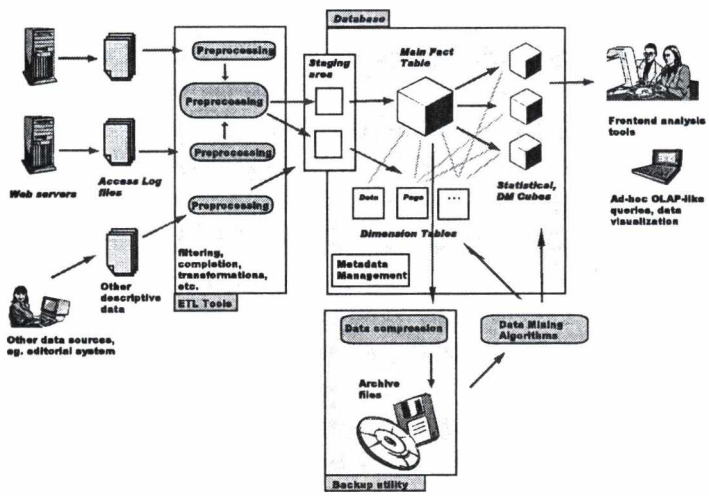


Fig. 1. The architecture of the web log analyzer

The main flow of the clickstream data is shown on the figure. The prime data sources are the web server log files. They contain all the requests the web servers get. A hit has standard attributes like client host, request time, request code, http request (containing the page url), referer, agent identification string and contains session identification cookie as well. Collecting cookies in log files simplifies the processing of session related information. The session identification with the use of the usual Common Log Format or the Extended Common Log Format, without cookie handling should be less accurate [4].

Additional data sources include the editorial system of the [origo] portal describing the web page hierarchy and the properties of the portal contents, like author, creation date and other descriptive attributes.

The first step is preprocessing the raw log data. We decided to handle clickstream data of the unique page-view hits as the finest data granularity, without the additional requests for parts of the pages, like for example requests for the embedded pictures. Thus we introduce an initial filtering stage that does as well cleaning of the raw data. In this step we also deal with the task of making data more understandable, completing missing information, doing reverse DNS lookups, identifying session-level attributes etc. After the first phase the data is ready for loading into the database, through a staging area.

The database is designed according to the traditional multidimensional star/snowflake schema [11,10]: page hit data is collected in a central fact table, and all the descriptive dimensional data are stored in dimension and dimension hierarchy tables. During the construction of the schema we have paid attention to the huge data volumes. It was necessary to evaluate the element counts and element count growths of the dimensions. Processing a sample log helped to determine the problematic dimensions, and to estimate the fact table size. The central fact table contains only a short time period (up to a few weeks or months) of data of the finest granularity.

To enhance the analytical performance of the system we build statistical data cubes that contain aggregated data from the main fact table. The main idea is to store high resolution data only for a limited time period. The statistical cubes contain data for a long time period, but with less dimensions or filtered data. The determination of the aggregation schemas is based on the needs of analysts and on the evaluation of the cube size and the cube data density. We tried to omit storing sparse sub-cubes and tried to find interesting and informative aggregates. We designed a metadata-structure for data propagation to the statistical cubes. Appropriate descriptive metadata on the cubes can be useful for developing a general OLAP-like analyser tools as well.

For the long period web log storage we use a novel compression module. The module is designed to store all kind of log files efficiently, compress the log data with a compression rate exceeding that of the general data compression utilities by using particular properties of the log file structure. The module receives a general, descriptive meta-structure for log data (field type definitions and other field properties) in order to compress the data by selecting the appropriate methods from a large set of compression methods, like gamma-code, delta-code, Huffman-code etc. As a final step the compressed data is archived. The compression utility gets processed data from the database, from the main click fact table, and can restore this data later.

The compression module has a number of desirable properties. Decompression runs in linear time and may begin by almost random access in the middle of the compressed file. Accidental errors of the compressed data remain local. As the compression is made separately on fields, and the decompression results a parsed memory image of the data set, the compressed data is directly ready for data mining tools.

Data mining algorithms such as association rule and sequential pattern mining or clustering can be implemented using the compressed, long range archive data as special, standalone applications. We implemented a general purpose pipelined data mining tool, which can use the compression utility as input, and can write back the result into the database. Even though mining log data inside the database is feasible [18], mining data outside the database seems nowdays more effective and flexible.

The clickstream database provides a good opportunity to store and analyze the results of these methods. The goal is to make data mining results accessible in the OLAP-like query framework, to make this knowledge visualizable, easy to use. Hence we design data mining cubes, data-mining dimensions or dimension hierarchies like frequent page sequences or user clusters. Using these new elements in data collection and in the analytical framework can make data mining results more understandable, the OLAP-like ad-hoc queries more flexible. An easy-to-use web-based frontend tool makes these statistical and data mining reports accessible for the users.

The general framework and the experiences through building the weblog architecture would give us a good base to design other log file processing modules as well, like mailserver logs or search engine logs.

3. EXPERIMENTS

3.1. Download and session statistics

For the sake of illustration we describe the October 2001 web logs that we have permission to publish. Notice that current figures became roughly ten times larger since then! The total number of unique IP addresses was 875,651, having a total of 18,795,106 requests which can be grouped into 5,702,643 sessions. (There should be at least 30 minutes between two sessions.) The site contains 77,133 unique pages which have been downloaded at least at once.

Pages over the [origo] site consist of a dynamically generated frame with fixed elements (links to other topics and services) as well as links to the hot news of the current and other topics. The URL of a particular content remains valid forever and with the current frame generated at request; links to old pages are, however, disappear after a while though they remain searchable in the archive.

In Figure 2 we show the distribution of page access count, with the count on the horizontal axis and the corresponding density on the vertical axis. The distribution is Zipf-like with a $\gamma = -1.2$ exponent. The dashed line is the fitted line using least squares method.

Figure 3 shows a similar graph for the number of pages visited by a user. The distribution appears to be a combination of two Zipfians, in the lower range with a larger exponent $\gamma = -1.5$, while in the higher range a smaller $\gamma = -2.5$.

Next we consider user session statistics. We define a new session to start after 30 minutes of idle time or more. In Figure 4 we show the distribution of session length with density on the vertical axis and a logarithmic scale on the horizontal axis.

The experiment shows that in the Origo web log, the average sesion length is about 7 minutes with 1919 seconds of deviation. The median session length is 0 second, which means that half of the session contains one request, and less than 50% of them contains more than one.

The distribution of time elapsed between adjacent hits by a user is seen in Figure 5. The graph consists of several peaks. The first one corresponds to redirects. The next peak at 14 seconds might correspond to real user intention: after seeing the page, quickly selecting the desired link. The peak at 64 seconds might then characterize users who read the page

content before continuing. Peaks at 1745 and 3486 seconds are explained by the refresh meta-tag of pages (`www.origo.hu/index.html` has a refresh meta tag of 1740 second). Another maxima appear at multiples of days that may describe users who read [origo] news (or perhaps have Internet access) at a regular time of the day. The average time gap between two request is about 1 day, and the median time gap is about 10 minutes. In other words half of the visits come more than 10 minutes after the last access, and half of them come within 10 minutes.

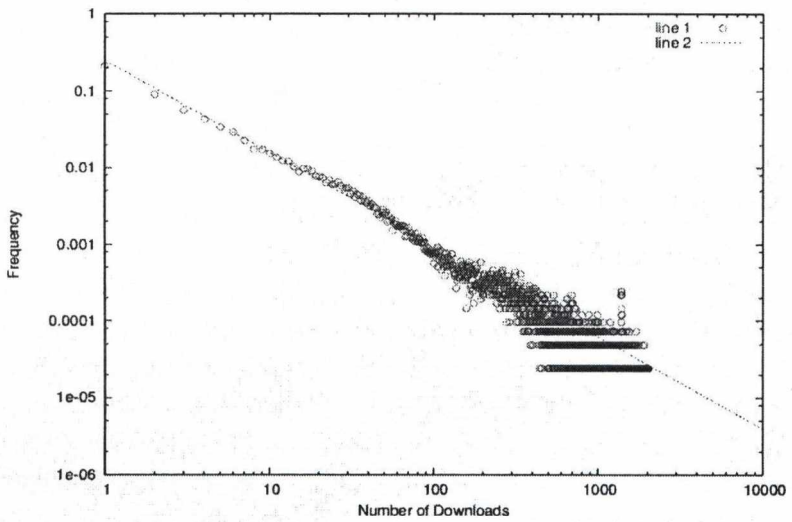


Fig. 2. Page Download with the count of pages on the horizontal axis and the corresponding density of pages on the vertical axis, both scales logarithmic.

3.2. Clickstream analysis

Click stream analysis has a vast literature; approaches falling into indentifying important sequences in the click pattern and modeling user behavior. A very early approach [6] utilizes association rule mining [2]; frequent sequences [13,19] and longest repeating subsequences [16] can be considered as variants of this approach.

Another approach models navigational behavior by assuming that the user has a limited finite memory of past pages and the choice of the next hyperlink is more or less determined by the last few pages visited. Various studies investigate how well length three [20] or longer [15] sequences predict user behavior, comparing patters from weblogs with Markovian processes of certain order or consider hidden Markov models [21].

First we describe our frequent sequence mining experiments including algorithmic issues and setting the minimum support. In section 3.3 we turn to the challenge of handling the huge size of frequent sequences that makes it impossible for a human to interpret the results. We briefly concern some useful space-saver specialization of the basic definition, as well as the

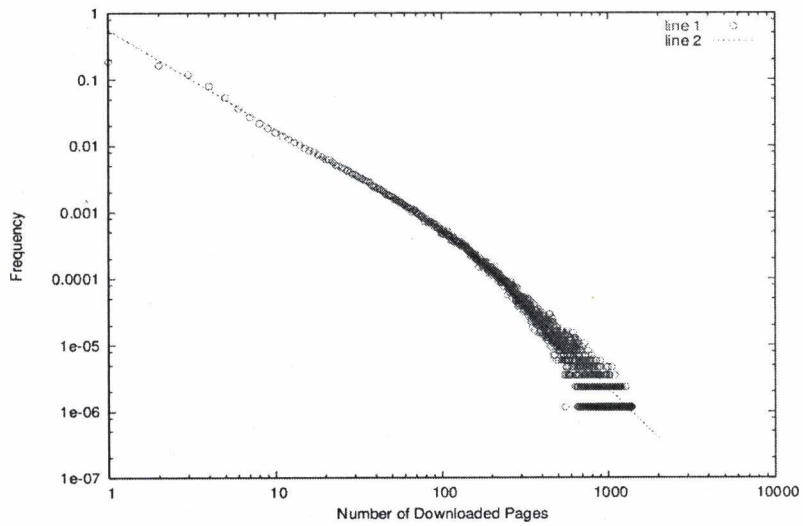


Fig. 3. User Download with the count of users on the horizontal axis and the corresponding density of users on the vertical axis, both scales logarithmic.

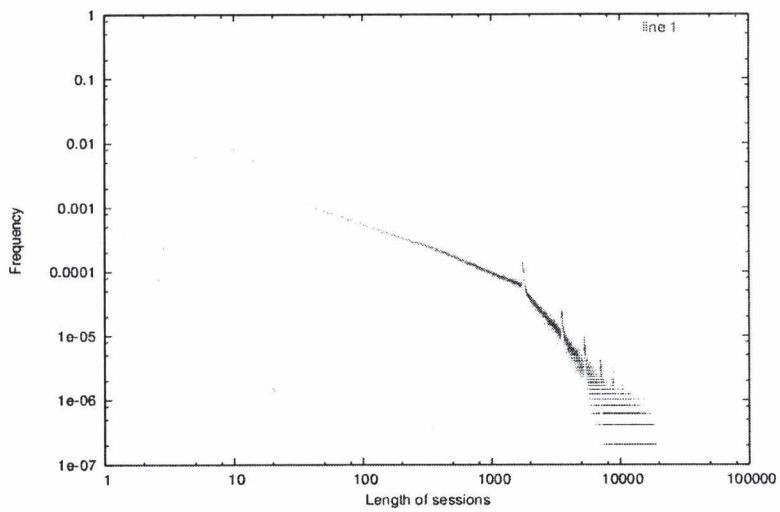


Fig. 4. Session length on the horizontal axis with density on the vertical axis, both scales logarithmic.

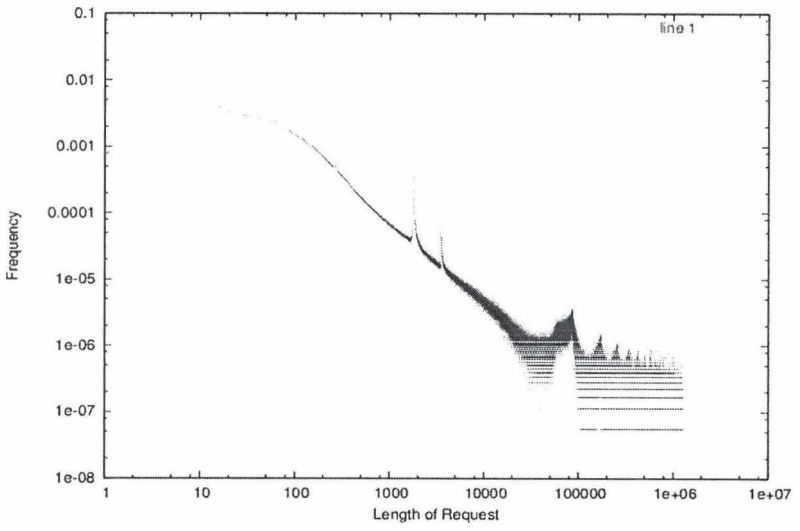


Fig. 5. Length of idle time between two requests on the horizontal axis, with a logarithmic scale density on the vertical axis.

usefulness of the 'closed frequent itemset' concept as well as those of association rules. Finally in Section 3.4 we sketch a novel approach to filter the relevance of longer sequences by measuring how much of their first element is "remembered" when the user makes the last selection.

While the output of frequent sequence mining gives meaningful insight of navigational patterns, it is typically used as the first step of *association rule discovery*. An association rule $X \rightarrow Y$ holds with *support* s and *confidence* c, or (c, s) in short, if s percentage of all the users visited all pages in set X, and c percentage of these users also visited all pages in set Y. Usually we set a minimum constraint for both parameters: min_supp and min_conf respectively to select only the most relevant rules of the dataset.

Challenge: infeasible Apriori data structures

We find that algorithm *Apriori* [2], a descendant of algorithm *AIS* [1] using ideas from [12], is infeasible for web log sequence mining with low minimum support and *heavy users* with a large number of page visits. We overcame this difficulty by designing a new internal data structure that holds considerable supplemental information aside the trie, and redesigning the counting process around this data structure. Due to space limitations we omit discussion and use the results of the modified algorithm.

The key problem with a textbook implementation of algorithm *Apriori* is that a combinatorial explosion is observed starting from the 5-element frequent sequences. Since on the average 50–60% of the candidates proved to be frequent sequences themselves, the explosion is not explained by the time spent on the counting of non-frequent candidate support. A

more detailed analysis showed that over 85% of the time was spent on 20% of the log file which belonged to the accesses of *heavy users* with a large number of page downloads.

The combinatorial explosion for medium-size sequences caused by the heavy users can be explained on a theoretical base. The counting process of Apriori, while examining the sequence of one user, for each element of the sequence uses as many trie descend-lookups as the number of perviously visited trie nodes. With short user sequences (like market baskets at the time of the development of Apriori) this poses no problems at all. However, web logs can hold user sequences of up to several thousands of accesses, making the counting process infeasible for web logs.

pass	count of		run time		
	candidates	frequents	(new)	(original)	(or., 80%)
1	14718	358	-	0:00:04	-
2	128164	4030	0:11:12	0:00:45	0:00:25
3	62452	13801	0:09:07	0:07:09	0:01:36
4	62324	27017	0:15:13	0:25:09	0:04:03
5	72050	37939	0:18:31	0:59:04	0:08:20
6	74273	42521	0:22:50	1:43:52	0:14:09
7	70972	43465	0:24:46	2:26:37	0:18:55
8	70541	45373	0:20:05	3:05:05	0:22:30
9	75082				0:24:57

Table 1. Comparison of encoding and decoding time on the sample web logs

The statistics of a sample run are summarized in Table 1. Run times were measured on a 1.4 GHz AMD Athlon based machine with 3.5 GB of RAM running linux. The minimum support was set according to the next subsection so that there are numerous nontrivial frequent documents.

Setting the minimum support

The key problem of mining a news site's web log is that the data set is extremely heterogenous. The most relevant difference is between the static pages (main page, directory indices) and the content (news, articles) themselves. *The support of a static page is higher by one to three orders of magnitude.* This can be explained by on one side the aging of the news and articles, on the other side the recurrence of the index pages in one user's clickstream.

This has a dramatic impact on the reasonable *min_supp* values. The main page gets 28% of the hits, but there are only 8 documents with over 1% of support. The 50 documents with over 0.1% support contain only 3 content page, everything else is either a directory index or some ad support page. On the other hand to get a fair coverage of the data set by frequent documents one has to set the minimum support to as low as 0.01%. This is quite far from the textbook or synthetic dataset definition of 'frequent'.

The impact of the minimum support on the number of frequent documents and the total hits collected by them can be seen on figure 6.

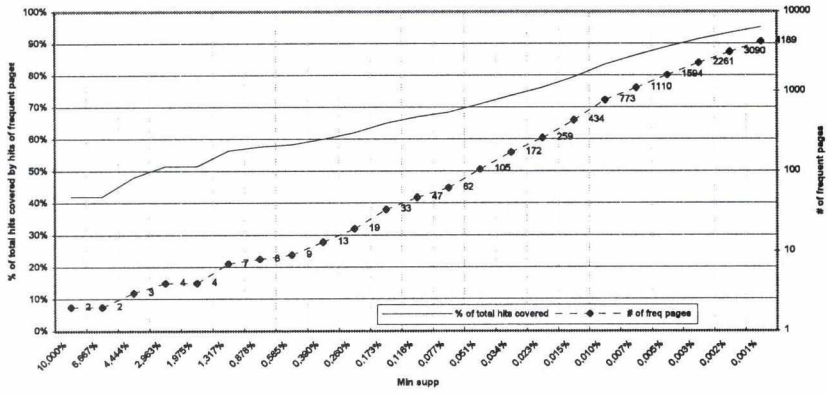


Fig. 6. Number of frequent pages for different *min_supp* values

3.3. Challenge: frequent sequences are too frequent

As seen in Figure 7, there are just *too many* frequent sequences for a human to read through in hope of understanding the data set. Also, the information in frequent sequences is highly redundant. Due to the low *min_supp* value, the output is flooded with all combination of the directory index pages. We have to either manually filter the output or look for better concepts than plain frequent sequences.

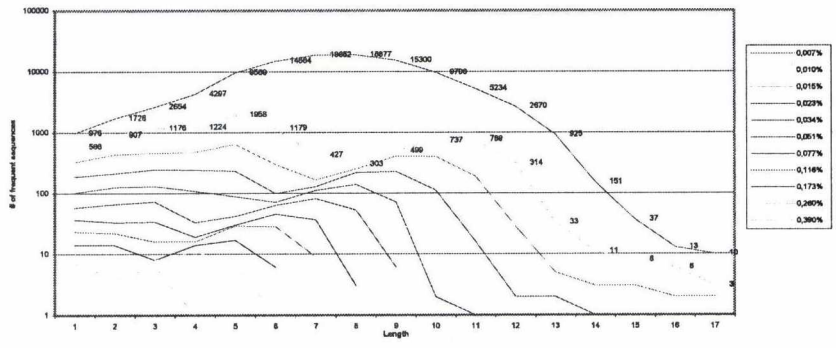


Fig. 7. Number of frequent sequences for different *min_supp* values

Note that the minimum support values are also logarithmic (with a quotient of 1.5), so the equidistant partition of column '1' shows the power law of the 1 element frequent sequences (e.g. frequent documents).

First we give ideas of alternative definitions that may reduce the number of frequent sequences. A novel relevance filtering method is descibed in the next subsection.

Closed frequent itemsets

A frequent itemset X is *closed* if when taking all users who downloaded all pages in X we cannot find any more pages all of them downloaded too. This is equivalent to the following:

we cannot find a $Y \supset X$ larger set which had the same support. A third equivalent definition is that a frequent itenset is closed if it does not stand on the left hand side of a 100% association rule. In theoretical studies this is a very important concept, as there are examples where the count of non-closed freqent itemsets is exponential in the count of closed frequent itemsets. However, in real world datasets *there are no 100% association rules valid*[4], so by the the third definition all frequent itemsets are closed.

Weakening closedness
We could weaken the requirement of the association rule being 100%. If we do this to the extreme and filter out all frequent itemsets which are on the left hand side of an association rule of any confidence, then we get the *maximal* frequent itemsets (or sequences). This approach can reduce the number of frequent sequences by 50-70%, but still cannot reach the human understandable size. On the other hand, this gives us an idea on how many association rules hold (without considering the possibility of a certain itemset to be on the left hand side of more association rules).

Referer-strict sequences
We get a more refined and weblog-specific definition if we require the user download equence to contain a frequent sequence with each element's referer being the previous element of the frequent sequence. PLease note that this way we still note require the frequent sequence to be a consecutive subsequence of the containing download sequence of the user. We allow 'browse loops', e.g. when the user turns into a dead end or looses his way through the site navigation and returns to a previously visited page and continues on the path of the frequent sequence.

3.4. Memory of navigation

Both Markovian and frequent sequence mining approaches suffer from time and storage limitations, as noted for the former approach in [17,15]. When mining frequent sequences we set the minimum support threshold and receive, as output, all sequences visited by at least the prescribed percent of users. By setting the threshold high, however, we get too little and obvious information, those related to visiting various index.html pages. And as we decrease the threshold we very quickly face an exceedingly large amount of output that needs data mining within the data mining output such as relevance measuring [5], clustering [8] or base selection [23].

We suggest a different combined approach. As suggested in [13], we consider the support of frequent (strict referrer) sequences. We compare their support to the postfix shorter by one element, thus measuring how much information is remembered of the first Web page when making the decision to visit the k-th. Formally stated, we measure the *bias* from a memoryless Markov (or order $k-1$ Markov) process,

$$r(x_1, x_2, \ldots, x_k) = \Pr(x_k|x_1, \ldots, x_{k-1}) - \Pr(x_k|x_2, \ldots, x_{k-1}) \cdot \Pr(x_2, \ldots, x_{k-1}|x_1) \; .$$

[4] ...usually, of course. Whenever there is the slightest applicability of an independence-based probabilistic model on the data source (for example where users actions are independent of each other) the probability of an itemset to be closed is exponentially small in the support of it. As we set a minimum constraint on the support, this is practically zero.

For example consider the frequent strict referrer sequence

```
www.origo.hu/szoftverbazis/jatekok/akcio/index.html
www.origo.hu/szoftverbazis/jatekok/akcio
www.origo.hu/szoftverbazis/jatekok/akcio
```

with value $r = 0.22$, meaning that visiting the action game index page makes it more likely to follow a second game page. The meaning of such a rule needs careful understanding. The user is very unlikely jump directly to a particular game page, unless using search. The likely entry point is thus the index—or another game page. Hence this rule in fact means that it is *less* likely that the user will visit a third game page. As another example,

```
www.origo.hu/search/index.html
www.origo.hu/search/index.html
www.origo.hu/index.html
```

has a negative bias -0.088, meaning that after subsequent searches we are less likely to go back to the index page and continue with reading news articles—we are likely to have a target in mind and are less intended for browsing the site. Pages

```
www.origo.hu/adsminisites/usa
```

consist of image sequences; all such sequences are frequent of length up to 12 and have positive bias that exponentially diminishes (-0.07, 0.07, 0.004, 0.013, 0.013, 0.011, 0.010, 0.0087, 0.0082, 0.0072, 0.0067). Note that the length three sequence has negative bias. The length two sequence has a positive bias for continuation with another world news article, these two observations together meaning that a large number of users quit after seeing two pictures, however those who continue are even more likely to continue as they see more and more. These observations, for the example of the above image sequence, explain a large tail in session length and may partly explain the Zipfian distribution observed in Section 3.1.

The method of selecting sequences with large (positive or negative) bias filters out most of the frequent sequences. For example we may safely discard rules such as "many users immediately quit reading world news and turn to local news",

```
www.origo.hu/index.html
www.origo.hu/nagyvilag (world news)
www.origo.hu/itthon (home news)
```

since the bias 0.0003 means a purely random bahavior that may equally likely happen if they have already seen a number of news, regardless of topical categories.

As indicated by Figure 8, the bias of frequent sequences decays exponentially as the length of sequences increases. Sequences with positive bias appear to have a larger tail, however the maximum in this case is attained at the above sequence of images and is characteristic to such sequences.

3.5. Clustering

We use clustering by singular value decomposition, a method well described in the literature when applied to graph partitioning [3,14,9] and latent semantic indexing over the document-word matrix [7].

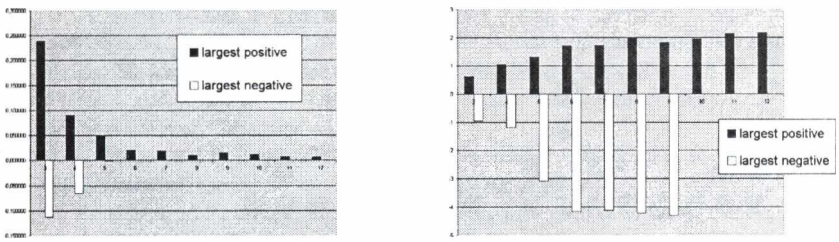


Fig. 8. Value of the largest positive and negative bias for a sequence of length 3 to 12, on a regular (left) and logarithmic (rigth) scale

We adapt the algorithm described in [14] for the document-user matrix instead of a graph adjacency matrix. The algorithm first computes the first k singular vectors and projects users or, respectively, the documents into a k-dimensional space. Over the reduced size projection we apply the k-means clustering algorithm. While singular vector computation is expensive, few vectors yield clusters of poor quality. Typically for k clusters at least k dimensions are suggested; however even projecting to higher dimensions we often obtain a single huge cluster.

Since k-means often fails to produce balanced size clusters, we use two heuristics to produce a "reasonable" split of the data. First, index pages with very large number of visits often keep points close and result in huge clusters; these elements can safely be discarded when found. Second, by starting with a certain dimension k, we continue computing increasing number of singular vectors until we reach the desired split in k-means.

In case of document-clustering we compared the results with the prefixes of the URLs (the domainname and the first tag of the path). A cluster is homogeneous if a few prefixes dominate the cluster. Figure 9 shows an increasing homogenity as we move down in the hierarchy. Each color in the pie charts corresponds to an URL prefix downloaded by the users in the cluster (e.g. software `www.origo.hu/szoftverbazis` is red, sport `www.origo.hu/sport` is green).

A cluster of users, on the other hand, can be represented by the documents downloaded by the members of the cluster. These documents shows the interest of the set of users. On the bottom of Figure 9 we see the homogenity increasing down in the hierarchy, similar to document clusters. Notice however that clusters of users are less homogeneous, since a typical user is interested in several topics and is easily distracted—nevertheless the prime goal of a portal site design is to make users spend more time there.

ACKNOWLEDGMENT

Our thanks to the team at Axelero Inc., especially to István Szakadát, Gábor Kiss and Tibor Takács.

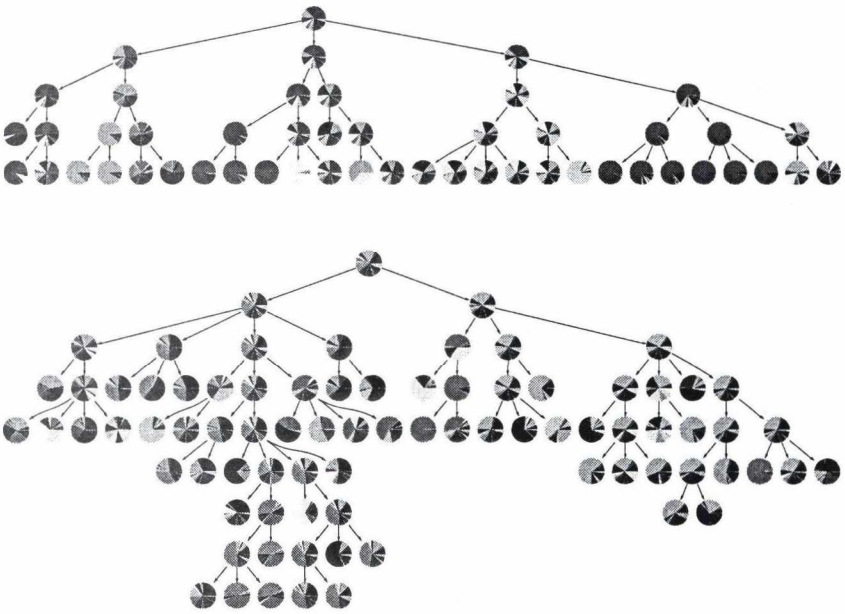


Fig. 9. The top of the hierarchy of the document (top) and user (bottom) clusters. Each color means different prefix in the URLs of the given cluster, and downloaded by the users in the cluster, respectively

References

1. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.
2. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.
3. C. Alpert and S. Yao. Spectral partitioning: The more eigenvectors, the better, 1994.
4. J. Andersen, A. Giversen, A. H. Jensen, R. S. Larsen, T. B. Pedersen, and J. Skyt. Analyzing clickstreams using subsessions. In *Proceedings of the third ACM international workshop on Data warehousing and OLAP*, pages 25–32. ACM Press, 2000.
5. S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: generalizing association rules to correlations. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press, 1997.
6. M. S. Chen, J. S. Park, and P. S. Yu. Data mining for path traversal patterns in a web environment. In *Sixteenth International Conference on Distributed Computing Systems*, pages 385–392, 1996.
7. S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
8. E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering based on association rule hypergraphs. In *Research Issues on Data Mining and Knowledge Discovery*, 1997.
9. R. Kannan, S. Vempala, and A. Vetta. On clusterings — good, bad and spectral. In *IEEE:2000:ASF*, pages 367–377, 2000.
10. R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, Inc., 2002.
11. M. Levene and G. Loizou. Why is the snowflake schema a good data warehouse design? *Inf. Syst.*, 28(3):225–240, 2003.
12. H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. In U. M. Fayyad and R. Uthurusamy, editors, *AAAI Workshop on Knowledge Discovery in Databases(KDD-94)*, pages 181–192, Seattle, Washington, 1994. AAAI Press.

13. B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Using sequential and non-sequential patterns in predictive web usage mining tasks. In *2002 IEEE International Conference on Data Mining (ICDM'02)*, 2002.
14. A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm, 2001.
15. P. Pirolli and J. E. Pitkow. Distributions of surfers' paths through the world wide web: Empirical characterizations. *World Wide Web*, 2(1-2):29–45, 1999.
16. J. E. Pitkow and P. Pirolli. Mining longest repeating subsequences to predict world wide web surfing. In *USENIX Symposium on Internet Technologies and Systems*, 1999.
17. A. Schechter, M. Krishnan, and M. Smith. Using path profiles to predict http requests. In *Proceedings of the 7th World Wide Web Conference, Brisbane, Australia*, 1998.
18. C. I. Sidló and A. Lukács. Shaping sql-based frequent pattern mining algorithms. In *KDID'05: Proceedings of the fourth International Workshop on Knowledge Discovery in Inductive Databases*. Springer-Verlag, 2005.
19. M. Spiliopoulou. The laborious way from data mining to Web log mining. *International Journal of Computer Systems Science and Engineering*, 14(2):113–125, 1999.
20. X. Sun, Z. Chen, W. Liu, and W.-Y. Ma. Intention modeling for web navigation. In *Proceedings of the 11th World Wide Web Conference (WWW)*, 2002.
21. A. Ypma and T. Heskes. Categorization of web pages and user clustering with mixtures of hidden markov models, 2002.
22. O. R. Zaïane, M. Xin, and J. Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Advances in Digital Libraries*, pages 19–29, 1998.
23. M. Zaki and M. Ogihara. Theoretical foundations of association rules, 1998.

LINGUISTICALLY MOTIVATED KNOWLEDGE REPRESENTATION FOR MODERN SEMANTIC WEB TECHNOLOGIES

Kamenka Staykova
Institute of Information Technologies – Bulgarian Academy of Sciences
Sofia 1113, Acad. G. Bonchev str., bl.2
staykova@iinf.bas.bg

Abstract
This paper presents an attempt a particular semantic web technology Magpie to be used with knowledge representation supported linguistically, instead of ontological construction usual for these kind of applications. The theoretical and technical challenges which make difficult the usage of linguistically motivated ontologies with modern semantic web technologies are discussed.

Keywords
Knowledge Representation, Ontologies, Semantic Web, Natural Language Processing

1. INTRODUCTION

Knowledge representation has been a challenging field since the time of birth of the ideas, formal languages and programming techniques called Artificial Intelligence. The computer science shows a remarkable development since the first attempts to present 'common sense' knowledge formally and to operate with it by such a languages as LISP and PROLOG. The big difference comparing 60-ties with today is the realized dream of freedom in information exchange which is achieved having Internet and Intranets. The technologies of information publishing, access, retrieval and usage place the user within a vast and deep data flow. The problem of formal presentation of data, information and knowledge is still a challenge of the present day. People are lost in too much irrelevant information, accessible in human-understandable form and the knowledge management techniques need to take in account the human beings' semantics in their algorithms. Information is not in proper form to be used by 'machines' (software agents) because it relies on human beings implicit associations, it is often fuzzy, contradictive and not complete.

Researchers from different areas of information management agreed that a basic construction behind the knowledge in particular field should exist, because of sharing and reusability needs. The definition "Explicit presentation of a shared conceptualization" is widely accepted following the ideas in [7]. Such formal constructions are given the name 'ontologies' after the philosophical notion 'ontology' and are the backbone of emerging Semantic Web.

It is clear now that the different fields of human knowledge will be presented by different ontologies; for example, the ontologies of mechanics, financial ontologies, ontologies of

medicine and so on. There are different paradigms of building ontologies, even in one and the same scientific field, for example Natural Language Processing[1]. Even within the same task of a modern web technology, for example E-Learning task, the need of "multiple ontological levels, frames and views" is recognized[2].

This paper presents one promising semantic web technology called Magpie (section 2) and discuses the possible usage of such a technology with different ontologies showing the challenges reached in the attempt to use linguistically motivated construction of knowledge (section 3). Finally some conclusions are offered (section 4).

2. MAGPIE - WHAT KIND OF TECHNOLOGY IS IT?

"Magpie is a suite of tools supporting a 'zero-cost' approach to semantic web browsing. It avoids the need for manual annotation by automatically associating an ontology-based semantic layer to web resources."[6] The authors of Magpie characterize it from different perspectives- as an interpreter of web pages, as a semantic web browser, as a framework for developing semantic web applications.

The most important feature of Magpie as *a web pages interpreter* is the availability of different viewpoints to web pages, each viewpoint depends on the particular ontology chosen by the user as her/his eyelet. Usually ontologies are too large constructions of knowledge and the user's eyelet is, in fact, a particular part of the chosen ontology, accessed in its form of 'ontological lexicon'. Lexicon comprises several classes of the chosen ontology and a number of instances, which represent items the Magpie extension is able to work with and find in the web page. This way Magpie semantically marks-up web documents on-the-fly.

"Another way to look at Magpie is as a *semantic web browser.* If we take this view, then Magpie can be seen as providing an efficient way to integrate semantic and 'standard' (i.e., non-semantic) web browsing, through the automatic association of semantics to web pages and the provision of user-interface support."[6] Magpie supports automatically updated semantic log and using browsing history management it enables collaborative Semantic Web browsing [4].

Most recent development of Magpie (as a framework supporting e-Learning) is focused on the dynamic semantic web applications- semantic web services. Magpie supports two main types of services: the first type are services, which depend on the particular ontology used by the user at the moment, the second group of services demonstrate the trigger functionality of Magpie.

It is easy to install Magpie from http://www.kmi.open.ac.uk/projects/magpie and to try it by one's own. Let's have an example with Magpie and a sample lexicon, which forms the

---

[1] John Bateman's page http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/info-pages/ontology/ontology-root.htm offers an useful ontologies classification.

[2] Prototype of advanced learning platform http://kmi.open.ac.uk/projects/kweb/resources/D3.3.3.pdf, page 7.

Magpie user's eyelet. The chosen ontology is the Portal Ontology of AKT project available in several formats and announced as "the main ontology, describing people, projects, publications, geographical data, etc."[3] The lexicon contains records of four classes with a number of instances and enables Magpie to operate with this ontologically supported information. ***Fig. 1*** below shows a snapshot of semantically enriched browser (Internet Explorer) with the Magpie buttons corresponding to the four classes: **SW technologies, SW activities, Other SW topics, Community**.

Two functionalities of the buttons are available at this level and they are offered via a pull-down menu visible on ***Fig. 1***. The first option is "Change Button Text" and makes a change of the button label possible for each of the top-level classes/categories. The change could simply customize the user's interface and does not effect the real ontological class or the ontological construction as a whole. The second option is "Change Highlight Colour" and concerns the colours in which the recognized entities of a particular class found on the web page will appear.

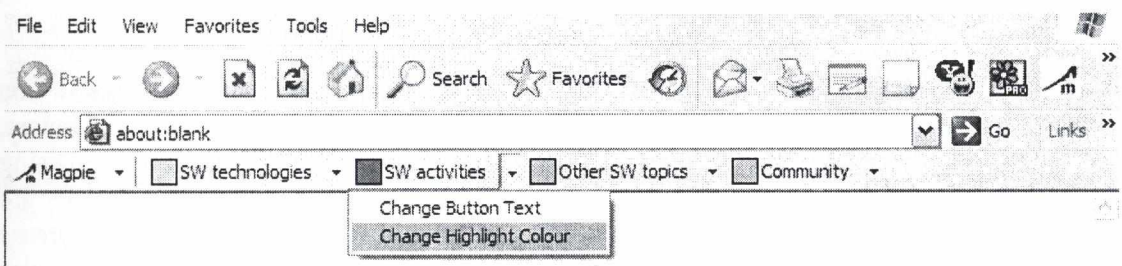


Fig. 1. Magpie with loaded ontological lexicon[4]

Fig. 2 gives an idea of the way Magpie is working. The loaded page (chosen by the user) is Magpie home page[5]. User's eyelet is formed by clicking on each/some of the four buttons, presenting ontological classes. Let's concentrate our attention on the class behind "Community" and search the loaded page for some instances of this class. Currently "Community" consists of instances of **kw-person**. The snapshot given on the ***Fig. 2*** shows that three instances of **kw-person** are recognized on the page. Clicking on any of these highlighted instances, we are able to invoke ontology dependent semantic menu offering information services, which are different for the different classes of concepts. For example, instances of **kw-person** class could be further investigated in respect to "Find community of referers", "Find co-authors", "Shares institution with", "Authored learning material", which is visible from ***Fig. 2***. The information accessed via these semantic services is not explicitly given on the page.

[3] http://www.aktors.org/publications/ontology

[4] The figure is taken from a report of the KnowledgeWeb Project publicly available at http://kmi.open.ac.uk/projects/kweb/resources/D3.3.3.pdf

[5] http://www.kmi.open.ac.uk/projects/magpie/main.html

Magpie user's eyelet. The chosen ontology is the Portal Ontology of AKT project available in several formats and announced as "the main ontology, describing people, projects, publications, geographical data, etc."[3] The lexicon contains records of four classes with a number of instances and enables Magpie to operate with this ontologically supported information. ***Fig. 1*** below shows a snapshot of semantically enriched browser (Internet Explorer) with the Magpie buttons corresponding to the four classes: **SW technologies**, **SW activities**, **Other SW topics**, **Community**.

Two functionalities of the buttons are available at this level and they are offered via a pull-down menu visible on ***Fig. 1***. The first option is "Change Button Text" and makes a change of the button label possible for each of the top-level classes/categories. The change could simply customize the user's interface and does not effect the real ontological class or the ontological construction as a whole. The second option is "Change Highlight Colour" and concerns the colours in which the recognized entities of a particular class found on the web page will appear.

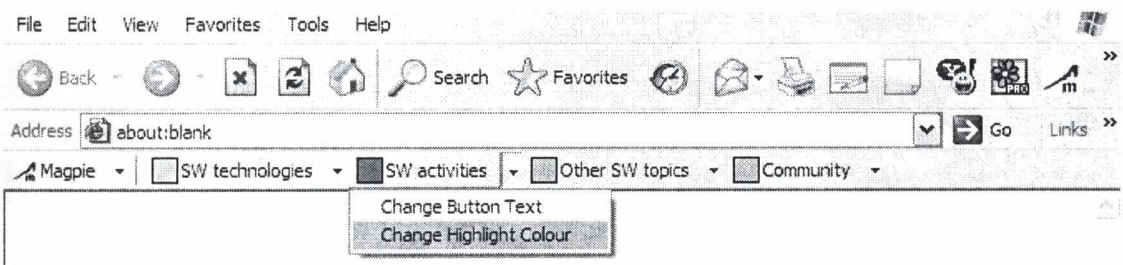


Fig. 1. Magpie with loaded ontological lexicon[4]

Fig. 2 gives an idea of the way Magpie is working. The loaded page (chosen by the user) is Magpie home page[5]. User's eyelet is formed by clicking on each/some of the four buttons, presenting ontological classes. Let's concentrate our attention on the class behind "Community" and search the loaded page for some instances of this class. Currently "Community" consists of instances of **kw-person**. The snapshot given on the ***Fig. 2*** shows that three instances of **kw-person** are recognized on the page. Clicking on any of these highlighted instances, we are able to invoke ontology dependent semantic menu offering information services, which are different for the different classes of concepts. For example, instances of **kw-person** class could be further investigated in respect to "Find community of referers", "Find co-authors", "Shares institution with", "Authored learning material", which is visible from ***Fig. 2***. The information accessed via these semantic services is not explicitly given on the page.

---

[3] http://www.aktors.org/publications/ontology

[4]The figure is taken from a report of the KnowledgeWeb Project publicly available at http://kmi.open.ac.uk/projects/kweb/resources/D3.3.3.pdf

[5] http://www.kmi.open.ac.uk/projects/magpie/main.html

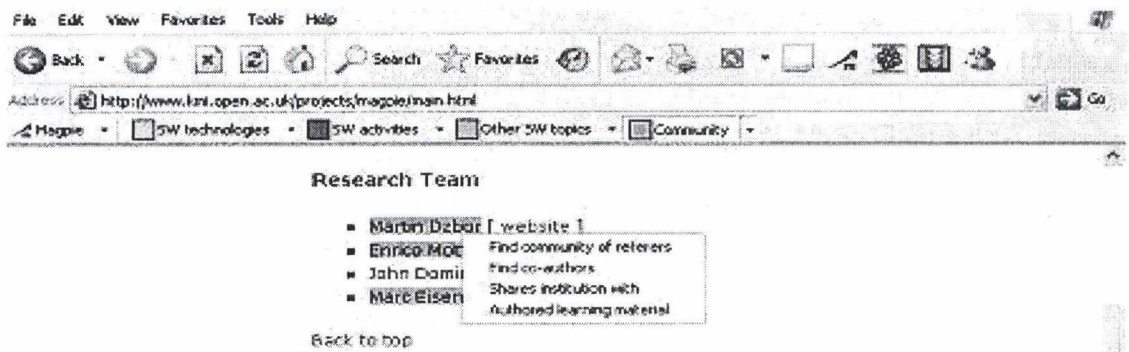


Fig. 2. Magpie in work - contextual semantic services

The nature of semantically (ontologically) dependent services and the degree of their sophistication doesn't depend on Magpie architecture, which considers all services as black boxes. The services could generate textual explanations, could attempt to show a related image or scheme if such exist in the repository of annotated materials connected to the ontology in use; the services could compute the needed answer at the moment, rather than provide links to any relevant documents.

Magpie provides support for trigger services also, which are based on "subscribe&acquire" user-system interaction. Dedicated interfaces, called *collectors*, visualize the result of such services. Trigger services use a semantic log ('browsing history') collected by Magpie during the browsing session. The recognized entities are asserted as facts into the Magpie semantic log and several watchers monitor *the patterns* in the asserted facts. When the relevant assertions have been made for a particular watcher, a semantic service response is triggered and applicable knowledge delivered to the Magpie plug in that in turn displays it in a dedicated window next to the user's web browser. [6] The patterns come from the user's subscription to a particular trigger service. This way the collectors show a semantically filtered view of the semantic log providing a structured record of user's browsing history.

Magpie seems to be a promising semantic web technology owing to all the features described above. The vigorous development of the area of knowledge representation and semantic knowledge applications raises the problem with the different understandings and implementations of ontologies inevitably.

3. MAGPIE WITH A LINGUISTICALLY MOTIVATED ONTOLOGY

Magpie offers the user a choice of ontology. It would be very comfortable to pick the ontology, which supports particular viewpoint to the web for the particular task. The idea behind the following discussion is that the choice of linguistically motivated ontology will be a good choice given that most of the information available from the web is (still) presented in textual form. The hypothesis is that a linguistically motivated ontology could bridge the gap between the text and its machine-understandable form easier than any kind of other ontology.

The experiment presented here is performed by use of Generalized Upper Model (GUM), ontology inspired and rooted deeply in the Systemic Functional Linguistics, which is a functional theory of language. The ontology is languages and tasks independent. It has been used as basic knowledge representation construction in the field of Natural Language Generation. Generalized Upper Model is "compared and contrasted with both non-linguistic and linguistic ontologies" [1]. Recently it has been translated into the Web Ontology Language (OWL)[6]. As a linguistically motivated ontology GUM is built with respect to natural language constructions at each level of natural language chunks. The inner organization of Generalized Upper Model comprises two hierarchies, one of CONCEPTs (or THINGS of the World), the second one of RELATIONS. The most general abstract entity at CONCEPTS hierarchy is named UM-THING. There are three major subtypes of UM-THING: ELEMENT, CONFIGURATION and SEQUENCE. ELEMENTs are standalone concepts or ontological items. CONFIGURATIONs involve ELEMENTs, which participate in particular event or state of affairs. SEQUENCEs are larger constructions presenting "complex situation where various activities or CONFIGURATIONs are connected by some relation." [1].

Currently it is impossible to choose Generalized Upper Model as Magpie ontology and the related theoretical and technical challenges are discussed below.

The First Challenge appears immediately from the fact, that there is no existing Knowledge Base or 'world' of facts and statements, which formal representation is based on Generalized Upper Model. GUM is top-down ontology with theoretically grounded construction and contains higher level concepts only. Formal representation of lower level expressions is strongly required for the experiment (and during the real browsing). The particular ontology behind the example given in Section 2, the Portal Ontology, is a bottom-up organized one and the creation of its classes is supported by classification of instances according to some of their features.

Despite this extremely discouraging difference in paradigms we could try to find in both ontologies some concepts which are supposed to carry the same meaning. Let's have for example the concept "set". The meaning of concept "set" is quite abstract and it is used as an organizational notion in ontologies as such. In GUM "set" is named "UM-SET" and appears quite near to the top notion of the ontology – UM-THING. This very abstract level of the concept UM-SET suggests the understanding that if there are defined some (particular or abstract) UM-THINGs they could be grouped to form a UM-SET: "An abstract assemblage of elements". The formal definition of UM-SET is "DECOMPOSABLE-OBJECT and ABSTRACTION". There are two subtypes of UM-SET in GUM: DISJUNCTIVE-SET and ORDERED-SET, with comments attached correspondingly: "A set of alternatives" and "A set whose elements are ordered". This high level conceptualization aims to introduce further time intervals, space as ordered points and so on. If we have a look at the Portal Ontology, we'll discover that probably the OWL terms **Class** and **SubClassOf** with the specified **Restriction** are nearer to the meaning of GUM concept UM-SET. In general, the task of different ontologies comparison is important within the society of knowledge engineers. It is clear already

[6] available from http://www.fb10.uni-bremen.de/ontology

that this task is not simple and solid mathematical apparatus and philosophical knowledge are needed to describe and maintain the possible analogies (see, for example, one attempt to be compared Euro Wordnet Top Ontology and UpperCyc Ontology in [8]). The recent work on ontologies comprises all the stages of ontologies lifecycle- creation, change, deployment, evaluation, analysis, merge, refine, assemble, and maintenance versioning of ontologies.

Back to the attempt to use Magpie with Generalized Upper Model, we could simply look at the pragmatic site of the problem. We need a formal representation of some facts only as minimal data needed to experiment with Magpie working with a linguistically motivated ontology. An ontological lexicon with GUM concepts and expressions has to be built to present the particular classes, instances and relations, which support the facts of the particular example given above. At this point we face ***The Second Challenge*** which could be given the name Attitude to Labelling.

It was shown on ***Fig. 1*** that Magpie offers functionality to change the labels on the buttons, which constitute the chosen viewpoint at the moment. There are only two alternatives: first, the labels of the ontological classes are tightly connected to the nature of classes as ontological constructions or second alternative, the labels to be easily attached and changed not breaking the ontological construction. The linguists presume that "word is important" even used as label; they tend to search for nuances of meaning in almost equal natural language expressions and believe that any (linguistically motivated) ontology should make difference of these nuances. That's why they will be surprised that the labels like COMMUNITY and KW-PERSON could be mutually commuted. On the other hand, the Magpie creators belong to a different scientific society, where the researchers don't have the same attitude to the natural language. The labels themselves don't possess such an importance in knowledge management area; logically and pragmatically thinking, 'A4509', 'ATTENTION' and '->!!!<-' could be the same if behind the label there is represented the same thing or notion.

Currently the two perspectives are brought closer. From the linguists' side the step is taken to Computational Linguistics. This is an area where the significance of words and natural language expressions is respected and diversity of computational methods are involved in natural language processing. Part of this work is creating and using linguistically motivated ontologies. From the knowledge engineers' side, the step is in direction of taking in account human-beings' associative thinking. Probably all novice computer programmers receive the invaluable advice when giving names to an entity within their programs- variables, files, folders etc. to choose word groups which "mean something". All software companies take care on 'user friendly' interfaces, which rely on signs, images, sound, but mainly on natural language expressions. Often the labels are meaningful. For example, this was the hypothesis of the researchers who experimented to extract semantic information from the web directories of Yahoo! and Google [9]. It could be stated that the labels are simply strings even when used within a linguistically motivated ontology, but when viewed as natural language chunks the bridge to the area of Linguistics is passable in both directions.

Back to the example and functionality of Magpie shown on ***Fig. 2*** we could conclude that the change of the label is useful option, when the user is aware that this change won't alter the ontological construction or, with other words, the formal semantic representation of the item. "Change Button Text" concerns the human-user and her/his comfort in interface interaction. This functionality could be much more useful if the change IS connected to the ontological construction supporting the task and if the manual change of a label CAUSES an 'automated' change in user's viewpoint. This idea is clarified further.

The task of our experiment is defined as follows: The meaning of KW-PERSON as it is given in the particular Magpie example has to be represented in GUM concepts. First, the right meaning of **KW-Person** should be found. The formal definitions of **KW-Person** and **KW** (Knowledge Web) are not available from the Portal Ontology. **Person** is defined as follows:

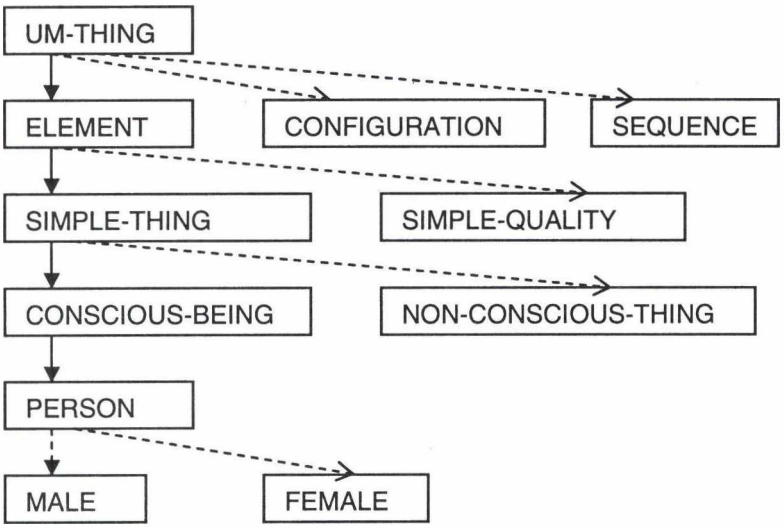
Person is a
OWL Class,
SubclassOf **Legal-Agent**,
DatatypeProperties are **full-name, family-name, given-name;**
ObjectProperties are **has-gender, has-academic-degree, has-appellation.**

If the label KW-PERSON is taken as a meaningful phrase and we assume that **KW** means "the project Knowledge Web"[7], the following semantic representations could be established for the notion, without claiming that the proposed list is exhaustive:

PERSON related to KW-PROJECT
PERSON participating in KW-PROJECT
PERSON involved in KW-PROJECT
PERSON working for KW-PROJECT
PERSON sponsoring KW-PROJECT
PERSON who believes in KW-PROJECT

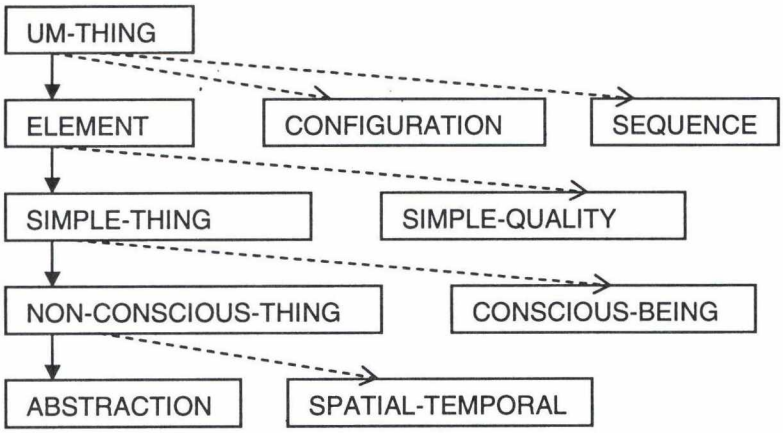
The meanings of **Person** and **KW** are substantially important for the formal representation. According to GUM ontology concept named PERSON is defined quite near the top concept UM-THING:

[7] http://knowledgeweb.semanticweb.org/



It is important from linguistic point of view, that some of the basic CONFIGURATION TYPES are dependent of 'consciousness', such CONFIGURATIONs are all the concepts belonging to SAYING&SENSING branch of the ontology. Gender is very important, too, because it could be marked explicitly in natural language expressions, e.g. by pronominalization. So, the need to have **ObjectProperty has-gender** for each **Person** within the Portal Ontology could be seen as supported by the linguistic reasons used in GUM building.

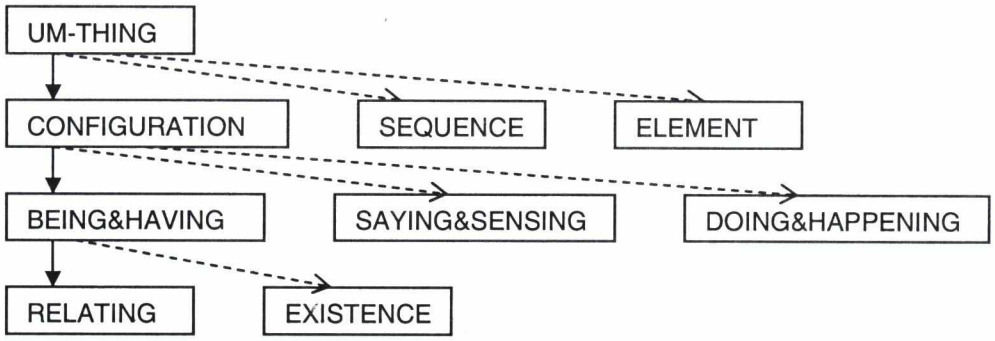
If **KW** is the particular "Knowledge Web project", and we take "Knowledge Web" as a name without analysis, the concept could be defined in GUM following the PROJECT semantics:



ABSTRACTIONs are described as "something that exists in metaphorical or qualitative space rather than in physical space", which fits the understanding of PROJECT like "a plan or proposal". PROJECT like "an undertaking requiring concerted effort" or "an

extensive task..."[8] could be a SPATIAL-TEMPORAL or an ELEMENT with characteristics from 'physical world', especially connected to the main concept there - DOING. ***The Third Challenge*** in our experiment is the problem with formal representation which is required to be expressive enough and easy for automated reasoning at the same time. It is demonstrated with the KW-PROJECT definition above and the KW-PERSON definition which follows.

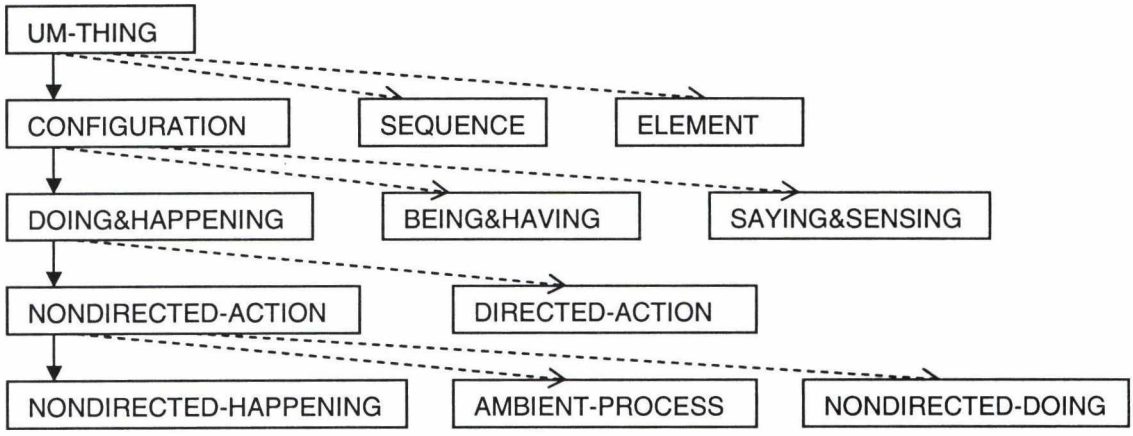
KW-PERSON = PERSON related to KW-PROJECT



This is a general definition which gives no information about the type of relation between the PERSON and the KW-PROJECT. RELATING is a wide area of concepts in GUM, it comprises CONFIGURATIONS of possession (have something), positioning (be somewhere) and intensive relations (be something). Most usual reading of KW-PERSON is probably "PERSON participating in KW-PROJECT", which is presented in GUM concepts again in RELATING area in the sense near to 'belong to'. Sometimes it is useful to have general formal representation for the semantics of the particular concept and GUM presents the generalization natural for human languages. When more particular meaning has to be expressed the lower branches of the ontology are used. KW-PERSON expressed as "PERSON involved in KW-PROJECT" is presented as follows:

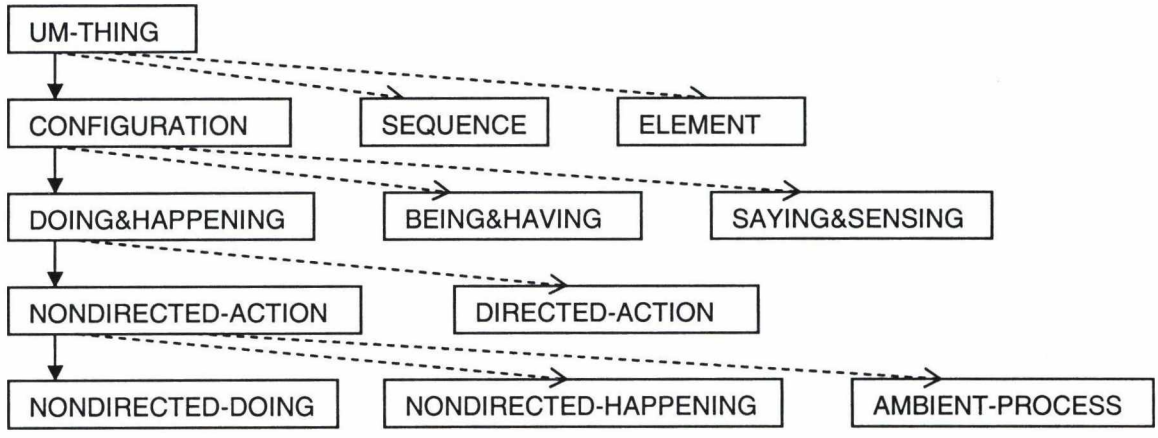
[8] http://www.thefreedictionary.com/project

KW-PERSON = PERSON involved in KW-PROJECT



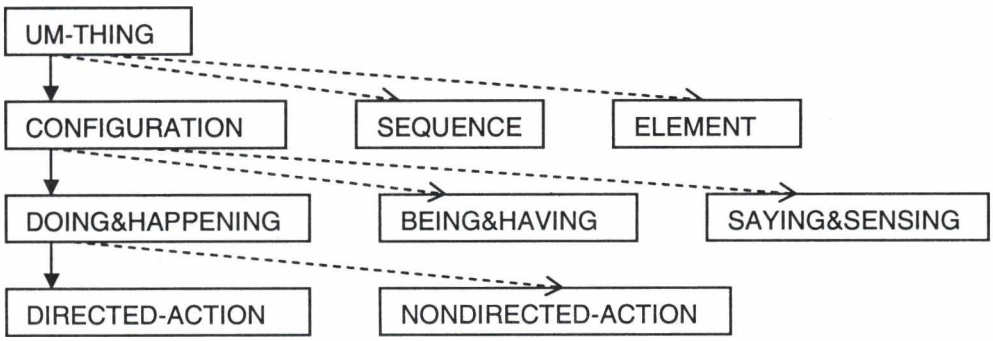
Typical NONDIRECTED-HAPPENINGs are CONFIGURATIONs which have one participant in the process only. CONFIGURATION INVOLVE is typical DIRECTED-ACTION, but the meaning of the passive phrase "PERSON involved in KW-PROJECT" is nearer to HAPPENING from the PERSON's point of view, when the involvement is not a volitional act concerning the PERSON. Formal representations of ACTIVE-PASSIVE pairs raise an interesting discussion among linguist, but it is not a topic of this presentation. If it is important to be shown that the PERSON is involved consciously and conscientiously in the KW-PROJECT, it is better to read KW-PERSON as "PERSON working for KW-PROJECT".

KW-PERSON = PERSON working for KW-PROJECT



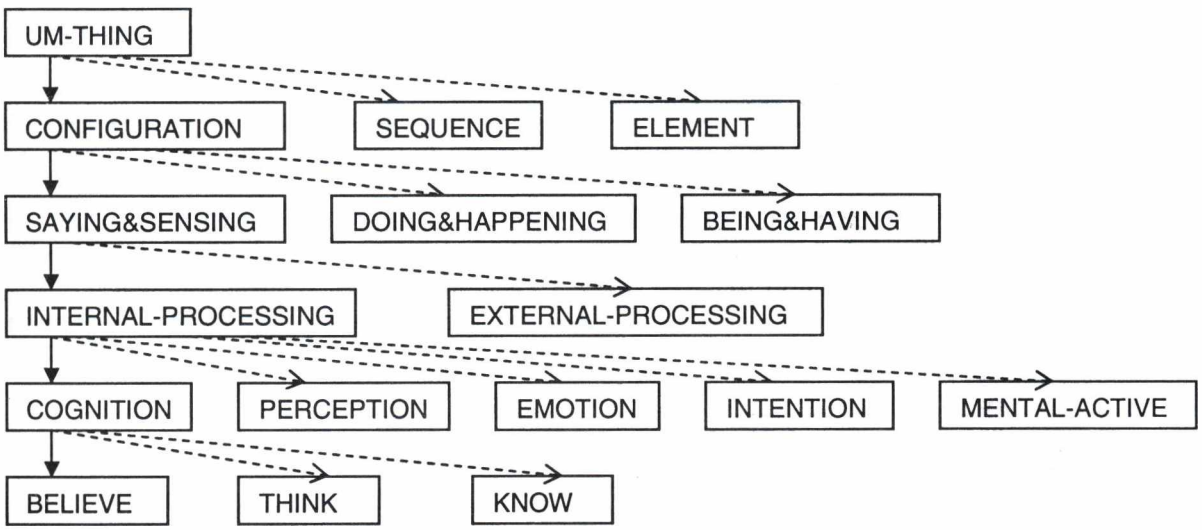
Different type of relation between PERSON and KW-PROJECT could be SPONSORING:

KW-PERSON = PERSON sponsoring KW-PROJECT



KW-PERSON could be defined as somebody who believes in ideas of Knowledge Web. In this case the CONFIGURATION is expressing mental activity and the concept KW-PROJECT should have the meaning of ABSTRACTION rather than of 'work', 'task' or 'undertaking'.

KW-PERSON = PERSON who believes in KW-PROJECT



The variety of definitions shown in the simple example above demonstrates the expressive power of linguistically motivated ontologies and the problems with formal representations of this kind. For example, the two-fold meaning of KW-PROJECT mentioned above could be a source of confusion if the rules to build formal expressions are not clear. On the other hand, not always the full (and too heavy) definitions of concepts are needed, especially when labels are 'meaningful'.

KW-PERSON could be defined in general (using "PERSON related to KW-PROJECT"), or in much more particular concepts. The generalization presented explicitly by the different ontological levels could be used substantially.

If we go back to the example shown on ***Fig. 2*** above and if we assume that the KW-PERSON behind the COMMUNITY button is supported by the expression PERSON related to KW-PROJECT then the instances of KW-PERSONs found by Magpie on the loaded webpage: "Martin Dzbor", "Enrico Motta", "Mark Eisenstadt" are highlighted. If the button's label "KW-PERSON" could be substituted with its representation "PERSON related to KW-PROJECT" and if the three concepts PERSON, RELATE and KW-PROJECT are presented within the ontology, the user could change her/his eyelet. substituting KW-PERSON with each of the three concepts. If PERSON is chosen, probably all the four names of the Research Team on the page will be recognized as PERSONs and highlighted.

The real challenging task for knowledge engineers is the effect of the choice of concept RELATE within the user's eyelet instead of the concept KW-PERSON. It will be useful to have an access from the CONFIGURATION

PERSON related to KW-PROJECT

to the other semantically related CONFIGURATIONs within the ontology:

PERSON participating in KW-PROJECT
PERSON involved in KW-PROJECT
PERSON working for KW-PROJECT
and so on.

The biggest difficulty is not only the lack of formal representation of knowledge supported by linguistically motivated ontology, but also the lack of mechanisms to map the texts available from the web to such a representation. Natural Language Processing systems are still too heavy and still do not show the expected efficiency of performance.

CONCLUSIONS

It is quite challenging, but seems not impossible to use modern semantic web technologies with linguistically motivated ontology. The question is whether the price of the planned efforts is not too high for the expected benefits, listed below.

The user expresses her/his knowledge in ontological constructions nearer to natural language ones much naturally than to an artificially constructed ontologies.
The user's point of view while browsing the web with a browser extension like Magpie is supported by particular ontology. Possible change of this eyelet will be much easier if the supporting ontology is linguistically motivated.
It is expected the information extraction from natural language texts to linguistically motivated ontologies to be faster and cheaper than to other types of ontological constructions.

Dynamic services and software agents of the emerging Semantic Web don't have to be effected if their ontologies are well formulated, even linguistically motivated ones.

Contrary, the services working with natural language texts are expected to improve their performance.

It seems rewarding and even necessary natural language technologies and linguistically motivated ontologies to be involved in the Semantic Web activities more intensively.

REFERENCES

[1] Bateman, John, Renate Henschel, Fabio Rinaldi (1995), *The Generalized Upper Model 2.0*, IPSI, Darmstadt http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/gum/index.htm

[2] Domingue J. B., Dzbor M. (2004), *Magpie: Browsing and Navigating on the Semantic Web*, In Proc. of the Conference on Intelligent User Interfaces, Portugal

[3] Domingue J. B., Dzbor M., Motta E. (2003), *Semantic layering with Magpie*, book chapter in Handbook on Ontologies in Information Systems (edited by Staab, S. & Studer, R.), Springer Verlag

[4] Domingue J. B., Dzbor M., Motta E. (2004), *Collaborative Semantic Web Browsing with Magpie*, In Proc. of the 1st European Semantic Web Symposium (ESWS), Greece

[5] Dzbor, M., Domingue J. B., Motta E. (2003), *Magpie - towards a semantic web browser*, In Proc. of the 2nd Intl. Semantic Web Conf., Florida, USA

[6] Dzbor, M., Motta E., Domingue J. B. (2004), *Opening Up Magpie via Semantic Services*, In Proc. of the 3rd Intl. Semantic Web Conference, Japan

[7] Gruber, T. R. (1992), *Ontolingua: a Mechanism to Support Portable Ontologies,* Report KSL, Stanford University, pp 61-66

[8] Kiryakov, A. and Simov K. (2000), *Mapping of Euro Wordnet Top Ontology into UpperCyc Ontology,* Proceedings of EKAW Workshop on Ontologies and Text

[9] Magnini, B., L. Serafini, and M. Speranza (2003), *Making Explicit the Semantics Hidden in Schema Models*, In Proc. of the 2nd Int. Semantic Web Conf., Florida, USA

INFORMATION RETRIEVAL TECHNOLOGIES FOR REAL WORLD TASKS

Angel Veselinov Velikov
Institute of Information Technologies – Bulgarian Academy of Sciences
Sofia 1113, Acad. G. Bonchev str., bl.2
avelikov@dorsum-bg.com

Abstract

The problem of the Information Retrieval is subject to long years of scientific research. We will try to review some of the presently available technologies and features found in tools that can be used for real world tasks.

Keywords

Information retrieval, search engine

1. INTRODUCTION

Every researcher, no matter the field he is working in, faces eventually the need to find particular information buried in a heap of non-relevant documents. Every enterprise that maintains vast databases and archives of documents and other kinds of media that is not fully structured or is structured in an inappropriate fashion struggles to find an effective solution to manage it.

Information Retrieval software is such solution. It has evolved from simple exact and partial word matching algorithms to intelligent systems using advanced techniques to analyze and classify the media and profile the search by continuously adapting to the user, while solving purely technical problems like the increasing need of storage space and processing power and handling the media formats.

Information retrieval is concerned with identifying documents in a collection that best match a description of a searcher's information need and are likely to contain the information one is looking for. Central to any effective retrieval system is the identification and representation of document content, the acquisition and representation of the information need, and the specification of a matching function that selects relevant documents based on these representations.

A typical user interface presents a way to enter some constraints on the search, based on the properties of the media, e.g. for files it can be the size, the type, the date etc.

The more advanced software may have the possibility to constrain properties that are not intrinsic and are based on some rating method or classification. Such classification can be done automatically based on different technologies and heuristics, but some service providers base the classification on human decision. Both ways are questionable and while the humans can give more accurate classification, the machines are faster and cheaper to use. The classification can be used indirectly in other approaches to determine the relevance of the document.

The queries usually have a text field which contents are the subject of the most advances in the field of the Information Retrieval. Basically making a partial or exact word matching, the search software retrieves a list of possible documents. A more fuzzy approach is to sort all the documents by their relevance to the query. The middle way that is most common is to present sorted only the most relevant documents.

Actually the relevance of a document is the most important measure, but unfortunately it cannot be defined and is based solely on the user's opinion. To estimate the performance of a particular approach to the Information Retrieval problem we evaluate their recall and precision. Recall is the fraction of all relevant documents that are found by the software. Precision we call the fraction of the relevant documents among all returned from the search. The ultimate aim is to make these 100%.

It is obvious that the word matching is far from such results. The Boolean search gives the user the opportunity to improve the results by combining the words in the query with Boolean operators. When the Boolean search allows using of parenthesis to alter the priority of the operators it is referred to as faceted search.

Another improvement is to search by phrases of words usually enclosed by quote marks. This phrase search generally decides about the document's relevance first by the phrase words found and then by their order in it. This needs the software to store the word positions too. But if we have the word positions we can implement the proximity search. It allows defining the maximum distance between the words searched. If the software allows defining the order too, then we have a very good extension of the phrase search.

When the software is big enough it can use some statistics of the user behavior to update the document relevance. This approach is called click tracking because it is usually based on what user clicks, in what order and how long it keeps his attention.

Another mathematical approach is the vector search. It calculates the relevance with complex formulas based on the query, document contents and the statistics derived from them. A common statistic used is the Term Frequency/Inverse Document Frequency (TF/IDF). It considers the most relevant the documents with many instances of the search terms which are rare in the rest of the collection.

These techniques are useful for bigger queries or when trying to find similar documents. A "more like this" interface is used when the query is too short. It makes an initial list of documents from which the user can select the most relevant and then the search continues based on the relevance of this document. A possible next step can be the link tracking which tracks the links to the selected document.

According to [19] an IR model is characterized by four parameters:

- representations for documents and queries,
- matching strategies for assessing the relevance of documents to a user query,
- methods for ranking query output, and
- mechanisms for acquiring user-relevance feedback.

IR models can be classed into four types: set theoretic, algebraic, probabilistic, and hybrid models.

A classification is provided in [14] for the different software solutions for IR:

- Any arbitrary method of Internet searching we call a "*search service*" (the most general term).
- A manually administered database we call a "*catalog*" or a "*directory*" (or in simple cases a "*list*").
- An automatic robot, which indexes Internet data mainly by itself, is called a "*searchengine*".
- An automat, which scans searchengines in parallel and merges the results, will be called a "*meta-searchengine*".
    - If this automat is running as a client on the user PC, we call it "*client-based*" meta-searchengine.
    - If this automat is running as a server, answering queries of many users, it's a "*server-based*" meta-searchengine.
- A WWW form which allows to query several searchengines one after another is called an "*all-in-one-form*" (and not a meta-searchengine).

2. DATA RETRIEVAL AND INFORMATION RETRIEVAL

In the classic book of Rijsbergen [6] he makes a clear distinction between the two terms, providing a table:

Table 1. Data retrieval or information retrieval?

	Data Retrieval (DR)	**Information Retrieval (IR)**
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

"In data retrieval we are normally looking for an exact match, that is, we are checking to see whether an item is or is not present in the file. In information retrieval this may sometimes be of interest but more generally we want to find those items which partially match the request and then select from those a few of the best matching ones.

The inference used in data retrieval is of the simple deductive kind, that is, aRb and bRc then aRc. In information retrieval it is far more common to use inductive inference; relations are only specified with a degree of certainty or uncertainty and hence our confidence in the inference is variable. This distinction leads one to describe data retrieval as deterministic but information retrieval as probabilistic. Frequently Bayes' Theorem is invoked to carry out inferences in IR, but in DR probabilities do not enter into the processing.

Another distinction can be made in terms of classifications that are likely to be useful. In DR we are most likely to be interested in a monothetic classification, that is, one with classes defined by objects possessing attributes both necessary and sufficient to belong to a

class. In IR such a classification is on the whole not very useful, in fact more often a polythetic classification is what is wanted. In such a classification each individual in a class will possess only a proportion of all the attributes possessed by all the members of that class. Hence no attribute is necessary nor sufficient for membership to a class.

The query language for DR will generally be of the artificial kind, one with restricted syntax and vocabulary, in IR we prefer to use natural language although there are some notable exceptions. In DR the query is generally a complete specification of what is wanted, in IR it is invariably incomplete. This last difference arises partly from the fact that in IR we are searching for relevant documents as opposed to exactly matching items. The extent of the match in IR is assumed to indicate the likelihood of the relevance of that item. One simple consequence of this difference is that DR is more sensitive to error in the sense that, an error in matching will not retrieve the wanted item which implies a total failure of the system. In IR small errors in matching generally do not affect performance of the system significantly." [6]

3. INDEXING

Indexing is the process of developing a document representation by assigning content descriptors or terms to the document. These terms are used in assessing the relevance of a document to a user query.

Indexing in general is concerned with assigning nonobjective terms to documents. The assignment may optionally include a weight indicating the extent to which the term represents or reflects the information content.

Nonobjective terms are intended to reflect the information manifested in the document and there is no agreement about the choice or degree of applicability of these terms. They are also known as content terms.

On the contrary, the objective terms are extrinsic to semantic content, like author name, document URL, and date of publication.

The effectiveness of an indexing system is controlled by two main parameters. Indexing exhaustivity reflects the degree to which all the subject matter manifested in a document is actually recognized by the indexing system. When the indexing system is exhaustive, it generates a large number of terms to reflect all aspects of the subject matter present in the document; when it is nonexhaustive, it generates fewer terms, corresponding to the major subjects in the document. Term specificity refers to the breadth of the terms used for indexing. Broad terms retrieve many useful documents along with a significant number of irrelevant ones; narrow terms retrieve fewer documents and may miss some relevant items.

The two parameters directly affect the recall and the precision of the search results. Ideally, you would like to achieve both high recall and high precision. In reality, you must strike a compromise. Indexing terms that are specific yields higher precision at the expense of recall. Indexing terms that are broad yields higher recall at the cost of precision. For this reason, the effectiveness is measured by the precision parameter at various recall levels.

We can divide the indexing approaches in two categories: atomic and complex indexing. The atomic indexing works on the level of the single term while the complex indexing gets

advantage of the relations between the terms in a bigger structure. When indexing a text, the two approaches are called single-term and multi-term or phrase indexing.

3.1 Single-Term Indexing

The term set of the document includes its set of words and their frequency. Words that perform strictly grammatical functions are compiled into a stop listand removed. The term set can also be refined by stemming to remove word suffixes.

Approaches to assigning weights for single terms may be grouped into the following categories: statistical, information-theoretic, and probabilistic. While the first two categories just use document and collection properties, the probabilistic approaches require user input in terms of relevance judgments.

3.1.1 Statistical methods

Assume that we have N documents in a collection. Let tf_{ij} denote the term frequency, which is a function of the frequency of the term T_j in document D_i.

Indexing based on term frequency fulfills one indexing aim, namely, recall. However, terms that have concentration in a few documents of a collection can be used to improve precision by distinguishing documents in which they occur from those in which they do not. Let df_j denote the document frequency of the term T_j in a collection of N documents, which is the number of documents in which the term occurs. Then, the inverse document frequency, given by $\log(N/df_j)$, is an appropriate indicator of T_j as a document discriminator.

The term-frequency and inverse-document-frequency components can be combined into a single frequency-based indexing model, where the weight of a term T_j in document D_i denoted w_{ij} is given by

$$w_{ij}=tf_{ij} \log(N/df_j)$$

Another statistical approach to indexing is based on term discrimination. This approach views each document as a point in the document space. As the term sets for two documents become more similar, the corresponding points in the document space become closer (that is, the density of the document space increases) and vice versa.

Under this scheme, we can approximate the value of a term as a document discriminator based on the type of change that occurs in the document space when a term is introduced to the collection. We can quantify this change according to the increase or decrease in the average distance between the documents. A term has a good discrimination value if it increases the average distance between the documents; in other words, terms with good discrimination value decrease the density of the document space. The term-discrimination value of a term T_j, denoted dv_j, is then computed as the difference of the document space densities before and after the term T_j is introduced. The net effect is that high-frequency terms have negative discrimination values, medium-frequency terms have positive discrimination values, and lowfrequency terms tend to have discrimination values close to zero. A term-weighting scheme such as $w_{ij}= tf_{ij}dv_j$ is used to combine term frequency and discrimination values.

3.1.2 Information-theoretic methods

In information theory, the least-predictable terms carry the greatest information value. Least-predictable terms are those that occur with smallest probabilities. Information theory concepts have been used to derive a measure, called signal-noise ratio, of term usefulness for indexing. This method favors terms that are concentrated in particular documents. Therefore, its properties are similar to those of inverse document frequency.

3.1.3 Probabilistic methods

Probabilistic approaches require a training set of documents obtained by asking users to provide relevance judgments with respect to query results. The training set is used to compute term weights by estimating conditional probabilities that a term occurs given that a document is relevant (or irrelevant). Assume that a collection of N documents of which R are relevant to the user query, R_t of the relevant documents contain term t, and t occurs in f_t documents. Two conditional probabilities are estimated for each term as follows:

Pr [t in document|document is relevant] = R_t/R;

Pr [t in document|document is irrelevant] = $(f_t - R_t)/(N - R)$.

From these estimates, Bayes' theorem is used, under certain assumptions, to derive the weight of term t as

$$w_t = \log \frac{R_t/(R - R_t)}{(f_t - R_t)/(N - f_t - (R - R_t))}$$

The numerator (denominator) expresses the odds of term t occurring in a (irrelevant) relevant document. Term weights greater than 0 indicate that the term's occurrence in the document is evidence of the document's relevance to the query; values less than 0 indicate its irrelevance.

3.2 Multi-term or phrase indexing

Single terms are less than ideal for an indexing scheme because their meanings out of context are often ambiguous. Term phrases, on the other hand, carry more specific meaning and thus have more discriminating power. Phrase generation is intended to improve precision; thesaurus-group generation is expected to improve recall. A thesaurus assembles groups of related specific terms under more general, higher level class indicators.

Methods for generating complex index terms or term phrases automatically may be categorized as statistical, probabilistic, or linguistic.

3.2.1 Statistical methods

A term phrase consists of the phrase head, which is the principal phrase component, and other components. A term with document frequency exceeding a stated threshold, such as df>2, is designated as the phrase head. Other components of the phrase should be medium- or low-frequency terms with stated co-occurrence relationships with the phrase head, for example, that the phrase components should occur in the same sentence as the phrase head within a stated number of words.

Term groupingor term clusteringmethods are used to generate groups of related words by observing word co-occurrence patterns in a document collection. Given a term-document matrix as a 2D array, one method compares the columns of the matrix to each other and assesses whether the terms are jointly assigned to many documents in the collection. If so, the terms are assumed to be related and are grouped into the same class.

3.2.2 Probabilistic methods

Probabilistic methods generate complex index terms based on term-dependence information. Since this requires considering an exponential number of term combinations and, for each combination, estimating the probabilities of coincidences in relevant and irrelevant documents, only certain dependent term pairs are considered in practice. In theory, these dependencies can be user specific.

In the statistical and probabilistic approaches, terms that occur together are not necessarily related semantically. Therefore, these approaches are not likely to lead to high-quality indexing units.

3.2.3 Linguistic methods

Assigning syntactic class indicators such as adjective, noun, or verb to terms can enhance the statistical method described above. Phrase formation is then limited to sequences of specified syntactic indicators (for example, noun-noun, adjective-noun). A simple syntactic analysis process can be used to identify syntactic units. The phrase elements can then be chosen from within the same syntactic unit.

Linguistic approaches for generating term relationships usually involve the use of an electronic lexicon. There are also proposals for generating term relationships based on user feedback. Though various automatic methods for thesaurus construction have been proposed, their effectiveness is questionable outside the special environments in which they are generated.

4. SET THEORETIC MODELS

The Boolean model represents documents by a set of index terms, each of which is viewed as a Boolean variable and valued as True if it is present in a document. No term weighting is allowed. Queries are specified as arbitrary Boolean expressions formed by linking terms through the standard logical operators: AND, OR, and NOT. Retrieval status value (RSV) is a measure of the query-document similarity. In the Boolean model, RSV equals 1 if the query expression evaluates to True; RSV is 0 otherwise. All documents whose RSV evaluates to 1 are considered relevant to the query. This model is simple to implement and many commercial systems are based on it. User queries can employ arbitrarily complex expressions, but retrieval performance tends to be poor. It is not possible to rank the output, since all retrieved documents have the same RSV, and weights can not be assigned to query terms. The results are often counter-intuitive. For example, if the user query specifies 10 terms linked by the logical connective AND, a document that has nine of these terms is not retrieved. User relevance feedback is often used in IR systems to improve retrieval effectiveness. Typically, a user is asked to indicate the relevance or irrelevance of

a few documents placed at the top of the output. Since the output is not ranked, however, the selection of documents for relevance feedback elicitation is difficult.

The fuzzy-set model is based on fuzzy-set theory, which allows partial membership in a set, as compared with conventional set theory, which does not. It redefines logical operators appropriately to include partial set membership, and processes user queries in a manner similar to the case of the Boolean model. Nevertheless, IR systems based on the fuzzy-set model have proved nearly as incapable of discriminating among the retrieved output as systems based on the Boolean model. The strict Boolean and fuzzy-set models are preferable to other models in terms of computational requirements, which are low in terms of both the disk space required for storing document representations and the algorithmic complexity of indexing and computing query-document similarities.

5. ALGEBRAIC MODELS

The vector-space model is based on the premise that documents in a collection can be represented by a set of vectors in a space spanned by a set of normalized term vectors. If the vector space is spanned by n normalized term vectors, then each document will be represented by an n-dimensional vector. The value of the first component in this vector reflects the weight of the term in the document corresponding to the first dimension of the vector space, and so forth. A user query is similarly represented by an n-dimensional vector. A query-document's RSV is given by the scalar product of the query and document vectors or the cosine of the angle between them. The higher the RSV, the greater is the document's relevance to the query.

A second major achievement of the researchers that developed the vector space model is the introduction of the family of TF/IDF term weights. These weights have a term frequency (TF) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency (IDF) factor measuring the inverse of the number of documents that contain a query or document term. If the cosine measure is used, the vector lengths are normalised. The three components TF, IDF and length normalisation can be calculated by various formulas and are often reported upon by a three letter combination [20]. One of the recent weighting algorithms LNU/LTU [21] uses a combination of the document length and the average document length instead of the cosine measure for length normalisation. This algorithm outperforms the cosine versions on the TREC collections, but lacks the metaphor of measuring the angle between two vectors.

The strength of this model lies in its simplicity. Relevance feedback can be easily incorporated into it. However, the rich expressiveness of query specification inherent in the Boolean model is sacrificed.

6. PROBABILISTIC MODELS

The vector-space model assumes that the term vectors spanning the space are orthogonal and that existing term relationships need not be taken into account so they are statistically independent. Furthermore, the model does not specify the query-document similarity, which must be chosen somewhat arbitrarily. The probabilistic model takes these term

dependencies and relationships into account and, in fact, specifies major parameters such as the weights of the query terms and the form of the query document similarity.

The model is based on two main parameters—Pr(rel) and Pr(nonrel), the probabilities of relevance and non-relevance of a document to a user query—which are computed using the probabilistic term weights and the actual terms present in the document. Relevance is assumed to be a binary property so that Pr(rel) = 1 - Pr(nonrel). In addition, the model uses two cost parameters, a1 and a2, to represent the loss associated with the retrieval of an irrelevant document and non-retrieval of a relevant document, respectively.

The model requires term-occurrence probabilities in the relevant and irrelevant parts of the document collection, which are difficult to estimate. If no relevance information is available, the model behaves like the vector space model using IDF weights only (ignoring TF and length normalisation), that is, much poorer than most TF/IDF weights. However, this model serves an important function for characterizing retrieval processes and provides a theoretical justification for practices previously used on an empirical basis (for example, the introduction of certain term-weighting systems).

7. HYBRID MODELS

As in the case of the vector-space model, the extended Boolean model represents a document as a vector in a space spanned by a set of orthonormal term vectors. However, the extended Boolean (or p-norm) model measures query-document similarity by using a generalized scalar product between the corresponding vectors in the document space. This generalization uses the well-known L_p norm defined for an n-dimensional vector, d, where the length of d is given by

$$|d| = |(w_1, w_2, ..., w_n)| = \left(\sum_{j=1}^{n} w_j^p\right)^{\frac{1}{p}}$$

where $1 <= p <= \infty$, and $w_1, w_2, \ldots, w_n$ are the components of the vector d.

Generalized Boolean OR and AND operators are defined for the p-norm model. The interpretation of a query can be altered by using different values for p in computing query document similarity. When p = 1, the distinction between the Boolean operators AND and OR disappears, as in the case of the vector-space model.

When the query terms are all equally weighted and $p = \infty$, the interpretation of the query is the same as that in the fuzzy-set model. On the other hand, when the query terms are not weighted and $p = \infty$, the p-norm model behaves like the strict Boolean model. Varying the value of p from 1 to ∞ offers a retrieval model whose behavior corresponds to a point on the continuum spanning from the vector-space model to the fuzzy and strict Boolean models.

The best value for p is determined empirically for a collection, but is generally in the range $2 <= p <= 5$. [19]

8. RELEVANCE FEEDBACK TECHNIQUES

Unlike a database environment, an IR environment lacks precise representations for user queries and documents. Users typically start with an imprecise and incomplete query, and improve the query specification - hence, the retrieval effectiveness - iteratively and incrementally. The user is asked to provide evaluations or relevance feedback on the documents retrieved from the initial query. This feedback is used in subsequently improving the retrieval effectiveness.

Relevance feedback is elicited in either two-level or multilevel relevance relations. In the former case, the user simply labels a retrieved document as relevant or not; in the latter, a document can be relevant, somewhat relevant, or not relevant. Multilevel relevance can also be specified in terms of relationships. For example, for three retrieved documents d1, d2, and d3, d1 can be more relevant than d2, and d2 more relevant than d3.

For simplifying our presentation of relevance feedback, we assume two-level relevance and the vector-space model. The set of documents deemed relevant by the user constitute positive feedback; those deemed irrelevant constitute negative feedback.

The two major approaches to utilizing relevance feedback are modifying the query and modifying the document representations. Methods based on modifying the query representation affect only the current user-query session and have no effect on other user queries; methods based on modifying the representation of documents in the collection affect the retrieval effectiveness of future queries. In all the methods, more than two or three iterations may result in minimal improvements.

The basic assumption for relevance feedback is that documents relevant to a particular query resemble each other - in the vector-space model, their corresponding vectors are similar. Using relevance-feedback techniques in Web search engines requires document representations to be more descriptive and semantically rich than just indexing the title or abstract of the document. One way to achieve this is to index the entire document. The vector-space model readily accommodates all relevance feedback techniques, whereas the probabilistic model needs special extensions to accommodate query expansion.

8.1 Modifying the query representation

There are three ways to improve retrieval effectiveness by modifying the query representation. The first, modification of term weights, involves adjusting the query term weights by adding document vectors in the positive feedback set to the query vector. Optionally, negative feedback can be used to subtract document vectors in the negative feedback set from the query vector. The reformulated query should retrieve additional relevant documents similar to the documents in the positive feedback set. This process can be carried out iteratively until the user is satisfied with the quality and number of relevant documents in the query output.

Experimental results indicate that positive feedback is more consistently effective. This is due to the fact that documents in the positive feedback set are generally more homogeneous than documents in the negative feedback set. However, an effective feedback technique, termed dec hi, uses all documents in the positive feedback set and

subtracts from the query only the vectors of highest ranked irrelevant documents in the negative feedback set.

The second method, query expansion, modifies the original query by adding new terms to it. The new terms are selected from the positive feedback set and sorted using measures such as

- noise (a global term-distribution measure similar to idf),
- postings (the number of retrieved relevant documents containing the term),
- noise within postings,
- noise.frequency within postings (frequency is log2 of the total frequency of the term in the retrieved relevant set),
- noise.frequency.postings, and
- noise.frequency.

A predefined number of top terms from the sorted list are added to the query. Experimental results show that the last three sort methods produced the best results and that adding only selected terms is superior to adding all terms. There is no performance improvement by adding more than 20 terms [19].

In some cases, the above two techniques do not produce satisfactory results because the documents in the positive feedback set are not homogeneous (that is, they do not form a tight cluster in the document space) or because the irrelevant documents are scattered among certain relevant ones. One way to detect this situation is to cluster the documents in the positive feedback set to see if more than one homogeneous cluster exists. This method is called query splitting. If the documents cluster, the query is split into subqueries such that each sub query represents one cluster in the positive feedback set. The weight of terms in the subquery can then be adjusted or expanded as in the previous two methods.

8.2 Modifying the document representation

This approach involves adjusting the document vectors in the collection based on relevance feedback. It is also referred to as user-oriented clustering. It is implemented by adjusting the weights of retrieved and relevant document vectors to move them closer to the query vector. The weights of retrieved irrelevant-document vectors are adjusted to move them farther from the query vector. Care must be taken to ensure that individual document movement is small, since user-relevance assessments are necessarily subjective.

8.3 Personalization and Content-Based Image Retrieval (CBIR)

We can give the personalized multimedia management tool EGO as a representative of this technique, developed in the University of Glasgow [7].

It is concentrated on the "more like this" approach which seems like the most appropriate for the multimedia retrieval. When facing the lack of semantics and meta-data for the items that are retrieved, like images, sounds and movies, the reasonable relevancy information source that is left is the feedback from the user. It is used to fill the semantic gap by developing a more interactive user interface. It follows the organization of the user and personalizes the interface. EGO uses ideas that are already implemented in the past and improves them with additional mathematical models and better user interface. Its

implementation is limited to the images but the ideas can be used for other types of media too.

"The images are represented according to the hierarchical object model proposed in [8]. This model makes a distinction between the various visual features extracted. Rather than representing an image by a single stacked feature vector, it is composed of a set of feature vectors, one for each distinct feature implemented." [7].

The features used are low-level statistics of the image from the theory of the computational geometry and morphology. They are weighed with a feature transformation matrix and their Euclidean distance is measured. The distances are then combined linearly.

"The recommendation system is based on a relevance feedback algorithm, that attempts to learn the best query representation and feature weighting for a selected group of images. As far as the learning system is concerned, each group image is regarded as a positive training sample. The proposed group-based learning scheme involves (1) updating the system's matching parameters, (2) creating a multi-point query representation and computing a ranked list for each query point based on the learnt parameters, and (3) combining the individual result lists for the new recommendations." [7]

The learning of the feature weights is according to [9] using a weighted covariance matrix of the examples. Here the hierarchy is used to make difference between inter- and intra-feature weights.

The search /"recommendation"/ is then based on retrieved clusters for every example: "The clusters are computed by an agglomerative hierarchical clustering algorithm, using Ward's minimum variance criterion [10]. The ideal number of clusters is automatically estimated using the method presented in [11]. The query points are selected as the image closest to each cluster centroid. Each query point is associated with a weight relative to the cluster size..."

The results are combined using a rank-based voting approach.

A specialized review on the personalization techniques is [12]. It presents the "Fetch" tool that is concentrated on the user feedback methods and interfaces. Similar is the article [13] which concentrates on the "more like this" approach.

9. LINK TRACKING AND CLICK TRACKING

The link tracking is the subject of the link analysis area of the information retrieval. It studies the information that can be extracted from the links between the documents. Linking between web pages is their basic idea which makes this method very effective in searching Internet space. There is a common sense behind the linking heuristic:

- A link from page A to page B is a recommendation of page B by the author of page A.
- If page A and page B are connected by a link the probability that they are on the same topic is higher than if they are not connected. [15]

Such logic is followed in other fields like bibliometrics and sociology. The links can be studied on more abstract level as references. This means that we can apply additional

methods to extract from the document relations that are other than the hyperlinks. Such can be the citing of another document. This approach is connectivistic and uses graph representations and algorithms.

The first assumption of connectivity based techniques immediately leads to a simple query-independent criterion: The larger the number of hyperlinks pointing to a page the better the page. The main drawback of this approach is that each link is equally important. It cannot distinguish between the quality of a page pointed to by a number of low-quality pages and the quality of a page that gets pointed to by the same number of high-quality pages. Obviously, it is therefore easy to make a page appear to be high-quality – just create many other pages that point to it.

This issue is the basis of the work of Brin and Page [16]. Their PageRank measure works very well in distinguishing high-quality web pages from low-quality web pages. It assigns a score to each document independent of a specific query. This has the advantage that the link analysis is performed once and then can be used to rank all subsequent queries.

While PageRank is query-independent measure, Carriere and Kazman [17] proposed an indegree-based ranking approach to combine link analysis with a user query. They build for each query a subgraph of the link graph limited to pages on the query topic. More specifically, they use the following query-dependent neighborhood graph. A start set of documents matching the query is fetched from a search engine. This set is augmented by its neighborhood, which is the set of documents that either point to or are pointed to by documents in the start set. Since the in-degree of nodes can be very large, in practice a limited number of predecessors of a document are included. The in-degree-based approach ranks the nodes by their in-degree in the neighborhood graph. This approach has the same problem that each link counts an equal amount.

Combining the two approaches, Kleinberg [18] proposed the HITS algorithm. Given a user query, the algorithm first iteratively computes a hub score and an authority score for each node in the neighborhood graph. The documents are then ranked by hub and authority scores, respectively.

Nodes, i.e., documents that have high authority scores are expected to have relevant content, whereas documents with high hub scores are expected to contain hyperlinks to relevant content. The intuition is as follows. A document which points to many others might be a good hub, and a document that many documents point to might be a good authority. Recursively, a document that points to many good authorities might be an even better hub, and similarly a document pointed to by many good hubs might be an even better authority.

There are two types of problems with this approach: First, since it only considers a relatively small part of the graph, adding edges to a few nodes can potentially change the resulting hubs and authority scores considerably. A second problem is that if the neighborhood graph contains more pages on a topic different from the query, then it can happen that the top authority and hub pages are on this different topic. This problem was called topic drift.

Apart from ranking, link analysis can also be used for deciding which web pages to add to the collection of web pages, i.e., which pages to crawl. A crawler (or robot or spider) performs a traversal of the web graph with the goal of fetching high-quality pages. After fetching a page, it needs to decide which page out of the set of non-crawled pages to fetch next. One approach is to crawl the pages with highest number of links from the crawled pages first.

10. STANDARDS

We will mention several standards used in the different information retrieval systems. Most of them are related to the taxonomy and classification approaches.

10.1 ANSI Z39.19 – Thesaurus Construction

This standard helps for building universally compatible thesauri. It provides a series of standard operators that allow for the definition of a lexicon in a hierarchical fashion. There is support for broader meanings and narrower meanings, as well as synonyms. The "preferred term" operator, used to denote the preferred term amongst synonyms is typically a candidate for a taxonomy node label. The standard allows the user to decide between a pure hierarchy or a networked approach to word relationships. Availability of the "Scope Note" operator allows for intelligence about the language and specific words and phrases to be captured as well.

10.2 Dublin Core

This standard provides an approach to meta tagging unstructured content. Dublin Core came from the online library sharing area. It describes a standard set of meta tags that are used to track objects in a library collection, making the collections objects searchable in a standard manner. For organizations struggling with defining an internal standard for meta tagging, Dublin Core can provide a starting point. In practice, the standard tags are typically modified or customized for internal use.

10.3 RDF (Resource Description Framework)

RDF is a product of the semantic web and the W3C project. RDF specifies a syntax and schema for describing web content in XML. While focused on providing a standard means to exchange content and processes across web sites, the standard could be utilized as a way to construct a taxonomy. The operators provided enable the construction of an intelligent taxonomy, meaning the relationships between topics can be self described.

10.4 DAML (DARPA Agent Markup Language)

This standard is also a product of the semantic web. It provides an extension to RDF. With wider support for self-describing tags, ontologies can be defined in DAML. DAML allows the designer to specify not only the relationship of the link between two topics (e.g.. parent-child, "was born in," "is author of" links), but also to specify rules about the nature of the values in the link. For example, "disjointed" classes specify that a member of one node cannot also be the member of another node.

10.5 ISO/IEC 13250 Topic Maps

Topic maps are a standard XML schema, XTM, which utilizes meta tags to build a self-describing ontology, thesaurus or taxonomy. Topic Maps can be integrated to search features to provide intelligent search and/or to enable web site-to-web site automated communication and exchange.

The components of a topic map are:

- Topics – A declared "subject" (or more generally any "thing") associated with the topic is the name. There are also distinct types of names – base names, display names, and sort names.
- Occurrences – declarative links to web sites, other topics, or other forms of online content that represent (e.g. are examples of) the topic.
- Associations – An association is a link element that asserts a relationship between two or more topics, e.g. 'Grapes of Wrath' (a topic type book) is linked to topic "John Steinbeck" via the "was written by" association.

11. CONCLUSION

We can see that the leading role in the Information Retrieval research is now taken by the commercial software providers. They have the implemented and used in practice the latest technologies that are taxonomy and semantic oriented. There are tools for image and audio files searching that are now out of the labs and on the market.

These technologies are more oriented to the classification approach. Algorithms from the Natural Language Processing and Statistics theories are used to make automatic generation and update of taxonomies. Such approaches are Bayesian probability, Neural Networks, Support Vector Machine, Semantic and Linguistic analysis.

When available, the taxonomy is used to improve the document relevance or to give the user the opportunity to search in a more interactive fashion – by browsing. Sometimes this is more effective because the user doesn't have exact idea what he is searching for.

Similar to the taxonomy are the ontology and thesaurus based software. The incorporation of several methods is frequently used to get better results.

REFERENCES

[1] *Search Tools for Web Sites and Intranets*, http://www.searchtools.com

[2] *Search Engine Watch*, http://www.searchenginewatch.com

[3] Special Interest Group on Information Retrieval of the ACM (Association for Computer Machinery), http://www.sigir.org

[4] Glasgow Information Retrieval Group, http://ir.dcs.gla.ac.uk

[5] Delphi Group (2004), *Information Intelligence: Content Classification and the Enterprise Taxonomy Practice*

[6] Rijsbergen, C.J. *Information retrieval*

[7] Urban, J. and Jose, J. *EGO: A Personalised Multimedia Management Tool* Department of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK

[8] Rui, Y., Huang, T.S., Ortega, M., Mehrotra, S. (1998). Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.* 8, pp. 644-655

[9] Rui, Y., Huang, T.S. (2000). Optimizing learning in image retrieval. *IEEE Proc. Of Conf. on Computer Vision and Pattern Recognition,* pp. 236-245

[10] Theodoridis, S., Koutroumbas, K. (1999). *Pattern Recognition.* Academic Press

[11] Salvador, S., Chan, P. (2003). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Technical Report CS-2003-18*, Florida, Institute of Technology

[12] Martin, I. and Jose, J. *A Personalised Information Retrieval Tool* Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland

[13] White, R., Jose, J. and Ruthven, I. *A System Using Implicit Feedback and Top Ranking Sentences to Help Users Find Relevant Web Documents*

[14] Sander-Beuermann, W. and Schomburg, M. (1998) Internet Information Retrieval - The Further Development of Meta-Searchengine Technology. *Proceedings of the Internet Summit, Internet Society*, University of Hannover, Germany

[15] Henzinger, M. *Link Analysis in Web Information Retrieval*, Google Incorporated, Mountain View, California

[16] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998) The PageRank citation ranking: Bringing order to the Web. *Technologies,Working Paper* 1999-0120, Stanford Digital Library

[17] Carriere, J. and Kazman, R.. (1997) Webquery: Searching and visualizing the web through connectivity. *In Proceedings of the Sixth International World Wide Web Conference*, pp. 701–711.

[18] Kleinberg, J. (1998) Authoritative sources in a hyperlinked environment. *In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 668–677

[19] Gudivada, V., Raghavan, V., Grosky, W. and Kasanagottu, R. (1997). Information retrieval on the World Wide Web, *IEEE INTERNET COMPUTING*

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

[21] Singhal, A., Buckley, C. and Mitra, M. (1996). Pivoted document length normalization. *In Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21–29

MIXED AND WEIGHTED MEASURES FOR CLIENT BEHAVIOR PREDICTION IN A PROACTIVE VIDEO SERVER

Csaba Domokos
University of Szeged, Institute of Informatics
domokos@inf.u-szeged.hu

Erika Széll
University of Szeged, Institute of Informatics
Szell.Erika.1@stud.u-szeged.hu

Péter Kárpáti
University of Klagenfurt, Department of Information Technology
kpeter@itec.uni-klu.ac.at

László Böszörményi
University of Klagenfurt, Department of Information Technology
laszlo@itec.uni-klu.ac.at

Abstract

The precision of the predictors used in the ADMS [1] can be determined by similarity. There are already such measures [2] given (Kendall's tau, Spearman's footrule, Ulam's distance), but we do not know exactly what efficiency they have and how well they show the difference between two lists.

We examined the characteristics of these similarity measures and developed some more measures that fit better our needs. One of the main goals is to consider the similarity more important at the begin of list, than at the end of list. Because the clients at the begin of the list probably will request more videos. During our work we defined 20 special ordered lists with 17 elements each. We tested the different measures on these lists. We also tested the Kemeny distance, which was defined in paper [3]. We modified the Spearman's footrule and the Ulam's distance according to the goal defined above (the top of the list considerate with higher weight (Weighted Spearman's footrule, Weighted Ulam's distance). Using the already known measures we developed a more complex, mixed measure, which uses more factors when defining the similarity. Finally we compared the 7 different measures using the artificially defined lists.

With using the similarity measures we can tell how good the predictors [2] work in ADMS project. We could order the predictors by goodness, testing them on a real database (the World Cup '98 Website's access log).

Keywords

Similarity measures, predictor, ADMS.

1. INTRODUCTION

For efficient solved of an ADMS [1] in a large network like Internet there are some techniques but they are not perfect in case of large size files. Therefore it seems to be a good idea to replicate videos and make place where these are needed. It will be good to know the best place for a video, so should acquire this knowledge from the past behavior of clients. To predict the future behavior of clients we have to use the request of clients of the past. There are some predictors but their correctness's are not known. We can tell how good the predictors are by using similarity measures.

We examined which film to be downloaded from a topic by a client. This is the incipient problem. Our task is to compare the predicted lists with the perfect list. On the first place of list stand the most important client, who request the most films. Accordingly the last client of list will request the less. We would like to know how similar is the real client's list and the predicted list. Naturally the front of the list is more important than the end of list because these elements (clients) would like to request more films. Therefore the similarity in the first part of list is more relevant. This paper shows some methods to compare sorted list, which present the similarity between the whole two sorted lists.

We use orderly lists. Some caveats refer to the list. A list is not compared with another arbitrary list. Both lists contain the same elements and an element must not return in the same list. The human brain can simply decide subjectively the similarity between two lists, but more different harder decision. In general two people do not say the same similarity in case of two difficult lists. Our target is to find the best method. This problem seems to be not so difficult but it is. The number of elements of lists is constant. If this condition is not realized we will have to add more elements to the shorter list. The elements may be whatever type because we can number equal forms. Three new similarity measures were developed based the original measure. Therefore 7 measures were comparing in our examination.

Our test has two parts.

- In the first step we examine special sorted lists one at a time with measures. We lined up the given results so there are 7 (number of measures) new lists.
- In the second part these new lists were examined one more time with the measures. We made a sorted list from the special artificial lists by our opinion so it would be the perfect list in the second step.

Finally we examined the results from the world cup database. We can say the goodness of predictors with our special measures. Another conclusion is how the results change beside different parameters. Parameter may be the long of period and the number of tested periods.

2. DESCRIPTION OF SIMILARITY MEASURES

In this chapter we are going to introduce four similarity measures. Moreover it is possible we also introduce some modified versions of these measures.

1. *Kendall's tau*: Examines the order of elements in two ranked list. One at a time only the order of two elements examined.
2. *Spearman's footrule*: This measure analyses the distance of an element from its correct places of the original list.
3. *Ulam's distance*: It gives the similarity based on the longest ordered subsequence.
4. *Kemeny distance*: It counts how many pair of elements must be exchange to get the original list.

In every case we compare two lists. The similarity measures defined between these two lists. We also investigate the precision ($\in [0;1]$) of similarity measure to let them be comparable.

2.1 Kendall tau

Kendall's tau operates with the concept of the conflict between elements. There is a conflict between two elements when they are in different order in two ranked lists. Kendall's tau (N: non-conflict C: conflict between two elements).

Let C be the number of conflicts and N the number of non-conflicts in two lists. Since an element cannot be conflicted with itself we get $N + C = n \cdot n - n$. The similarity measure is defined:

The similarity measure [2]:

$$x = \frac{N - C}{N + C}$$

The precision of the measure [2]:

$$p = \frac{x + 1}{2} \quad .$$

2.2 Spearman's footrule

This measure examines the distance of an element of the first list from its correct place of another list.

It is defined as follow [2]:

$$x = \frac{1}{2} \sum_{i=1}^{n} \left| \pi_1(i) - \pi_2(i) \right|$$

where $\pi_j(i)$ (j = 1,2) means the place of the i element in the j. list.

The precision of measure [2]:

$$p = \frac{w - x}{w} \quad ,$$

where

$$w = n \cdot \left\lfloor \frac{n}{2} \right\rfloor - \left\lfloor \frac{n}{2} \right\rfloor^2$$

Spearman's footrule calculate the difference between the two ranks of an element in the two ordered lists. The sum of these differences gives the measure of similarity. The closer an element is situated to its correct places the less is the sum.

In some special cases the Spearman's footrule gives wrong result. Let's examine these cases through examples (see Table 5.1) the worst case for us is the reversed list (B20). In this case the precision of similarity measure is 0. However some list very similar to the original list results 0 measure.

For example:

B20 = 17, 16, 15, 14, 13, 12, 11, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1
B14 = 17, 16, 15, 14, 13, 12, 11, 10, 9, 1, 2, 3, 4, 5, 6, 7, 8
B15 = 9, 10, 11, 12, 13, 14, 15, 16, 17, 8, 7, 6, 5, 4, 3, 2, 1
B16 = 9, 10, 11, 12, 13, 14, 15, 16, 17, 1, 2, 3, 4, 5, 6, 7, 8

These are special cases, where by characteristics of Spearman's footrule the sum of the distances of the element from estimated and correct place results the same their value as in the worst case. That's why the similarity is 0. However the full reversed list is a worst case any worst case list can be build with two steps (Figure 2.1):

Algorithm

- Move the first m elements in reversed order to the end of the list. Based on this greedy method an element comes up to the furthest place.
- Move the remaining elements from $m+1$ to the end of the list to front of the new list without any change in order.

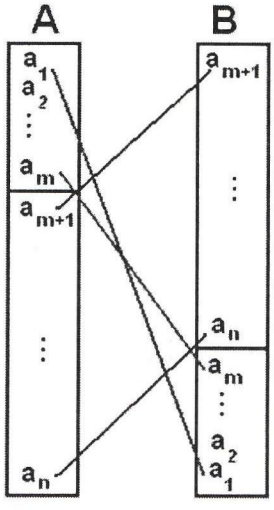


Figure 2.1 The worst list for the Spearman's footrule.

To get a worst case we should determine the place of an element where we divide the list into two separated lists. This place (m) can be calculated as:

$$m = \left\lfloor \frac{n+1}{2} \right\rfloor$$

where m that element where we have to divide the list into two part (see the Proof [8]).

2.3 Ulam's distance

Ulam's distance search subsequences in the predicted list in which the items are the same order then the good list. The elements need not to be neighborhood. The measure needs the number of the longest subsequent.

It is defined as follow [2]:

$$x = n - l$$

where l is the maximum number of items ranked in the same order in the two

The precision of the measure [2]:

$$p = \frac{n - x}{n}$$

The worst case for the Ulam's distance is the reserved list again.

2.4 Kemeny distance

It [3] calculates how many pair of elements must be exchange to get the original list. Given a list with an element the maximum number of exchanges is $n-1$ at most. (see example B7 in Table 5.1).

The precision defined as follow:

$$p = 1 - \frac{x}{n-1}$$

where x is the minimum number of exchanges to get the original list.

3. MODIFIED SIMILARITY MEASURES

We made some modifications on the already examined measures, we introduced the weights. The bigger the weight is, the closer the element is situated to the head of the good list concern to the first elements of good list.

3.1 Kendall's tau

We did not change the Kendall's tau because has not enough significant by the weight function.

3.2 Spearman's footrule

For this measure we defined the maximum weight to be the number of the elements. This way we distributed the precision of a list in the [0,1] interval. Other weights were also tried but there was no really change in the results. The used weight for the Spearman's footrule in Table 3.1.

weight	n	n-1	n-2	...	2	1
number	1.	2.	3.	...	n-1.	n.

Table 3.1 The used weight for the Spearman's footrule.

For the precision of a measure we need also worst case. We could consider using the original Spearman's footrule, that the worst case was not the same as the really worst case. The worst is B5 by Weighted Spearman's footrule.

B5 = 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 7, 6, 5, 4, 3, 2, 1

It is easy to see that the head of the list is not correct, however the major part of the list was good estimated. So there are similarities between B5 and the good list but the precision is 0. Unfortunately this is the Spearman's footrule's mistake. We can generate a worst list w the same way as we show earlier (see section 2.1). Because of weights the place of cut *(m)* changes. We need the w to count the precision and also m. Give w as follow (see the Proof [8]):

$$w = \frac{7}{6}m^3 - \left(\frac{5}{2}n + 1\right)m^2 + \left(\frac{3}{2}n^2 + n - \frac{1}{6}\right)m$$

We know the value of w. But we need where have we to divide the list. We have to compute the first derivation the previous function and find where it is equal to 0. We have to round again.

Give m as follow (see the Proof [8]):

$$m = \left\lfloor \frac{5n + 2 - \sqrt{\left(2n(2n+17) - \frac{5}{3}\right)}}{7} + \frac{1}{2} \right\rfloor$$

The similarity measure and the precision:

$$x' = \sum_{i=1}^{n} (n - i - 1)\left|\pi_1(i) - \pi_2(i)\right|$$

$$p' = \frac{w' - x'}{w'}$$

3.3 Ulam's distance

We weight the Ulam's distance too. This case the weight is 1/x because the x weight effected not enough change between the precisions (see Table 3.2).

weigth	1	$\frac{1}{2}$	...	$\frac{1}{n-1}$	$\frac{1}{n}$
element	1.	2.	...	n-1.	n.

Table 3.2: The elements of list with weight

The modified formula is:

$$n' = n \cdot \sum_{i=1}^{n} \frac{1}{i}$$

$$x' = n' - l \cdot sw$$

where l is the longest subsequence, sw is the sum of the weights of the elements in the list.

The precision:

$$p' = \frac{n' - x'}{n'}$$

3.4 Kemeny distance

It is easy to see, there is no meaning of using weights in Kemeny distance. The cause is similar than the Kendall's tau.

3.4 Complexity

The next table shows the complexity of all used similarity measures. There is here also the Complex method which will be presented in the next chapter.

	Time comlexity	Space complexity
Kendall	$\mathcal{O}(n^2)$	$\mathcal{O}(1)$
Spearman	$\mathcal{O}(n)$	$\mathcal{O}(1)$
W.Spearman	$\mathcal{O}(n)$	$\mathcal{O}(1)$
Ulam	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$
W.Ulam	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$
Kemeny	$\mathcal{O}(n)$	$\mathcal{O}(1)$
Complex	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$

Table 3.3 Time and space complexity for the measures.

4. COMPLEX METHOD

The above presented algorithms use only one property of the lists to decide their similarity. In the abstract we already mentioned that we need to use more complex. We used to consider more factors to decide as human.

These factors are:

- Is the order of elements correct? (In fact it is Kendall's tau)
- How far are the items of their correct place? (Spearman's footrule)
- How many exchanges need we to get the correct order? (Kemeny distance)
- How many pair of elements situated in the correct order side by side?

This method is defined as follow:

$$x = Ch \cdot (n - R) \cdot \sum_{i=1}^{n} S_i \cdot (n - i + 1) \cdot \left| \pi_1(i) - \pi_2(i) \right|$$

where *Ch* is the number of necessary exchanges, R is the number of neighboring pair of elements in correct order. The S_i (i=1,...,n) contains the number of conflicts of *i* th element, $\pi_j(i)$ (j = 1,2) means the place of the *i* element in the *j* list. We also need for the precision the worst case. We started from x form.

4.1 The precision of method

We defined the worst case with a computer program. The program counts all results of Complex method of all elements of n elements' permutation. (In our case n=11). The worst case is the list with the maximum value. In Figure 4.1 is the structure of the worst list for the Complex method. After we know the structure of worst list we can deduce the average worst case in formula.

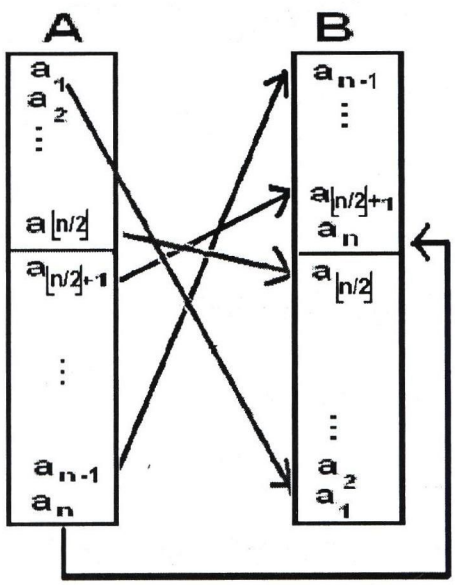


Figure 4.1 The worst list for the Complex method.

To compute the precision of this measure we need the worst case value. It gives as follow (see the Proof [8]):

$$w = \frac{1}{3}n^6 - n^4\left(\frac{m^2+m}{2} + \frac{11}{3}\right) + n^3\left(\frac{1}{3}m^3 + \frac{1}{2}m^2 + \frac{31}{6}m + 6\right) - n^2\left(\frac{m^3}{3} + \frac{m^2}{2} + \frac{43}{6}m + \frac{8}{3}\right) + n\left(\frac{m^2}{2} + \frac{5}{2}m\right)$$

The precision of the measure:

$$p = 1 - \frac{x}{w}$$

5. THE EXAMINED LISTS

In the Table 5.1 are the special ordered examined lists. The elements of the lists are $1,...,n \in N$. We got the list as follows. We divided at the m the list. The E is the begin and V is the end of the list.

- A : original list
- B1 : invariable list
- B2 : we get out every second element form the original list and this sublist we merge into its place reversed order.
- B3 : m = 7, E + reversed V
- B4 : m = 8, E + reversed V
- B5 : m = 7, V + reversed E
- B6 : m = 6, V + reversed E
- B7 : the last element go to the first place, and the other elements go back one position.
- B8 : m = 3, E + reversed V
- B9 : m = 4 (about 25% of 17), E + reversed V
- B10 : m = 5, E + reversed V
- B11 : m = 13 (about 75% of 17), E + reversed V
- B12 : m = 15 (about 90% of 17), E + reversed V
- B13 : m = 3, reversed V + E
- B14 : m = 8, reversed V + E
- B15 : m = 8, reversed V + reversed E
- B16 : m = 8, V + E
- B17 : m = 8, reversed E + V
- B18 : m = 8, reversed E + reversed V
- B19 : m = 9, we merged the second part of the original list the same order to the first part of the list.
- B20 : reversed list

A	=	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
B1	=	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
B2	=	1	17	3	16	5	14	7	12	9	10	11	8	13	6	15	4	2
B3	=	1	2	3	4	5	6	7	17	16	15	14	13	12	11	10	9	8
B4	=	1	2	3	4	5	6	7	8	17	16	15	14	13	12	11	10	9
B5	=	8	9	10	11	12	13	14	15	16	17	7	6	5	4	3	2	1
B6	=	7	8	9	10	11	12	13	14	15	16	17	6	5	4	3	2	1
B7	=	17	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
B8	=	1	2	3	17	16	15	14	13	12	11	10	9	8	7	6	5	4
B9	=	1	2	3	4	17	16	15	14	13	12	11	10	9	8	7	6	5
B10	=	1	2	3	4	5	17	16	15	14	13	12	11	10	9	8	7	6
B11	=	1	2	3	4	5	6	7	8	9	10	11	12	13	17	16	15	14
B12	=	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	17	16
B13	=	17	16	15	14	13	12	11	10	9	8	7	6	5	4	1	2	3
B14	=	17	16	15	14	13	12	11	10	9	1	2	3	4	5	6	7	8
B15	=	9	10	11	12	13	14	15	16	17	8	7	6	5	4	3	2	1
B16	=	9	10	11	12	13	14	15	16	17	1	2	3	4	5	6	7	8
B17	=	8	7	6	5	4	3	2	1	9	10	11	12	13	14	15	16	17
B18	=	8	7	6	5	4	3	2	1	17	16	15	14	13	12	11	10	9
B19	=	1	10	2	11	3	12	4	13	5	14	6	15	7	16	8	17	9
B20	=	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Table 5.1 The examined lists.

6. COMPARISON OF MEASURES

Using weights on Kendall's tau do not give appropriate results because the elements are already weighted with the number of elements in correct order.

In the other two cases the weights worked like we expected. These measures find more similarity with a list in which the front elements are correct then with a list in which the second half is in correct order. In the case of Spearman's footrule the precisions are more drawn aside. Unfortunately it finds little similarity with the reverse list, the problem comes from the failure of the original measure.

In the second step we built a list from the artificial lists based in our opinion. The most similarity lists stand on the first place. In the future this list name is **R.** We ranked the precisions of the measure. From the result of Kendall's tau was built the Kendall list, from result of Spearman's footrule was built the Spearman list etc. Following we used the measures in the new lists once more.

6.1 Analysis of results

The final compare is in Table 6.1. We got these results after we used similarity measures. In the graph (Figure 6.2) based on the results of Figure 6.1 we can see well which measure things which measure to be the best.

	Kendall	Spearman	W.Sp.	Ulam	W.Ulam	Kemeny	Comp.
Kendall l.	0,8790	0,8000	0,8142	0,7000	0,5944	**0,6316**	0,9889
Spearman l.	0,8842	0,8000	0,8407	0,7000	0,6041	0,2632	0,9813
W.Sp. l.	0,8790	0,8000	**0,8864**	0,6500	0,5578	0,5263	**0,9944**
Ulam l.	0,7105	0,5200	0,5988	0,5000	0,3060	0,3158	0,9011
W.Ulam l.	0,8316	0,7500	0,8046	0,6500	0,5012	0,2632	0,9623
Kemeny l.	0,7368	0,6900	0,7628	0,5500	0,4228	0,4737	0,9410
Complex l.	**0,9053**	**0,8500**	0,8860	**0,7500**	**0,6587**	0,4737	0,9914

Table 6.1: The precisions of the measures for the lists which ranked by the measures. See the related diagram in Figure

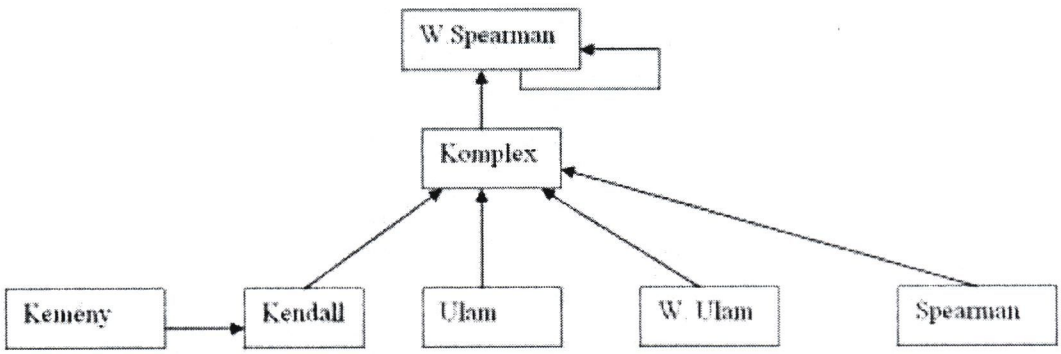


Figure 6.1: The graph of the best similarity measures shows which is the best measure.

- *Kendall's tau*: In our opinion this stands the third place on the ranking of similarity measure. The Complex is the best by Kendall's tau. But we can not trust in this result. According to transitivity this measure expect W.Spearman's footrule to be the best.
- *Spearman's footrule*: It shows Complex method is the best and also its modified version W.Spearman's footrule. We can be sure the W.Spearman's footrule gives more reliable result than the original Spearman.
- *Weighted Spearman's footrule*: This is on the top of the graph because the measures undirectly show this and Complex method directly votes it. This measure shows immodestly itself. In our opinion this is the most suitable similarity measure for our problem.
- *Ulam's distance*: It recommends Complex method but it is also not enough reliable. The cause of unreliablement is that it gives little different result like Kemeny. Therefore we won't use the results of this measure.
- *Weighted Ulam's distance*: It gives more different values than the Ulam's distance. But it's based on the original version. We also won't use it henceforward.

- *Kemeny distance*: It has other opinion than another, shows the Kendall's tau as the best measure. It gives a lot of same precision value (see Figure 6.3) so the results are also not enough reliable.
- *Complex measure*: It says the W.Spearman's footrule is the best but different of results between the Complex and W.Spearman's footrule is insignificant. In our rank it is the second best.

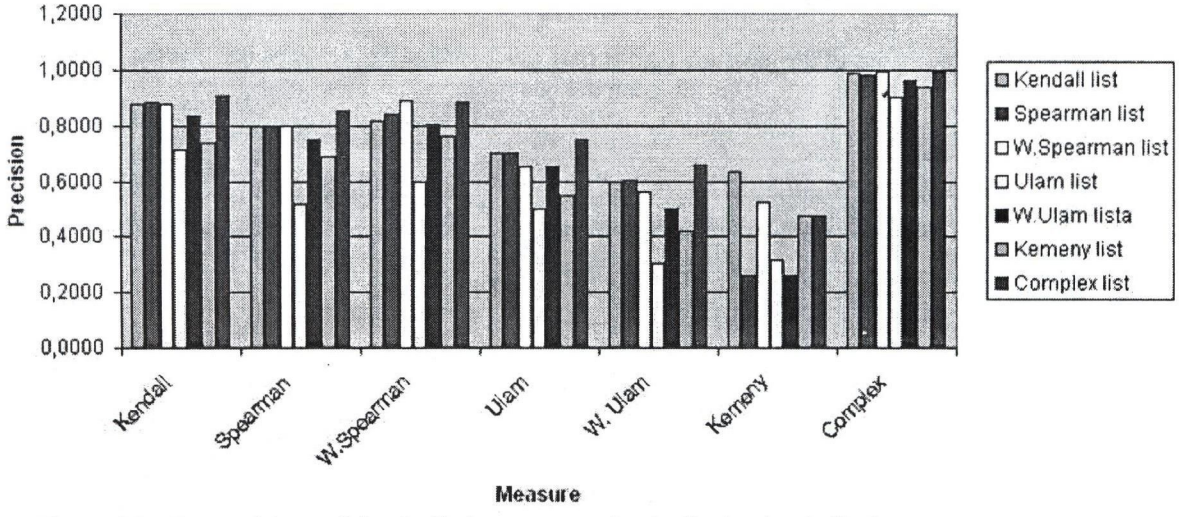
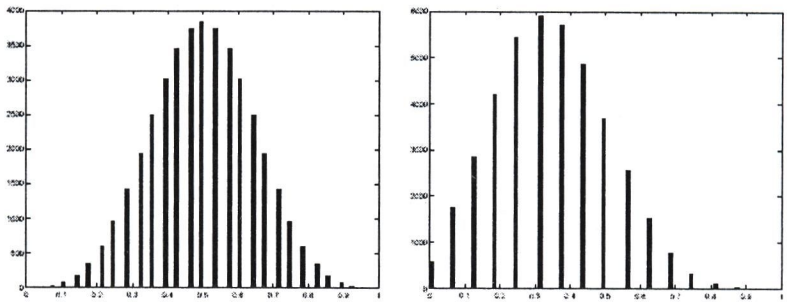


Figure 6.2: The precisions of the similarity measures for the list by the similarity measures.

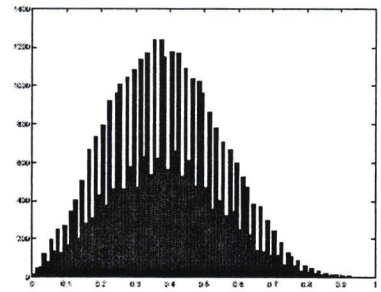
Finally the best is W.Spearman's footrule but we also use the Complex method and the Kendall's tau in the further work. The different between of the Spearman's footrule and the Complex method is very little, so we can use these measures. Because of the W.Spearman's footrule is the top of the graph we consider this is the best similarity measure.

The distribution of the precision of measures is in Figure 6.3. We applied the measures for lists with eight elements. We accorded the density functions of measures for lists (all 8-elements permutation). We can see in which range are the result of measures. Kendall's tau and the Spearman's footrule have close normal distribution. But the means of latter one isn't 0.5. The Weighted Spearman's footrule is more continuous than the simple version. In case of Ulam's distance and the Kemeny distance we also can discover the closed normal distribution, but they have little several results which stand far from each other. Weighted Ulam's distance is the odd one out because its distribution is not like bell-curve. The distribution of the precision of the Complex method is the most continuous, because it has the most results of precision. It is also closed normal distribution but the mean is about 0.8.

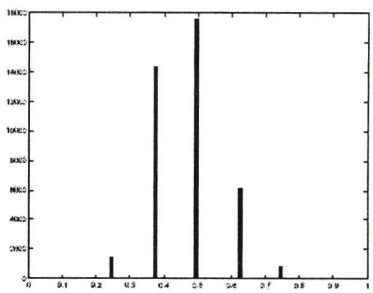


Kendall's tau

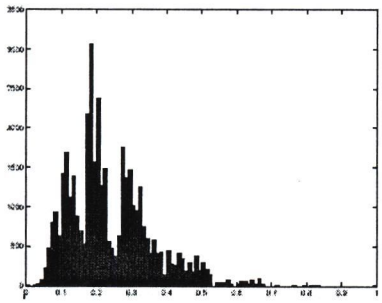
Spearman's footrule



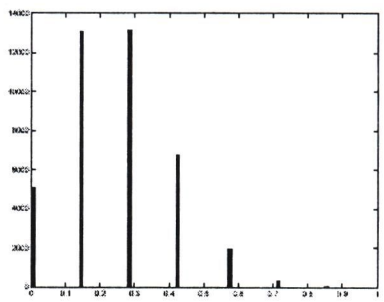
W. Spearman's footrule



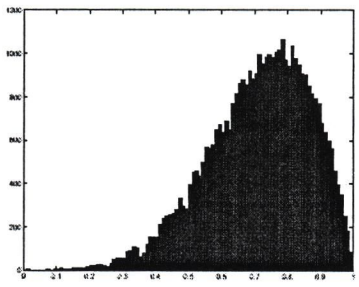
Ulam's distance



W. Ulam's distance



Kemeny distance



Complex method

Figure 6.3: The density functions of the measures for 8-elements permutations.

7. ACCURACY OF PREDICTORS BY WORLD CUP 98 DATABASE

We can say the accuracy of predictors by using similarity measures. We compared the result of predictors. We could order the predictors by goodness, testing them on a real database (the World Cup '98 Website's access log[7]). The amount of the downloaded

bytes gives the order of the clients. The judgment of the rank of the list is not subjective, because it is a real scenario. In this examination we used only Kendall's tau Weighted Spearman's footrule and the Complex method.

The predictors used the data of 91 days of World Cup database. They predicted download of several clients by the earlier download. Ordering the predicted values we get the lists. We examined all of 6 period type [2]. One period may contain 1, 4 or 7 days and the predictors may use also more periods. The best predictors are LPC, LastValueM and the BinomM0.75 by the results. The precisions of these are the most significant. These results do not depend from similarity measure or periods type. Furthermore the value of the BinomApp0.75 and the ANN-0.2 are also high.

The other paper [2] agree with us in LPC and the first places but suggested more other predictors. The cause of difference is that they used the average of three original similarity measures (Kendall's tau, Spearman's footrule and the Ulam's distance) while we used W.Spearman's footrule, Complex method and Kendall's tau. Weighted Spearman's footrule gives the finally order of predictors because it was the best measure by the results. So this measure can show how good the predictors are.

Probably we can examine more aspect, for example which period length is the best, but we did not examine this.

8. FUTURE WORK

In this paper you could see a real application of the similarity measures. We could examine in the future which period length fits the best to the predictions. We should look for other databases which suit better to our problem because this database is not a log file of a video server but of a web server. If we use another predictors, we will can say the new predictors' goodness with these measures, but we will have to use only these three best methods.

REFERENCES

[1] R. Tusch, (2004) *Design and implementation of an adaptive distributed multimedia streaming server*, Phd Thesis,University Klagenfurt, Information Technology Department,April

[2] Péter Kárpáti, András Kocsor, László Böszörményi, *Client Behaviour Prediction in a Proactive Video Server* Technical Reports of the Institute of Information Technology, University Klagenfurt

[3] Vu Ha, Peter Haddawy, (2003) *Similarity of Personal Preferences*: Theoretical Foundations and Empirical Analysis Elsevier Science

[4] M. G. Kendall (1962). *Rank Correlation Method.* Hafner Publishing Company, New York, Third Edition.

[5] C. Spearman. (1906). *Footrule for measuring correlation.* British Journal of Psychology, 2:89-108, June 6.

[6] S. M. Ulam. (1981) *Future applications of mathematics in the natural sciences.* American Mathematical Heritages: Algebra and Applied Mathematics. Texas Tech. University, Mathematics Series,13:101-114,

[7] URL: http://ita.ee.lbl.gov/html/contrib/WorldCup.html

[8] Cs. Domokos, E. Széll, P. Kárpáti, L. Böszörményi: *Mixed and Weighted Measures for Client Behavior Prediction in a Proactive Video Server*

[9] http://143.205.180.128/ITEC/Publications/pubfiles/pdffiles/2005-0172-CDES.pdf

DIGITAL LIBRARIES FOR PRESENTATION AND PRESERVATION OF EAST-CHRISTIAN HERITAGE

Desislava Paneva
Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, Sofia, Bulgaria
dessi@cc.bas.bg

Lilia Pavlova-Draganova
Laboratory of Telematics – Bulgarian Academy of Sciences, Sofia, Bulgaria
lilia@cc.bas.bg

Lubomil Draganov
Institute of Mathematics and Informatics – Bulgarian Academy of Sciences, Sofia, Bulgaria
lubo@cc.bas.bg

Abstract

This article aims to present the digital libraries with multimedia content as a technology for innovative presentation of cultural and historical values. It includes some basic concepts of digital libraries with multimedia content and a description of three types of architecture. Finally, it describes the ideas, conceptual decisions and strategies for implementation of digital libraries with multimedia content in two Bulgarian projects for presentation of East-Christian Icon art and cultural heritage in the global information space.

Keywords

Digital libraries with multimedia content, digital library architectures, cultural heritage, East-Christian Icon art.

1. INTRODUCTION

Preserving the cultural, historical and scientific heritage of various world nations, and their thorough presentation is a long-term commitment of scholars and researchers working in many areas. From centuries every generation is aimed at keeping record about its labour, so that it could be revised and studied by the next generations. New information and multimedia technologies have been developed during the past couple of years, which introduced new methods of preservation, maintenance and distribution of the huge amounts of collected material. There are various conceptual and technically feasible solutions available, such as digitalization of cultural and historical artefacts and creation of multimedia information archives, web presentations of valuable artefacts in virtual museums, galleries and digital libraries, 3D virtual realities, which present places of culture and history, digital modelling and simulation, aiding the conservation, restoration, storing and showing artefacts, etc.

This article aims to present digital libraries with multimedia content as a modern technological solution for innovative presentation of the variety of Bulgarian Icon art and cultural heritage (churches, monasteries, murals, etc.) from different artists, historical periods, and schools. The contemporary digital libraries have been made possible via the integration and use of a number of information and communication technologies, the availability of digital content on a global scale and the strong demand for users who are now online. It is expected that they should enable any citizen to access human knowledge

any time and anywhere, in a friendly, multi-modal, efficient, and effective way. A core requirement for such digital libraries is that they have a common infrastructure which is highly scalable, customizable and adaptive. This article describes three types of architecture with different complexity. Considering the specific needs and requirements of different cultural and historical heritage projects some of these architectures could be chosen. The last part of this article describes the ideas, conceptual decisions and strategies for implementation of digital libraries with multimedia content in two Bulgarian projects for presentation of East-Christian values in the global information space.

2. BASIC CONCEPTS OF DIGITAL LIBRARIES WITH MULTIMEDIA CONTENT

Digital libraries with multimedia content are a contemporary conceptual solution for access to information archives. According to an informal definition of digital libraries, they are managed collections of information, with associated services, where the information is stored in digital formats and accessible over a network. Digital libraries contain diverse hypertext-organized collections of information (digital objects such as text, images, and media objects) for use by many different users. The collected information is organized thematically and uses hyperlinks that allow the connection between any piece of data and additional data on the same topic. As an addition to the digital objects collection, there are many levels of metadata, indexes, hierarchical links, etc. [3]

The main characteristics of digital libraries are the following:

- Ability to share information;
- New forms and formats for information presentation;
- Easy information update;
- Accessibility from anywhere, at any time;
- Services available for searching, selecting, grouping and presenting digital information, extracted from a number of locations. Using these services depends on the user preferences, needs and wishes of the users, i.e. there is personalization available;
- Contemporary methods and tools for digital information protection and preservation;
- Ability to use different types of computer equipment and software;
- No limitations related to the size of content to be presented.

In the past digital libraries were isolated and monolithic systems limited to access to content of a single provider. The development of the technologies during the last years provides new functionalities and advanced services to contemporary digital libraries such as specialized services for

- Multi-layer and personalized search, context-based search, relevance feedback, etc.
- Resource and collection management;
- Metadata management;
- Indexing;
- Semantic annotation of digital resources and collection etc.

The new digital libraries will provide and manage complex services, processes and workflows on the basis of existing services. It is expected that these services be heterogeneous, autonomous and distributed. The flexibility, the automatic adaptation, the access anywhere and anytime, the decentralization, the wide variety of digital objects and collections, the information security, etc. will be of the some requirements. [1] [2]

Digital library architectures

A **Hypermedia digital library** can be considered as a database, storing data of different type (text, raster, vector, static and moving (video) images, animation, audio or other media), which is structured in a way to allow easy manipulation and use. Data is stored in the database in the form of objects, usually annotated to facilitate running search queries. To make these procedures automatic, the hypermedia library includes techniques for descriptive presentation of the data semantics as well as services for its management.
Web technologies help organizing hypermedia digital libraries by providing a means to structure and present them in a hypermedia manner. Hypermedia represents hypertext media; therefore it adheres to the hypertext information organization rules. Users are allowed to quickly move across subject-related topics in a non-linear way. These topics may include sets of objects, such as text, images, audio and other media, which relate to one another via hyperlinks.

The Hypermedia digital library in the Web space is a simplified conceptual solution for presenting complex multimedia content.

Grid-based infrastructures - The digital library is currently undergoing a transition from a statically integrated system to a dynamic federation of services. This transition is inspired by new trends in technology which include developments in technologies like Web services and grid infrastructures as well as by the success of new paradigms like Peer-to-Peer Networking and Service-oriented Architectures. The transition is driven by digital library "market" needs. This includes a requirement for a better and adaptive tailoring of the content and service offer of a digital library to the needs of the relevant community as well as to the current service and content offer, and a more systematic exploitation of existing resources like information collections, metadata collections, services, and computational resources.

Such new decentralized and service-oriented architectures for digital libraries make the library functionality available in a more cost-effective and tailored way and thus open up new application areas for digital libraries. Future digital libraries should enable any citizen to access human knowledge any time and anywhere, in a friendly, multi-modal, efficient, and effective way. A core requirement for such digital libraries is a common infrastructure which is highly scalable, customizable and adaptive.

A grid is a network or collection of distributed computer resources, which are accessible through local or global networks and are presented to the end user via an enormous virtual computer system, i.e. it is a virtual, dynamically changing organization of structured resources, which are shared among individuals, institutions and systems. Some of the main

advantages of the grid technology are: optimized and personalized access and enhanced management of digital resources; virtual resource organization; ability to be used worldwide, etc. The grid technology introduces essential improvements in the current distributed information systems, which are the proper basis for building contemporary digital libraries.

In essence, the creation of virtual digital libraries on the basis of grid-based infrastructures, support for the integration of metadata, personalization services, semantic annotation and the on-demand availability of information collections and extraction services will make digital libraries more useful and attractive to a wider clientele. Such a test-bed digital library infrastructure has been created for the DILIGENT project (Integrated project funded in part by the European Commission FP6 IST Programme), based on the grid technology [7]. Figure 1 depicts DILIGENT infrastructure.

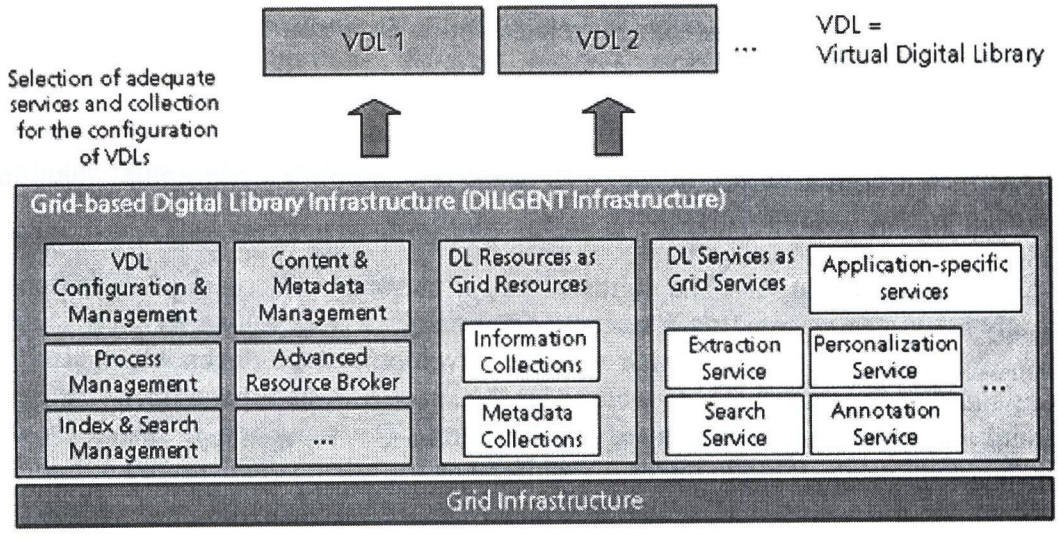


Figure 1: Grid-based digital library infrastructure

Hyperdatabase infrastructure - Future digital libraries should enable any citizen to access human knowledge any time and anywhere, in a friendly, multi-modal, efficient, and effective way. A core requirement for such digital libraries is a common infrastructure which is highly scalable, customizable and adaptive. Ideally, the infrastructure combines concepts and techniques from peer-to-peer data management, grid computing middleware, and service-oriented architectures. That infrastructure is offered in the project DELOS "A Network of Excellence on Digital Libraries" funded by the EU's Sixth Framework Programme. [5]

Peer-to-peer networks allow for loosely coupled integration of digital library services and the sharing of information such as recommendations and annotations. Grid computing middleware supports the dynamic allocation and deployment of complex and computationally intensive digital library services such as the extraction of features from

multimedia documents to support content-based similarity search. A service-oriented architecture provides common mechanisms to describe the semantics and usage of digital library services. Furthermore, it supports mechanisms to combine services into workflow processes for sophisticated search and maintenance of dependencies. As depicted in Figure 2, the digital library architecture envisaged consists of a grid of peers which provide various kinds of digital library services such as storage, extraction or retrieval services. These digital library services can be combined with processes. High scalability is achieved by executing the processes in a completely distributed, peer-to-peer fashion. For that, metadata about processes, services, and load of the peers is distributed and replicated over the grid. This is performed by a small hyperdatabase layer atop each peer. This layer also takes care of peer-to-peer navigation and execution of processes. Figure 2 depicts the execution of the process "Insert Image".

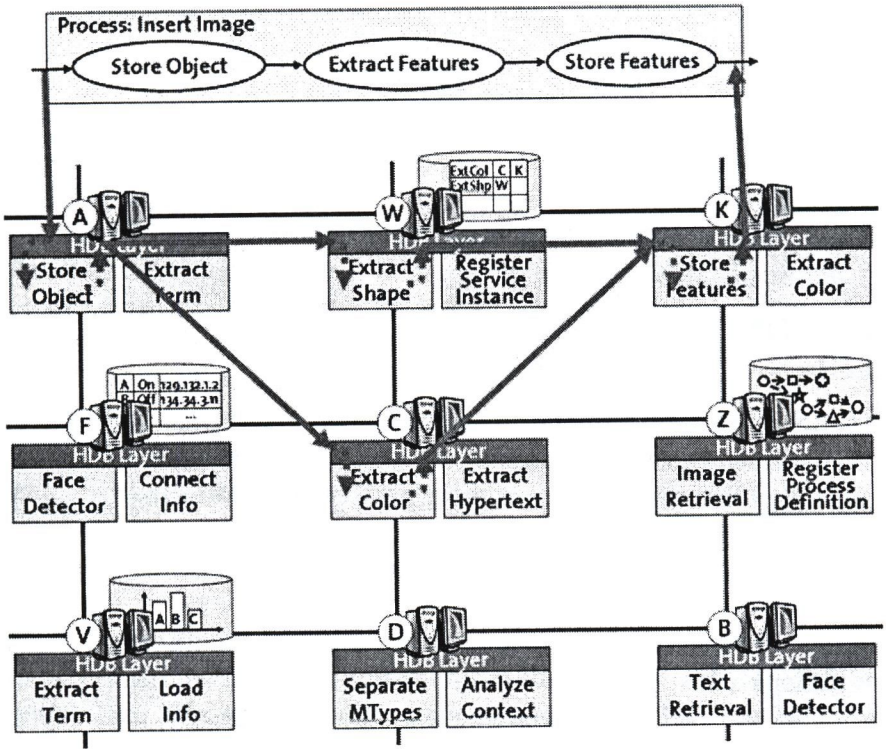


Figure 2: The execution of the process "Insert Image" in digital library architecture based on a hyperdatabase infrastructure

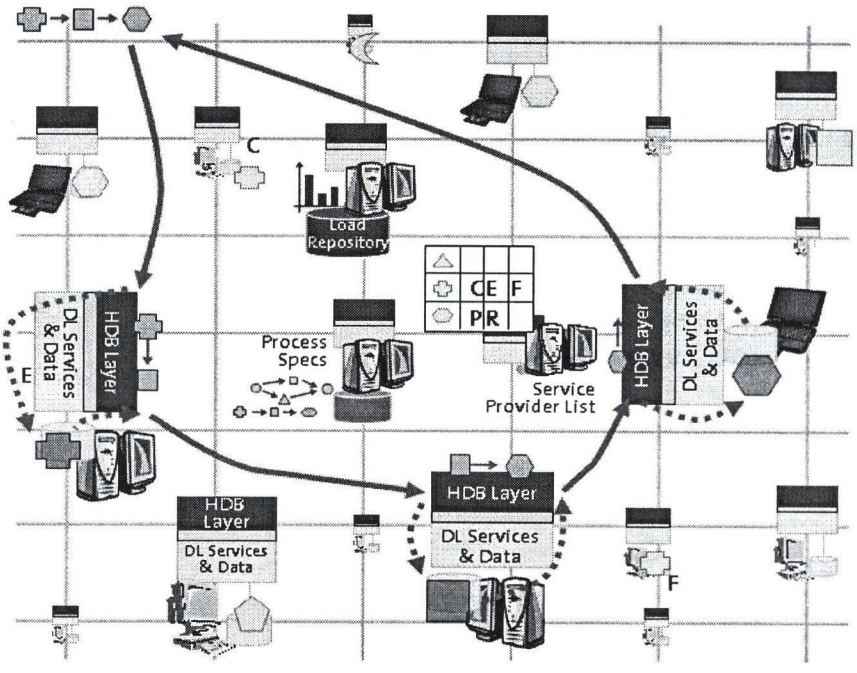


Figure 3: Digital library architecture based on a hyperdatabase infrastructure

3. DIGITIZATION OF EUROPEAN'S CULTURAL AND HISTORICAL HERITAGE

Europe's cultural, historical and scientific knowledge resources are a unique public asset forming the collective and evolving memory of our diverse societies. Resource discovery, accessibility, usability, interoperability, authenticity, quality and trust by all users of the Information Society are essential requirements for the delivery of digital cultural information and services. [4]

European libraries, archives and museums contain a wealth of information, representing the richness of Europe's history, its cultural diversity and its scientific achievements. The degree of access to this information determines how far people can experience their cultural heritage and benefit from it in their work or studies. By digitising their collections and making them available online, libraries, archives and museums can reach out to the citizens and make it easier for them to access material from the past. The online presence of this material from different cultures and in different languages will make it easier for citizens to appreciate their own culture heritage as well as the heritage of other European countries, and use it for study, work or leisure. [6]

The "i2010: Digital Libraries" initiative aims at making European information resources easier and more interesting to use in an on-line environment. The Commission adopted on 30/09/2005 the "i2010: Digital Libraries" communication outlining the vision of this initiative and addressing in particular the issues of digitisation, on-line accessibility and digital preservation of our cultural heritage. [9]

4. DIGITAL LIBRARIES OF EAST-CHRISTIAN ICON ART

East-Christian icon art is recognised as one of the most significant areas of the art of painting. Regrettably, it is still being neglected in the digital documentation and the registry of the art of painting. The accessibility to that large part of mankind's cultural and historical ancestry would be enhanced greatly if icons of all possible kinds and origins were digitised, classified, and "exhibited" in the Internet. That would allow the preservation and even the future digital restoration of a large number of rare specimens of the East-Christian art of painting. The need for a wide accessibility and popularisation is even bigger for the Bulgarian icons. Therefore, it is necessary that their idiosyncratic art and exceptional values be made available in the global information medium, so that they become accessible to both professional researchers and the wide audience.

The goal of the project "Virtual encyclopaedia of the art of the Bulgarian icon" is to develop the information content, structure, and the realisation of a digital multimedia library as a demonstrator of virtual encyclopaedia of the Bulgarian iconography". That library will include several hundred specimens of Bulgarian icons from different artists, historical periods, and schools. The chosen architecture represents a web-based hypermedia digital library, which means that presentation of a complex multimedia content in the Internet is simplified. The resources are digital objects of different formats – text, graphics, and other media. They will be structured in a hypermedia way, i.e., some digital objects point to other ones. In this way the user can navigate quickly, in a non-linear fashion, within areas of related topics, using the hyperlinks. The digital objects will be grouped according to their topics into thematic collections. For each object and collection, special meta-descriptions will be created. They will include data about the artist, the period, the school, the location, the material used, the category. Also, they will contain links to other digital objects and collections, keywords, and so on. That information will be used for the semantic annotation and indexing of the digital objects, which will facilitate their locating during search requests, and their web-based representation. A multitude of specialized services for metadata management, content management, indexing, metadata annotation of digital objects and collections, creation of requested document, different media types search, context-based search, multi-object, multi-feature search, etc, will be presented. Figure 4 depicts the architecture of hypermedia digital library in Web. The organisation of the media databases, the representation and description of the digital objects, and the classification of the artefacts, will be developed according to the recommendations of the international group of museum experts of East-Christian Art (UNESCO/I.DB.I) and the standards of CIDOC/MICMO. The project relies on the idea that the unity of the text information and the high quality of the digital images will represented the virtues of the Bulgarian icon in their entirety and will contribute to the its preservation, wider exhibition, and future potential restoration. The demonstrator that will be developed will be a tool for exploration of the techniques, styles, colours, and forms, as well as for the tracking and comparing of specimens and periods of the iconography and historical development of that art. The project will lay the foundations of the registration, documentation, and the exploration of a practically unlimited number of Bulgarian icons. The tools of the virtual encyclopaedia will give the users the opportunity to compare the icons in their historic context, so that some yet undiscovered treasured of the East-Christian iconography be

manifested.

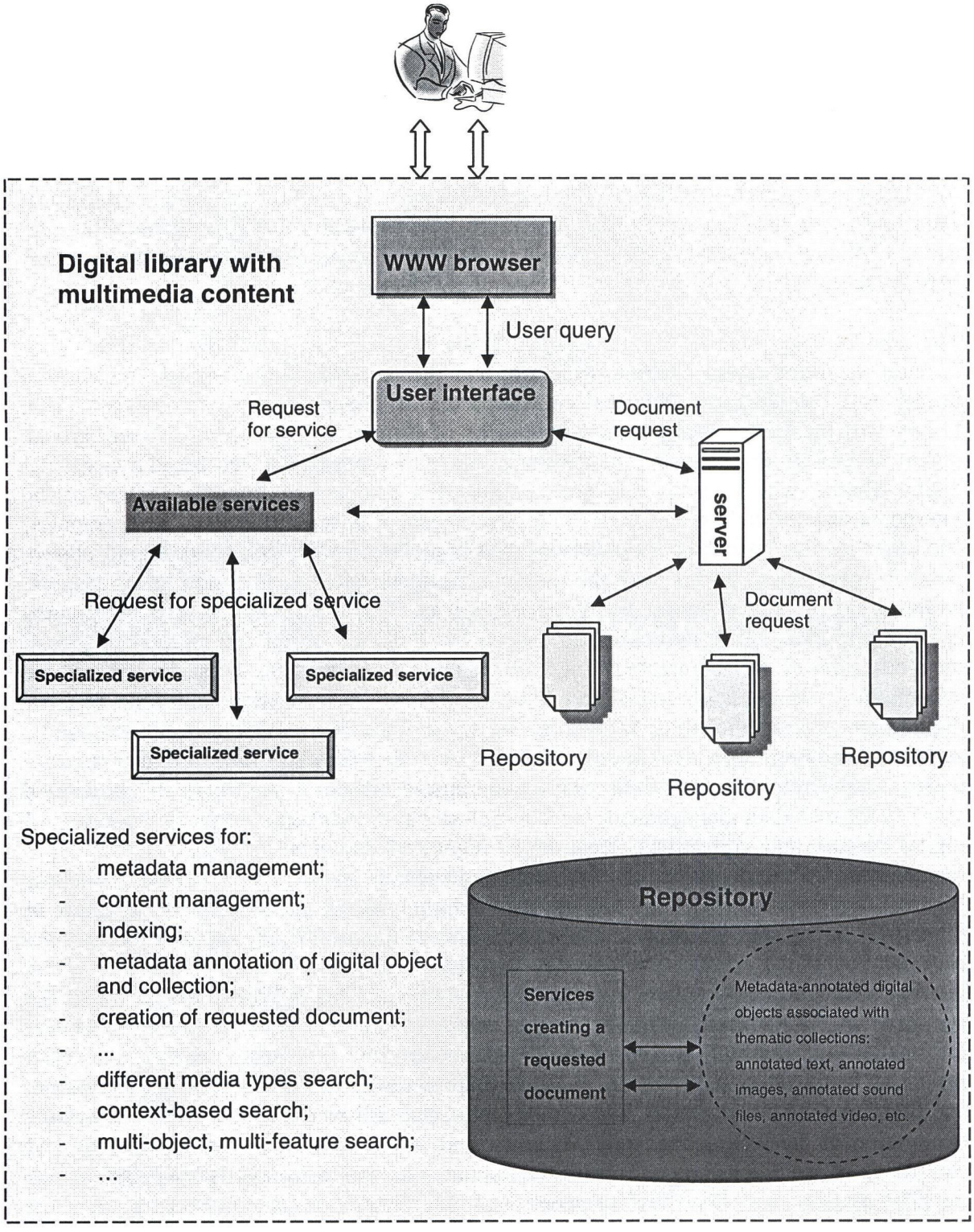


Figure 4: Hypermedia digital library

The project „Virtual encyclopaedia of the art of the Bulgarian icon" is a superstructure and further development of the European (FP5) project I.DB.I "International Database on East Christian Icon Art: Access to the World of Icon Art" [8]. A lot of work is done on that project for the creation of a high quality interface to a developed multimedia database of several hundred pieces of art of different authors, periods, and schools of the East-Christian iconography. That group includes painted icons and icons built with mosaics that are located in European museums, churches, monasteries, and private collections. The user interface offers novel tools and techniques for navigation, browsing, searching, and retrieval of digital representations of icons of the East-Christian iconography.

A project that has a similar subject is "Bulgaria SACRA". Its goal is to present the top achievements of the religious cultural-historical heritage of Bulgaria, too. However, its emphasis is on the architecture, the monuments, murals, etc. The technology of choice is yet again digital library with multimedia content, though in this case it aims at a complete description and multimedia representation of complex objects and their constituents, rather than stand-alone artefacts. For instance, an object-church with a general 2D and/or 3D view, a general and exterior architectural blueprint, a general interior blueprint, murals, iconostasis, icons, and so on. In this way we create a complete description of the chosen specimens. A special ontology describing completely the physical object of consideration and its constituents, their attributes, connections and relations between them, will be developed. We plan to accomplish a complex multi-object and multi-feature searching based on the semantic annotation and indexing of the complex objects. The final product of this project is planned to be an information artery for virtual representation of Bulgaria SACRA in the European and the global information space using the technologies of the digital multimedia libraries.

REFERENCES

[1] K. Kiernan, A. Kekhtyar (2003). EPT: Edition Production Technology for Multimedia Contents in Digital Libraries, Presented on Workshop on Multimedia Contents in Digital Libraries, USA

[2] IEEE Multimedia (2000), Virtual Heritage, April-June, Vol.7, No.2

[3] D. Krastev (2005). Central Library of Bulgarian Academy of Sciences – present and future, The libraries of Bulgarian Academy of Sciences, Reference book, Bulgaria

[4] eEurope and Digitisation - http://www.cordis.lu/ist/digicult/eeurope.htm

[5] Project DELOS: A Network of Excellence on Digital Libraries - http://www.delos.info/

[6] Project MINERVAEUROPE: Ministerial Network for Valorising Activities in digitalisation - http://www.minervaeurope.org

[7] Project DILIGENT: Digital Library Infrastructure on Grid Enable Technology - http://www.diligentproject.org/

[8] Project I.DB.I: International Database on East Christian Icon Art: Access to the World of Icon Art (2000). RTD Proposal: Description of scientific/technological objectives and workplan (Part B)

[9] "i2010: Digital Libraries", COM (2005) 465 final, Brussels, 30/09/2005

Analysis of Trust in Electronic Markets

Sviatoslav Braynov[†‡] Radoslav Pavlov[†]

[†] Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Bl. 8 Acad. G. Bonchev Str.
Sofia, Bulgaria

[‡] Department of Computer Science
University of Illinois at Springfield
Springfield, IL 62703, USA

Abstract

Trust is important for electronic markets due to the intrinsic uncertainty and risk associated with electronic transactions. The paper presents a game-theoretic and a decision-theoretic models of trust. The game-theoretic model is applied to the problem of learning trust through repeated interactions. We show that it is not beneficial for a person always to honor trust. Instead, it could be better to alternate between honoring and abusing trust so that to keep one's trustworthiness above a certain threshold.

The paper analyzes the impact of trust on electronic market efficiency. It is shown that if economic agents hold accurate trust estimates about one another, then the social welfare, the amount of trade and the agents' utility levels are maximized. We also show that market efficiency does not require complete trustworthiness.

The paper discusses various mechanisms that make agents truthfully reveal their level of trustworthiness before the beginning of every transaction. Honest reporting at the first stage of interaction informs other agents about possible risks and helps them form realistic expectations about possible outcomes.

1 Introduction

The concept of trust is essential to transactions in complex and dynamic systems with a high degree of interdependence. In such systems, the outcomes of an action depend not only on a single actor, but also on the actions of a number of external actors and chance events. Interdependencies often introduce an element of risk of failure, damage, loss,

inadvertent or malicious behavior, etc. Agents can fail to perform their tasks or to meet their commitments due to lack of incentives, lack of ability, or circumstances beyond their control. For example, a buyer may have to pay for a product before it is delivered and it may not be certain whether the product will be delivered on time. Problems could arise from events outside the control of the buyer. For instance, the seller may deliver a product of inferior quality, the product may get delayed because of transportation problems, or somebody can damage the product during transportation. In general, the actors affecting the final outcome of an action may be totaly unknown, their intentions and incentives could be difficult to predict, and their skills and knowledge could be difficult to envisage.

This naturally leads to the idea of trust. Everything that is impossible or costly to secure needs to be trusted. The concept of trust has been a subject of continuous interest in different research areas, including computer security [1], multiagent systems [2, 3, 4, 5], sociology [6], risk management [7, 8], economics and game-theory [9, 10, 11, 12].

Trust is usually considered a belief that an entity will perform in a favorable way. In other words, an actor takes the risk to depend on an entity which has partial or full control over a situation. Whether the outcome of the situation is favorable to the actor depends on the entity. For example, when a buyer is buying a second hand car, the buyer might need to trust the car dealer to sell him a good car. It is usually the case that the dealer has better information than the buyer about the quality and the condition of the car. If the buyer believes that the dealer is likely to withhold important information about the car, the buyer would certainly choose another car dealer.

Reputation is another concept [10] that is close to the notion of trust. It is worth noting, however, that the concept of reputation is more general than trust. An agent may have the reputation of being trustworthy, honest, aggressive, tough, etc. Therefore, the notion of reputation involves establishing and maintaining some individual characteristics which are publicly observable. Trust, on the other hand, could be based on private information. For example, an agent may be trustworthy without having any public reputation. Despite the difference between them, both trust and reputation can be used as a capital asset. Economic research [13] reveals that in their economic activity firms tend to convert their financial capital into reputational capital and vice versa. That is, in order to increase their reputation firms may be willing to invest some financial capital, or in order to increase their financial capital firms may be willing to sacrifice their reputation.

Reputation systems [14] have long been used as a means for measuring trustworthiness. Their use, however, is limited by several factors:

- In many cases, obtaining reputation rankings may be impossible. Internet provides vast opportunities to interact with total strangers for whom there is no history of previous transactions.
- Information obtained from reputation databases or recommender systems is usually too general to be applied to the context of a particular transaction.
- The coverage of current reputation systems is limited.
- There are problems with the aggregation of correlated reputation [15].

- Possibility for fake transactions [16].
- False identity and pseudonyms [17].
- Obtaining negative feedback from unsatisfied customers could be difficult [14]. Trust, on the other hand, could be based on private information.

2 Trust in electronic markets

The concept of trust is important to e-commerce because it affects the very essence of on-line business: the possibility to engage in a risky transaction. Internet users still fear the possibility of fraud, misuse of private information, identity change, deception, etc.

Need of trust in business transactions is usually explained by time asymmetry, lack of power, or inability to conclude perfect contracts. The *time asymmetry* argument draws on the fact that transactions are usually performed over a period of time and the actions of some agents temporally precede those of others. In other words, there is a time lag between the placement of trust and the corresponding action of the trusted party who can honor or abuse trust. For example, the delivery of goods and services by one party might occur only before the other party has made the payment. In the absence of appropriate control mechanisms, the party that acts first has to trust the other agents for fulfilling their part of the transaction.

The *contract argument* addresses the possibility of concluding definite and complete contracts between cooperating agents. The effect of legal contracts is at least threefold:

- It decreases the incentive for the trustee to abuse trust by providing appropriate sanctions.
- It compensates the trustor if trust is abused.
- It increases the trustor's expectations that the trustee will behave favorably.

Legal and economic experience [18], however, indicate that contracts usually are incomplete. Several factors contribute to contract incompleteness. The following are the most common. First, a contract may be ambiguous because the words explaining an issue are ambiguous. Second, the contract parties may fail to reach agreement about some issues, yet prefer to make a contract on the issue on which they agree. Third, the contract may have been left incomplete because the cost for the contract parties of drafting an issue may exceed the expected benefit. Finally, there may be asymmetric information about an issue because events in the world might not be mutually observable.

The *power argument* for placing trust is as follows. When an agent does not have the power to control actions of other agents (including nature) or when exerting the power is too costly for him and when other agents' actions have a bearing on his behavior or welfare, then he may be willing to place trust on the agents he cannot control. In game theory the notion of power is formalized by Harsanyi [19] The impact of power on multiagent planning is studied by Brainov and Sandholm [20]. The relation between power and trust is discussed by Luhmann [21].

It should be pointed out, however, that there is a significant difference between trust in physical markets and trust in e-commerce. Building and maintaining trust in electronic markets is more difficult without face-to-face interaction [22], partner identity, and clearly defined legal framework. In online markets it may not be possible to track down or even to identify a party in a transaction. A software agent, for example, may act on behalf of different human users at different moments at time, thereby making it difficult to relate its behavior to one physical entity.

Electronic markets also significantly reduce the costs of establishing new business contacts and changing business partners. This shortens the average life of business relationships, making it more difficult to build trust and consume the benefits of a long trust-based relationship.

Another factor that adversely affects trust is the dynamics and volatility of electronic markets. In e-commerce trust can be destroyed in an instant by misfortune or a mistake and the effect of it could be globalized throughout many interconnected markets.

3 Decision-theoretic model of trust

Trust has different connotations and has been used in different meanings in different contexts by different authors. Many authors [1] consider trust to be a belief or cognitive stance that could eventually be quantified by a subjective probability. We give a brief conceptualization of trust that will help avoid confusion and will facilitate further exposition.

In modeling trust, we take a rational-choice approach. We assume that: (i) There is a preference (utility) function which measures the desirability of an outcome; (ii) Different people may have different preference (utility) functions; (iii) Trust is only possible if, for the trustor; the expected outcome of placing trust is preferred over the expected outcome of not placing trust.

We assume that trust is a bilateral relation that involves an entity manifesting trust called the *trustor* and an entity being trusted called the *trustee*. Further, we assume that

- There is an event Γ that the trustor cannot control and that depends on the trustee. The trustee may have partial or full control over Γ.
- The trustor voluntarily decides to put himself in a position dependent on Γ in the sense that the trustor will benefit if Γ occurs, otherwise he will lose.
- The trustee can honor trust by bringing about Γ, and abuse trust by not bringing about Γ. Usually, the trustee is better off by abusing trust, since bringing about Γ may incur costs and inconvenience to the trustee.

In a trust relation, the trustor takes the risk to depend on the trustee for a certain event the outcome of which depends completely or partially on the trustee. We assume that trustworthiness could be measured by the probability of Γ. For example, the trustee could be a user and $\Gamma = \{$The user does not attempt to elevate his privileges$\}$. Another interpretation is $\Gamma = \{$The quality of the software delivered by the trustee meets the trustor's expectations$\}$.

Consider a trustor who uses a software product developed by the trustee. If the trustor cannot control the process of the design, implementation, and installation of the product, then the trustor may take the risk to trust the developer that the product meets its specifications. It is usually the case that the product developer has better information than the trustor about the security and reliability of the product. A problem may arise if the trustee decides to withhold information from the trustor in order to save production costs and deliver a product of inferior quality. Here, the trustee deliberately and intentionally decides to abuse trust. Moreover, additional problems may occur from events outside the control of the trustee and the trustor. For example, the product may contain a third party off-the-shelf component which happens to be unstable. Here, the problem does not arise because of untrustworthy behavior on the part of the trustee. In another scenario, the trustee is completely willing to deliver a high quality product and puts a lot of effort in the program implementation and testing. Unfortunately, because of lack of experience the trustee overlooks established security practices and delivers an insecure product.

In general, trust can be divided into two major categories:

- Trust based on good will: the trustor believes that the trustee has the good will not to abuse trust. That is, the trustee will bring about the favorable for the trustor event Γ even though it may not be beneficial for the trustee. In this case, the trustee has always the power and capacity (the full control over Γ) to honor trust. Whether he will honer trust depends on his incentives and benefits.

- Trust based on competence: the trustor believes that the trustee is competent enough to bring about the favorable event Γ. It could be the case that the trustee is willing, but not competent enough to honor trust (to bring about Γ).

Trustworthiness can also be classified as *perceived* or *actual*. Perceived trustworthiness is defined as the trustor's subjective belief in the competence or the good will of the trustee. Obviously, the perceived trustworthiness could be different from the objective trustworthiness, which is the actual incentive or capacity of the trustee. Perceived trustworthiness is measured with the thrustor's subjective probability of Γ, and the actual trustworthiness is measured with the objective probability of Γ. For example, a user may install a program, believing that the program failure rate is $\hat{\theta}$, while the actual failure rate is θ.

It should be noted that the concept of trust has two main meanings which are often confused. Trust could refer to both the trustfulness of the trustor and the trustworthiness of the trustee, where the *trustfulness* of the trustor is the extent to which the trustor places trust in the trustee, whereas the *trustworthiness* of the trustee is the extent to which the trustee honors trust. It is obvious that the trustfulness of the trustor largely depends on his expectation about the trustee's trustworthiness (the perceived or subjective trustworthiness). The higher the perceived trustworthiness, the more wiling the trustor is to place trust in the trustee.

In order to form a trusting belief, the trustor must have enough knowledge about the trustee's incentives or competence. Such knowledge usually comes from two sources: individual and public. Individual sources include the history of previous interactions with the trustee, recommendations of other agents, etc. Public sources usually include

the publicly established reputation of the trustee. In some cases, public and individual knowledge could be conflicting. For example, based on a series of positive experiences with a product developer, the trustor may decide to trust him even though it is well known that the developer has incentives to deliver inferior products.

The concept of trust is different from concept of reputation which relates to publicly established and recognized features of an entity. For example, an agent could be completely trustworthy without having any public reputation. Such an agent could efficiently interact with a few other agents that trust him and know him for a long time. Establishing public reputation, however, could be costly or could require long time.

In our interpretation of trust, the trustor does not have control over the trustee's behavior and the event Γ. In a sense, trust and control are mutually exclusive. The more control the trustor has over the trustee, the less the trustor needs to place trust. Obviously, trust relates to the state of being dependent, while control relates to the ability to keep someone in a dependent position.

Formally, the trustor's utility function can be denoted by:

$$U(\hat{\theta}, \Gamma(p_1, ..., p_n)) \tag{1}$$

where U is the trustor's utility, $p_1, ..., p_n$ are parameters describing the event Γ, and $\hat{\theta}$ is the degree of perceived trustworthiness, i.e., the probability with which Γ is expected to happen. A natural measure for the perceived trustworthiness is the subjective probability of Γ, i.e., the trustor's strength of belief that the trustee will bring about Γ. The degree of perceived trustworthiness $\hat{\theta}$ could differ from the degree of actual trustworthiness θ.

The event Γ is favorable to the trustor:

$$\frac{\partial U(\theta, \Gamma(p_1, ..., p_n))}{\partial \theta} \geq 0$$

That is, the trustor benefits from higher trustworthiness. The case of complete trustworthiness is represented by $\theta = 1$, and vice versa, the trustee is completely untrustworthy when $\hat{\theta} = 0$:

$$U(1, \Gamma(p_1, ..., p_n)) > 0$$

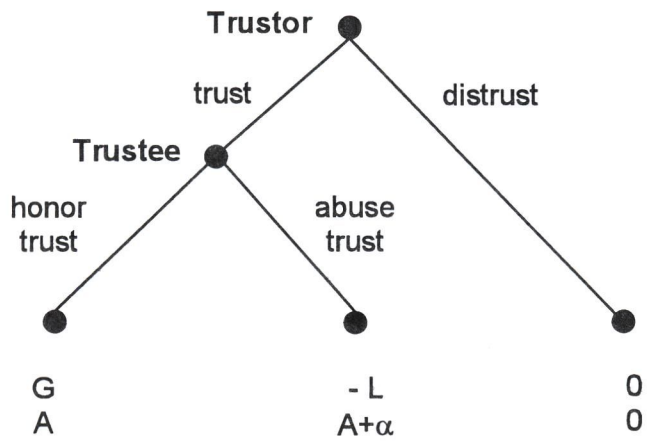
$$U(0, \Gamma(p_1, ..., p_n)) < 0$$

If we assume that utility is a continuous function of trustworthiness, then there is a threshold level θ_0, $\theta_0 \in [0, 1]$, that separates trustworthiness from untrustworthiness:

$$U(\theta, \Gamma(p_1, ..., p_n)) \geq 0 \qquad \text{for all} \qquad \theta \geq \theta_0$$

The trustor is always better off if the other agent's trustworthiness exceeds the threshold θ_0 which depends on the event Γ and its parameters $p_1, ..., p_n$.This defines a natural *participation constraint*: the trustor will place trust on the trustee (or will voluntarily agree to depend on the trustee) if the trustee's perceived trustworthiness exceeds θ_0. The participation constraint corresponds to the intuition that an agent will only engage in an interaction if the trustworthiness of the other party exceeds some threshold (the level of acceptable trustworthiness), which depends on the interaction context (through parameters $p_1, ..., p_n$) and on the trustor (through the trustor's utility function U). In

Figure 1: Extensive form of a Trust Game



other words, the threshold θ_0 is both objectively and subjectively determined. Such a formalization agrees with the threshold model of trust (Kee and Knox,1970; Coleman, 1990; Tan and Thoen, 1999).

Such a formalization of trust is domain independent and captures a wide range of applications, where the trustor believes that the trustee will behave in some expected way specified by the event Γ. Depending on the context the event Γ, can be given different interpretations.

4 Game-theoretic model of trust

A trust relation can be modeled as a one-shot extensive-form game shown in Figure 1. The game starts with a move made by the trustor, who must choose between placing trust and not placing trust. If the trustor decides not to place trust, the game is over and both the trustor and the trustee receive a payoff 0. If the trustor places trust, the trustee has two choices: to honor or to abuse trust. The trustee is usually better off by abusing trust (by not bringing about Γ), while the trustor is better off if trust is honored, i.e., Γ is brought about. If the trustee honors trust, the trustor and trustee receive G and A, respectively. If the trustee abuses trust, the trustor receives $-L$ and the trustee receives $A + \alpha$. Here, α represents the incentive of the trustee to abuse trust. We assume that these payoffs represent utilities for the trustee and the trustor that correspond to the outcomes of the game. The relationship between payoffs is as follows: $G > 0, L > 0, A > 0, \alpha > 0$.

If the trustor places trust, the trustee chooses between A and $A + \alpha$, so surely he will abuse trust. Anticipating this, the trustor chooses between lost L if he places trust and 0 if he does not. Therefore, there is a Nash (and subgame-perfect) equilibrium of the game in which the trustor does not place trust, and the trustee always abuses trust. The equilibrium is socially inefficient because both the trustor and the trustee are worse

off in the situation in which trust is not placed than in the case where trust is placed and honored.

This is one equilibrium of the game, but there is another in which the trustor places trust if the trustee honors trust, and the trustee always honers trust. The problem with this equilibrium is that it is not as credible as the first one. It would be irrational for the trustee to honor trust once the trustor has chosen to place trust. In other words, the second equilibrium is not subgame-perfect.

Which equilibrium will be chosen depends on the trustor's expectation of the trustworthiness of the trustee (the so called perceived trustworthiness). If the game is played only once and does not affect the interactions that the trustee is going to have with other agents, then the more plausible equilibrium is the non-cooperative one in which trust is withhold.

5 A Model of Trust Learning

In this section we use the game-theoretic model of trust developed in the previous section to analyze how trust could be learnt over time. A detailed description of the model can be found in [23].

Research on risk perception [7] indicates that trust typically is learned gradually over many repeated interactions. As we have seen in the previous section, one interaction is not sufficient to establish trust because the trustee has a short-run incentive to abuse trust. Hence, the non-cooperative equilibrium prevails.

In this section, we consider repeated interactions between the trustee and the trustor. A repeated game consists of infinite repetitions of the Trust game, shown in Figure 1. A repeated game allows the trustor to modify his expectations about the incentives and abilities of the trustee by observing his behavior in past interactions. Initially, the trustor starts with some prior beliefs about the trustworthiness of the trustee. As the game proceeds, the beliefs are adjusted. After a series of positive interactions, the trustor's trust in the trustee will gradually increase, and vice versa, multiple unsuccessful interactions in which trust is abused, may lead to such a low estimate of the trustworthiness of the trustee that the trustor no longer wishes to interact. In this case the game ends. Let θ_0 be the threshold, such that whenever the perceived trustworthiness of the trustee, $\hat{\theta}$, falls bellow θ_0, the trustor exits the game. This defines a natural *participation constraint*: the trustor will interact with the trustee (or will voluntarily agree to depend on the trustee) if the trustee's perceived trustworthiness exceeds the threshold:

$$\hat{\theta} \geq \theta_0$$

The structure of the game and the payoffs are common knowledge between the trustor and the trustee. To introduce reputation effects into the game, suppose that the objective trustworthiness of the trustee is θ. In other words, the trustee abuses trust with probability θ, if trust is placed. Further, assume that θ is private knowledge known only to the trustee. Therefore, the trustor will place trust if the trustor's expected benefit from placing trust exceeds the benefit from withholding trust:

$$\hat{\theta}G - (1 - \hat{\theta})L > 0$$

where $\hat{\theta}$ is the perceived trustworthiness of the trustee, i.e, the trustor's estimate of θ. Therefore, the threshold is:

$$\theta_0 = \frac{L}{L+G}$$

In this setting, the problem of learning trust is reduced to the problem of estimating θ based on a series of interactions in which trust is abused or honored. In order to preserve belief consistency, it is natural to assume that the trustor performs belief revision using Bayes' rule. To find a Bayesian estimator of the trustworthiness of the trustee, we use the Laplace succession rule [24]. The rule estimates the probability that the trustee will honor trust in the next interaction, given that he honored trust in h out of k interactions, a more general estimator can be used:

$$\hat{\theta} = \frac{h+1}{k+2} \tag{2}$$

Formula 2 suggests that there is a Nash equilibrium of the repeated Trust game in which:

- The trustor places trust if and only if:

$$\hat{\theta} = \frac{h+1}{k+2} > \theta_0$$

- The trustee alternates between honoring and abusing trust so that to keep the trustor's estimate $\hat{\theta}$ above the threshold θ_0.

The next Proposition follows naturally from the definition of the repeated Trust game and from players' equilibrium strategies:

Proposition 1 *In the equilibrium, the trustee alternates between honoring and abusing trust so that:*

$$\frac{h}{a} \geq \frac{L}{G}$$

where a is the number of times trust is abused, h is the number of times trust is honored, L is the lost to the trustor if trust is abused, and G is the gain to the trustor if trust is honored.

Proposition 1 suggests that the ratio of the number of times trust is honored to the number of times trust is abused depends on the ratio of the loss to the gain for the trustor. The result is intuitively reasonable. The more the trustor can lose in the game, the more frequently the trustee honors trust. Since by definition $L > G$, the trustee needs to honor trust more frequently than to abuse trust.

6 The effect of trust on market efficiency

In this section, we study the effect of trust on market efficiency. More specifically, we show that if agents hold accurate trust estimates about one another, then the social

welfare, the amount of trade and the agents' utility levels are maximized. We also show that market efficiency does not require complete trustworthiness. Untrustworthy agents could transact as efficiently as trustworthy agents, provided that they hold accurate estimates of one another. Therefore, what really matters is not the actual level of trustworthiness, but the accuracy of individual estimates. A market in which agents are trusted to the degree they deserve to be trusted is as efficient as a market with complete trustworthiness.

A detailed description of our analysis can be found in [25, 3]. The analysis is done within the framework of a bilateral negotiation involving a buyer and a seller. The seller produces some commodity and sells it to the buyer. We assume that the seller always delivers the commodity i.e., he is completely trustworthy. The buyer's trustworthiness, however, may vary. In other words, the buyer pays with some probability θ, $\theta \in [0, 1]$. We assume that the seller delivers first and after that the buyer pays. In this case, the seller depends on the buyer for the event $\Gamma = \{\text{the buyer pays}\}$, and the seller has to choose whether to enter into a transaction without being able to control Γ.

Let $\hat{\theta}$ and θ denote respectively the buyer's perceived and actual trustworthiness. That is, $\hat{\theta}$ is the seller's estimate of θ. One way to look at θ is to see it as an indicator of the buyer's willingness to pay. In another interpretation, θ could be the buyer's ability or capacity to pay (the buyer, for example, may be willing to pay, but may not have available funds). θ could also be the probability that the contract between the buyer and the seller can be enforced by an enforcement agency. That is, θ could be the probability of detecting particular violation, imposing particular type of sanctions, and using particular type of procedures for adjudicating disputes.

The seller's estimate, $\hat{\theta}$, could differ from the buyer's actual trustworthiness, θ. It is usually the case that the buyer is undertrusted. That is, $\hat{\theta} < \theta$. For example, many risk assessment firms treat the lack of credit history as a lack of trust. This is usually motivated by the fact that the marginal cost of obtaining additional evidence of trust exceeds the marginal benefit of the evidence. Undertrusting is a typical example of a market inefficiency produced by inaccurate trust estimates. In the extreme case, undertrusting could produce such a low estimate of the partner's trustworthiness, that an agent might decide not to participate in a transaction, even when the partner is completely trustworthy.

The following proposition shows the optimal amount of trust necessary to achieve market efficiency.

Proposition 2 *When the trust matches trustworthiness, $\hat{\theta} = \theta$, the seller and the buyer both maximize their individual utility functions. Moreover, the social welfare is maximized and the maximal possible output is produced and sold.*[1]

The next proposition shows that the undertrusting leads to market inefficiencies.

Proposition 3 *When the buyer is undertrusted, $\hat{\theta} < \theta$, the social outcome is not optimal. Namely, quantity exchanged and the utilities of both agents are smaller than those obtained in the case when the seller places the correct amount of trust on the buyer, $\hat{\theta} = \theta$.*

[1] The proofs of Propositions 2 and 3 can be found in [25]

According to Propositions 2 and 3, the case where the seller trusts the buyer to the extent that the buyer deserves to be trusted is optimal for society. By an optimal outcome we mean one that maximizes the social welfare, the quantity produced, and the agents' utility functions. In order to obtain efficiency it is not necessary that the buyer be trustworthy. The only relevant factor is the seller's accuracy in estimating the buyer's trustworthiness. Any underestimation of the buyer's trustworthiness tends to harm each agent and society as a whole.

It is worth pointing out that in the case when trust matches trustworthiness, $\hat{\theta} = \theta$, individual interests coincide with the social interest in the sense that agents maximize their utilities, maximizing at the same time the social welfare.

If the buyer is distrusted, $\hat{\theta} < \theta$, he may try to convince the seller of his trustworthiness. One possible way for a distrusted buyer to signal his trustworthiness is to make an advance payment to the seller, i.e., to pay some amount before the seller delivers the commodity. We assume that in the absence of payment the seller is relieved from his obligations. That is, if the buyer does not make an advance payment after he has promised to do so, the seller is relieved from his obligation to deliver.

Proposition 4 *If the buyer is distrusted, $\hat{\theta} < \theta$, and the agents choose an advance payment contract, then the social welfare is maximized. The quantity exchanged with an advance payment contract equals the quantity exchanged with an uncertain payment contract. Moreover, the advance payment contract gives each agent higher utility than the uncertain payment contract.*

According to Proposition 4 if the buyer is trustworthy, he should pay the entire price in advance. This corresponds to our intuition that paying before or after the delivery does not make a difference for a trustworthy buyer. Furthermore, Proposition 4 says that the advance payment contract is better than an uncertain payment contract in that it provides agents with higher utility levels.

From Propositions 4 it follows that when trustworthiness is underestimated, the advance payment contract is optimal. Advance payment contracts could also serve as a screening device. That is, they could help separate trustworthy agents from untrustworthy ones. If the buyer is trustworthy, he should not object to paying the whole price in advance. Therefore, a buyer who declines an advance payment contract is not as trustworthy as he claims to be.

7 Incentive Compatible Mechanism for Trust Revelation

In this section, we discuss a solution to the problem of obtaining trust estimates: an incentive-compatible mechanism in which agents truthfully reveal their trustworthiness at the beginning of every interaction. In such a mechanism agents always report their true level of trustworthiness, even if they are untrustworthy. Honest reporting at the first stage of interaction lets other agents know the interaction risk and form realistic expectations about possible outcomes. The mechanism works for both single-step and repeated interactions. A complete description of the mechanism is given in [4, 26]

To illustrate an incentive-compatible mechanism for trust revelation consider the following scenario. The trustor and the trustee take part in a transaction, $T(q, \theta)$, which can be described by a generic parameter q and the level of the trustworthiness of the trustee θ. The generic parameter q denotes the transaction terms which depending on the transaction context, could include price, quantity, services provided, or combination of them. The parameter θ denotes the trustworthiness of the trustee, i.e., the probability that the trustee will carry out his part of the transaction.

We assume that the trustor always executes the transaction i.e., he is completely trustworthy. The trustee's trustworthiness, however, may vary. In other words, the trustee carries out the transaction with some probability θ, $\theta \in [0, 1]$. The level of trustworthiness θ is a private value known only to the trustee. In other words, in the beginning of the transaction the random realization of the trustworthiness of the trustee is not observable by the trustor. Further, we assume that the trustor offers a transaction to the trustee by specifying the terms of the transaction q. The trustee could accept or reject the offer of the trustor.

Given this setting, what is the optimal transaction for the trustor to offer? In other words, the trustor faces the problem of designing an optimal (i.e, utility maximizing) offer. For example, the trustor might ask the trustee to declare his level of trustworthiness θ, and then based on the declaration, the trustor could choose the value of q. It is usually the case, however, that the trustee has a strong incentive to misrepresent his level of trustworthiness. By declaring more trustworthiness the trustee usually enjoys more benefits of future cooperation and more opportunities to abuse trust. By implementing an appropriate incentive-compatible mechanism the trustor can make the trustee truthfully reveal his level of trustworthiness.

An incentive-compatible mechanism for trust revelation works in the following way. First, at the beginning of the interaction the trustee declares his trustworthiness, $\tilde{\theta}$, and after that the transaction terms q are chosen by the trustor as a function q, $q = q(\tilde{\theta})$, of the trustee's declaration. In an incentive-compatible mechanism the trustee finds it optimal to report his trustworthiness truthfully, i.e., $\theta = \tilde{\theta}$.

The following proposition establishes the existence of an incentive-compatible mechanism for trust revelation.

Proposition 5 *There exists a single-transaction incentive-compatible mechanism* $q = q(\tilde{\theta})$ *that makes the trustee truthfully reveal his trustworthiness at the beginning of the transaction, if the following conditions hold:*

$$\frac{\partial^2 U(q, \tilde{\theta})}{\partial q^2} < 0$$

$$\frac{\partial U(q^*, \tilde{\theta})}{\partial q} = 0$$

for some q^*, *and*

$$\frac{\partial^2 U(q^*, \tilde{\theta},)}{\partial q \partial \tilde{\theta}} \neq 0$$

where $U(q, \theta)$ *is the trustee's utility function.*[2]

[2]The proofs of Propositions 2 and 3 can be found in [4, 26]

The basic idea behind the mechanism is that the trustee achieves maximum utility only if specific transaction terms are chosen. Since the transaction terms, q, depend on the trustee's declaration, $\tilde{\theta}$, by declaring his true level of trustworthiness the trustee maximizes his utility. The mechanism is designed in a way to give the trustee sufficient incentives to report truthfully even if it is untrustworthy.

The results of Proposition 5 can be carried over to the case of repeated transactions. By a repeated transaction between the trustor and the trustee we mean finitely or infinitely many repetitions of the single transaction in which both the trustor and the trustee discount their future payoffs at a constant rate.

In the case of a repeated transaction, it is not obvious that the trustee will truthfully declare his trustworthiness. For example, the trustee could lie in the first transaction in order to obtain more favorable transaction terms, q, in the subsequent transactions. the next proposition shows that trust revelation is possible in the case of a repeated transaction.

Proposition 6 *Under the conditions of Proposition 5, for any repeated transaction between the trustor and the trustee there exists an incentive-compatible interaction mechanism* $q = q(\tilde{\theta})$ *that makes the trustee truthfully reveal his trustworthiness in the beginning of the first transaction.*

An incentive-compatible mechanism for trust revelation provides several advantages. It does not rely on a third party for providing information of for backing up a transaction. It does not depend on collecting and analyzing information about untrustworthy agents. In addition, it does not require the estimation of other agents' trustworthiness. This solves the problem of inaccuracy of individual estimates and avoids many inefficiencies caused by inconsistent or inaccurate estimates. The mechanism also eliminates the need to speculate about other agents' intentions and beliefs.

8 Conclusions

The paper defines a game-theoretic and a decision-theoretic model of trust. The game-theoretic model is used to show how trust can be learnt over time. The equilibrium analysis demonstrates that instead of always honoring trust, the trustee alternates between honoring and abusing trust so that to keep the trustor's estimate above a certain threshold.

The decision-theoretic model of trust is used to analyze impact of trust on electronic markets. The paper shows how trust affects market efficiency, social welfare, the volume of trade, and the profits of market participants. The paper also demonstrates that complete trustworthiness is not necessary for market efficiency. Instead, every transaction has its optimal level of trustworthiness that is needed for the transaction completion.

The paper introduces the concept of trust-based mechanism design. The primary objective of trust-based mechanism design is to develop and implement optimal markets in which every agent discloses truthfully his/her trustworthiness before a transaction starts.

Most existing trust-building mechanisms assume complete or partial intervention of trusted third parties, such as enforcing agencies, reputation systems, etc. Such arrangements are usually costly and not always available to economic agents. The primary advantage of the trust-based mechanism design is that the mechanism is self-enforcing and does not rely on third parties, i.e., it is in the best interest of every agent to declare his/her true level of trustworthiness.

References

[1] A. Josang, "Trust-based decision making for electronic transactions," in *Proceedings of the Fourth Nordic Workshop on Secure Computer Systems (NORDSEC '99)*, Sweden, 1999, pp. 496–502.

[2] S. Marsh, *Formalizing Trust As a Computational Concept*, Ph.D. thesis, University of Stirling, U.K., 1994.

[3] S. Brainov and T. Sandholm, "Contracting with uncertain level of trust," in *Proceedings of the First ACM Conference on E-Commerce*, Denver, Colorado, 1999, pp. 15–21.

[4] S. Brainov, "An incentive compatible trading mechanism for trust revelation," in *Proceedings of the IJCAI'01 Workshop on Economic Agents, Models and Mechanisms*, 2001, pp. 62–70.

[5] C. Castelfranchi and R. Falcone, "The dynamics of trust: From beliefs to action," in *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies*, Seattle, 1999, pp. 41–54.

[6] J. Coleman, *Foundations of Social Theory*, Harvard University Press, 1990.

[7] P. Slovic, "Risk perception and trust," in *Trust: Fundamentals of Risk Analysis and Risk Management*, Vlasta Molak, Ed. Lewis Publishers, 1997.

[8] S. Kaplan B. Garrik, "On the quantitative definition of risk," *Risk Analysis*, vol. 28, pp. 11–27, 1981.

[9] P. Dasgupta, "Trust as a commodity," in *Making and Breaking Cooperative Relations*, D. Gambeta, Ed. Basil Blackwell, 1990.

[10] D. Kreps and R. Wilson, "Reputation and imperfect information," *Journal of Economic Theory*, vol. 27, pp. 253–279, 1982.

[11] C. Snijders, *Trust and Commitments*, ICS, 1996.

[12] V. Buskens, *Social Networks and Trust*, Kluwer, 2002.

[13] A. Boot, S. Greenbaum, and V. Thakor, "Reputation and discretion in financial contracting," *American Economic Review*, vol. 83, pp. 1165–1183, 1993.

[14] P. Resnick, R. Zeckhauser, F. Friedman, and K. Kuwabara, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[15] M. Schillo, P. Funk, and M. Rovastos, "Who can you trust: Dealing with deception," in *Proceedings of the Second Workshop on Deception, Fraud and Trust in Agent Societies*, Seattle, 1999, pp. 95–106.

[16] Zacharia and Maes, "Trust management through reputation mechanisms," *Applied Artificial Intelligence*, vol. 14, no. 8, pp. 881–907, 2000.

[17] E. Friedman and P. Resnick, "The social cost of cheap pseudonyms," Working paper, 1998.

[18] L. Werin and H. Wijkander, *Contract Economics*, Basil Blackwell, 1992.

[19] J. Harsanyi, "Measurement of social power, opportunity costs and the theory of two-person bargaining games," *Behavioral Science*, vol. 7, pp. 67–80, 1962.

[20] S. Brainov and T. Sandholm, "Power, dependence and stability in multiagent plans orlando, 1999.," in *Proceedings of the National Conference on Artificial Intelligence AAAI'99*, Orlando, Florida, 1999, pp. 11–16.

[21] N. Luhmann, *Trust and Power*, John Wiley and Sons, 1979.

[22] E. Rocco, "Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact," in *Proceedings of ACM CHI 98 Conference on Human Factors in Computing Systems*, 1998, pp. 496–502.

[23] S. Brainov, "Trust learning based on past experience," in *Proceedings of the IEEE International Conference on Integration of Knowledge Intensive Multiagent Systems (KIMAS)*, 2005, pp. 197–201.

[24] W. Feller, *In Introduction to Probability Theory and Its Applications*, vol. 1, John Wiley and Sons.

[25] S. Braynov and T. Sandholm, "Contracting with uncertain level of trust," *Computational Intelligence*, vol. 14(4), pp. 501–514, 2002.

[26] S. Braynov and T. Sandholm, "Trust revelation in multiagent interaction," in *in Proceedings of CHI'02 Workshop on The Philosophy and Design of Socially Adept Technologies*, 2002, pp. 55–60.

DISTRIBUTED SYSTEM FOR THE ACQUISITION OF SKILLS IN JAVA PROGRAMMING: A SUMMARY OF EXPERIENCES

Bandáková Jana
Čierny Marián
Samuelis Ladislav Ing., Ph.D.
Dept of Computers and Informatics
Faculty of Electrical Engineering and Informatics
Technical University of Košice
Letná 9, 0402 Košice,
Slovakia
Ladislav.Samuelis@tuke.sk

Abstract

This contribution describes briefly a distributed application (dedicated learning management system), which main aim is to support testing procedures of students' during their Java programming skills acquisition and to introduce them interactively into the component based programming. Application enables teachers (tutors) to automate repetitive tasks of building groups of random tests, assignments and their evaluation. It also supports archiving of the results for their later evaluation and tracking of students' progress.

Keywords

E-learning, three-layer architecture, Java tests

1. MOTIVATION AND THE PURPOSE OF THE SYSTEM

The motivation for the development of the application came from the observation that most effective way of learning a new programming paradigm for novice programmers is by doing experiments with the already existing code. This observation is not new in principle and it was utilized for various purposes in various contexts in the past (e.g. Psenakova I., 2000). The Java enabled distributed technology enables to build robust environment (learning management system with limited capabilities) for the provision of learning-by-experimenting. In other words the aim of this application is to support project-based learning environment for novice students. But a question still remains. What type of code (case-study) will we present to the learning community? We have decided that the case-study will "document" (or copy the architecture) of the LMS itself. I.e. building a typical 3 layer distributed application (automatic teller machine) will be the base motive of the case-study too. This approach also ensures that builders of the system gain deeper insight into the trickles of the code.

From the pedagogical point of view the purpose of the system is twofold.

1. The application, on one hand, automates some boring repetitive pedagogical tasks, i.e. preparation of tests, their random grouping and evaluation.
2. On the other hand, it provides a case-study and of chunks of codes for their later incremental change and this way to better understanding of the code behaviour.

These pedagogical tasks are not only automated as far as possible, but the system also provides a base for the later assessment of students' progress. We also state that a consequent evaluation of the progress of students' during their learning provides teachers with highly valuable information, which is necessary for raising the quality of the future learning materials. For this reason we have embedded archiving facilities for thorough tracking of the successfulness or failure of students' learning progress.

In the next paragraph we briefly describe the architecture and several selected functions of the system, which are available through the systems' graphical user interface.

2. ARCHITECTURE OF THE APPLICATION

The application implements the popular three-layer architecture (Horstmann C., 2002), as depicted on Figure 1. This architecture partitions the system into three logical layers:

- the user interface layer
- the business rules layer
- the database access layer

The three-layer architecture is used when an effective distributed client/server design is needed that provides (when compared to the two-layer) increased performance, flexibility, maintainability, re-usability, and scalability, while hiding the complexity of distributed processing from the user. The second layer (middle layer server) is between the user interface (client) and the data management (server) components. This middle layer provides process management where business logic and rules are executed and can accommodate hundreds of users by providing functions such as queuing, application execution, and database staging. There are a variety of ways of implementing this middle layer, such as transaction processing monitors, message servers, or application servers. In addition the middle layer adds scheduling and prioritization for work in progress. The three-layer client/server architecture improves the performance for groups with a large number of users. In the two-layer architectures database access functionality and business logic were often contained in the client component, any changes to the business logic, database access, or even the database itself, often required the deployment of a new client component to all users of the application. Multi-layer client/server architecture enhances the two-layer client/server architectures in two ways:

- It makes the application less fragile by further insulating the client from changes in the rest of the application.
- It allows more flexibility in the deployment of an application.

Multi-layer client/server reduces application fragility by providing more insulation and separation between layers. The user interface layer communicates only with the business rules layer, never directly with the database access layer. The business rules layer, in turn, communicates with the user interface layer on one side and the database access layer on the other. Thus, changes in the database access layer will not affect the user interface layer because they are insulated from each other. This architecture enables changes to be made in the application with no affects to the client component.

At the present time all three layers are implemented on Intel-based PCs and Linux type operating system. The application server hosts the Tomcat software, which is the open source implementation for servlets and the JSP technology, (http://phoenix.mis.cycu.edu.tw/Class/2001Su_J2EE). Tomcat also interprets Java commands and enables the communication with database through libraries.

The client is usually a browser such as Internet Explorer, Netscape, Opera etc. Browsers interact with the server through protocols. There are many protocols available on the Internet, which are used for this process. In our application the HTTP protocol (HyperText Transfer Protocol) is utilized. The architecture of the application is on the next Figure 1:

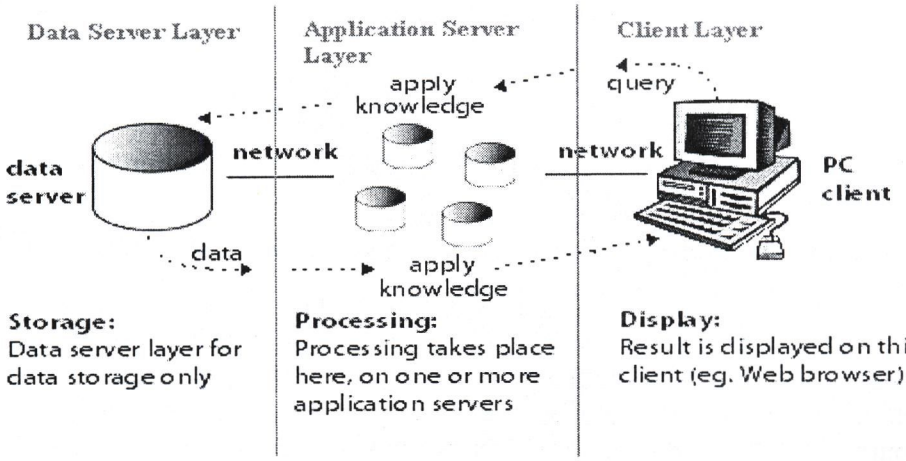


Figure 1. The three-layer architecture of the application

The database is processed by the MYSQL RDBMS, which may be located remotely. All pieces of information, which are necessary for the application, such as students' and teachers' usernames and passwords, personal information and students' test results, are stored in the database. The application server is used to process the queries to the database server. For example, if student or teacher enters into the system their usernames and passwords are sent to the application server, which forwards the query to the database server. The database server sends back the appropriate information.

The database contains:

- test questions and answers
- test results
- students' and teachers' personal information
- other pieces of auxiliary information

JSP source code and the compiled code, which manages the communication between server and database and between server and client, are stored in specific directories. The compilation runs on the server with authorized access. If the compilation is successful the information message will be displayed. Servlets are products of the JSP code compilation.

The recent version of the application was developed on local PCs with the installed Borland JBuilder 2005 Enterprise software. Development of the application requires

installation of the Tomcat software (Tomcat 4.0) for processing the JSP (Java Server Pages) technology (http://www.apl.jhu.edu/~hall/java/Servlet-Tutorial), (Hall M., 2000) and for the cooperation with the MYSQL database (http://dev.mysql.com/doc/, 2004).

3. GROUP OF FUNCTIONS PROVIDED FOR STUDENTS

Students enter the system through their "username" and "password". If the entry to the system is unsuccessful, an error message appears and the client can try to enter the system again. Successful entry into the system consists of two steps. Firstly, the system checks the entered "username" and "password" against their values in the database. Secondly, the system checks, whether the authorized client is already logged in. This approach allows the following:

- Only that student who is registered may enter the system.
- It is possible to define a definite time period for opening a session in the system.

The aim of the time interval definition is to control the access of students to tests during the practical labs and in this way to avoid cheating.

Function "Registration". This function is used for the registration of new students. Registration consists of two steps. Students register into the system and the basic information is stored into the database. Teachers later admit students to enrol into a specific course e.g. Java course. This facility enables teachers to control the access to specific courses. Teacher has exclusive rights to update the list of students who are eligible to take a specific course. Students are notified automatically about the registration via e-mail. In case of successful entry into the system, the following functions are available for students:

Function "Personal information". Student's personal information is stored in the database and he/she may modify them later.

Function "Test results". This function provides students with viewing the test results, percentage of successfulness, teacher's classification and history of particular tests with correct and incorrect answers. When student finishes the test, system grades the performance and delivers appropriate notes in accordance to given criteria. Teacher is able to modify the automatically generated grade.

Function "Type of the test". This function provides student with the selection of the test type, which he/she intends to accomplish. During the training phase student can choose more questions and at the end of the test is able to view the successfulness in percentage and the correct answers. Teacher controls the number and the type of the questions.

Function "Change password". Student who entered the system can change his/her login and password and later logout the system

4. GROUP OF FUNCTIONS FOR TEACHERS

Teachers enter the system through "username" and "password" like students. If the entry to the system is unsuccessful, an error message appears and the client can try to enter the system again. If the entry to the system is successful then the following functions are available:

Function "Students". This function displays the list of enrolled students.

Function "New account". Teacher is able to add new students to the system.

Function "Security". This function is used to specify student's "username" and "password" to ensure that a specific student makes the test in a definite time-period. Teacher may create these usernames and passwords manually or automatically by software generator. Teacher can also set the time interval, during which the "username" and "password" are set for the first entry. When the time interval elapsed, which was set for passing the test, the system resets the initial settings for the student. To sum up, teacher has the following rights:

- To view and update students' username and password.
- To view and update the time interval, during which the student is allowed to enter the system. If a student does not make the test in the defined time interval then it is necessary to generate new username and password for a new time-interval. This feature contributes to the enhancement of the system security.

The purpose of these teacher's rights is to prevent the following attempts for cheating:

- Student gets the username and password directly before he/she passes the test so it is difficult to send these data to other persons.
- Student's username and password is generated for a definite time-interval, during which student is allowed to pass the test.

Function "Search". Teacher can search students by name, by surname or by group.

Function "Presence". Teacher is able to control the presence of students in labs in number of weeks. The number of the weeks may vary for the Java courses.

Function "Questions". Questions are divided into three groups: exam questions, test questions and credit questions. Teacher can add new questions or modify the existing questions in the system.

Function "Results". Teacher is able to track the study progress of his/her students in sorted tables. System evaluates the results and offers noted grades, which are subject to the teacher's approval.

5. IMPLEMENTATION OF THE CASE-STUDY AND THE PROJECT-BASED APPROACH TO LEARNING

As it was already mentioned in the introduction, the case-study is a part of the dedicated e-learning system. The main purpose of the case-study is to help students in gaining practical skills in mastering Java programming language by developing a complete application,

which simulates the automatic teller machine example. The core architecture of the application is from the (Horstmann C., 2002).

The specification of the task is as follows. The application has three actors: Client of the bank, Employee of the bank and Administrator of the bank.

1. The Client is provided with the following abilities: Clients have their own customer number and a personal identification number (pin), which are necessary for the entry into the bank system. Every client has two accounts:
   a. Saving account and
   b. Checking account (puts some limits on the balance)

   The client can add or withdraw money from an account or to transfer between accounts.
2. The Employee of the bank enters the system through "username" and "password". The bank employee has the next functions available in addition:
   a. Registration of the new client;
   b. Clearing the account of the existing client (it removes the client from the bank)
3. The administrator of the bank also enters the system through "username" and "password" like an employee of the bank. The administrator has the next functions available in addition:
   a. Registration the new employees of the bank;
   b. Removal of bank employees from the database;
   c. Changing the personal information of the bank employee.

From technical point of view this case-study was prepared from a running java application developed in Borland JBuilder 2005 Enterprise environment, it uses the JSP (Java ServerPages) technology and cooperates with MYSQL database server. The main purpose of using these technologies was the necessity to provide dynamic HTML pages and to monitor and archive the students' learning progress.

This case-study's code and explanation is reachable from the dedicated learning management system, which is explained in chapter 3. It offers students an anatomized online view to the selected developmental phases of the case-study for the study purposes. It also suggests additional learning materials and exercises in order to support the project-based learning.

In the following text we briefly describe the available functionalities for students.

Button "**Visited chapter**" This button triggers a table with two columns "Name of chapter" a "Date of the last visit". The number of rows in the table is equal to the number of chapters of the case-study. If the chapter was not visited then the appropriate field in the row is marked as "Not visited". If the chapter was already visited then the background colour of the row of the table changes to grey colour and in the appropriate field of the row is updated with the actual datum of the visit.

Button "**Dictionary**" – enables access and detailed explanation of items, which are referenced in the case-study.

Button "**Links**" – points to web links that are related to topics dealt in the chapter. Web links are divided into two parts: Java related and MYSQL related links. In spite of the fact that the focus of the case-study is on the Java skills acquisition, it is necessary to know the basics of databases because of the case-study distributed nature. In case those students prefer other databases, they are free to study the appropriate database.

Button "**Quiz**" - provides very simple questions and refers to the most important topics dealt in the chapter. Not every chapter is finished with quiz.

Button "**Content**" – triggers a menu to access any chapter of the case-study.

The students' interface provides also a scroll menu for quick orientation and picking up chapters of the case-study.

6. CONCLUSION

To sum up, the main goal of the course is to uncover the developmental process of a non-trivial distributed application through the provision of the case-study. Students have at disposal the Java code in every step of the development. At the end of each module there are suggestions for tasks (or projects), which are constructed as requirements for minor incremental changes in the existing case-study code. This approach assures that a student (or group of students) is not "lost" in the main thread of the application development. Teacher consults the specification of the projects and later conducts discussions around the observed problems. This project-oriented learning aim is to make learning more efficient for students. We plan to evaluate the obtained data statistically later.

REFERENCES

[1] Hall, M. (2001). Core Servlets and JavaServer Pages (JSP). Neocortex spol. s.r.o., 2001

[2] Horstmann, C. (2002). Big Java. RR Donnelly & Sons Company, 2002

[3] http://dev.mysql.com/doc/, (2005-02-07)

[4] http://phoenix.mis.cycu.edu.tw/Class/2001Su_J2EE/J2EE01, (Client_Server_Architecture).pdf (2005-02-08)

[5] http://www.apl.jhu.edu/~hall/java/Servlet-Tutorial (2005-02-07)

[6] Pšenáková, I. (2002): Internet v dištančnom vzdelávaní. In: *Sborník příspevku z mezinárodní konference Modernizace vysokoškolské výuky technických předmetů.* 1. vyd. Hradec Králové: Gaudeamus, 2000, s. 171-173. ISBN 80-7041-723-4.

COMMERCIAL LMS VERSUS OPEN SOURCES LMS

Tatiana Urbanová
Novitech Partner Ltd.
Education Centre TeleDom
Moyzesova 58, 040 01 Košice, Slovak Republic
Phone:+421 55 7274428,Fax:+421 55 7274 429,E-mail:Urbanova@teledom.sk

Dana Šiškovičová
Department of Applied Mathematics and Business Informatics,
Faculty of Economics, Technical University of Košice
Němcovej 32, 040 01 Košice, Slovak Republic
Phone: +421 55 602 3269, Fax: +421 55 63 309 83, E-mail: dana.siskovicova@tuke.sk

Abstract

The paper presents the experience with e-learning through LMS of two education institutions - Faculty of Economics, Technical University of Košice and Teledom, Educational centre, Košice. We want to propose the comparison of three used learning management systems: eDoceo, ulern and Moodle from different point of view.

Keywords

e-learning, LMS, sharing of course content

1. INTRODUCTION

In today's knowledge economy, learning is needed to survive and to thrive. This is true for individuals, for organizations, for communities, for nations. In this context, distance learning has become an imperative. The nature of our society and economy drives the need for learning. The demand for education – formal, informal, lifelong – is necessary to survive and develop in knowledge society. The use of alternatives to the typical classroom setting has been ongoing for more than 100 years – from correspondence courses in paper form through video and computer access.

Today's student population can benefit from valuable learning experiences by interacting with their peers and tutors online, without the inconvenience of commuting to a physical location and adhering to a rigid timetable to learn. In the world of e-learning and course design a lot of effort has been put into research and ideas on how to organize and structure on-line course environments. Helping to fulfil this objective, the Learning Management Systems (LMS) are used. LMS are also referred to as Course Management Systems (CMS) or Virtual Learning Environments (VLE). LMS allow institutions to provide online environments for distance, on-campus and hybrid learning. A wide range of commercial and open source products are now available (WebCT, Blacboard, ANGEL, eCollege, Intralearn, PostNuke, etc.). One of such LMS is also u-Lern, eDoceo and Moodle with which we have practical experiences. In this paper we present brief comparison of these three LMS.

Creation of an e-learning base by individuals can be mentioned as one of the possible ways, how to do it. The exchange of experiences and knowledge throughout the institution

or between the institutions seems to be more efficient. Because in Slovakia exists a relatively well-developed partnership between private companies and the educational sector, primarily the universities; we have possibility to do some decisions in field the ICT based learning using the information and practices of our partners. The mentioned practices are also presented in second part of this paper - e-learning practices.

2. E-LEARNING PRACTICIES

2.1 Faculty of Economics, Technical University of Košice

To accomplish the requirement of the e-learning, the top management of the Faculty of Economics determined as one of the main priority introduction of the distance learning methodology to all types of the education provided by Faculty – bachelor and master study programmes and continuing education courses [1].

Since the interest of the students for regular and external form of study in the offered study branch "Finance, Banking and Investment" exceeds every year the possible capacity of the Faculty more than 10 times, there are good motives to propose to the applicants another possibility of study – distance learning.

The introduction of LMS system and their usage in Faculty of Economics, Technical University of Košice dates back in the end of days of last century. The Faculty builds on knowledge, skills and experiences gained during the realized distance-learning project as:

- **"Distance Education in the area of Finance, Banking systems and Investment"** was focused on the development and pilot run of the two-term distance education course "Postgraduate distance study in the field of Finance, Banking systems and Investment". The aim of the project was to develop postgraduate distance education course and it was accredited at the Ministry of Education of the Slovak Republic.
- **"Effective Methods of Teaching"** was focused on the modern pedagogy, effective communication skills and coping with stress in pedagogical process, IT, telecommunication and multimedia in the teaching process and distance learning methodology.
- **"Preparation, management and strategic planning of the educational projects"** was focused on the strategic planning of the educational projects, preparation and writing of the project proposals, project management and communication skills and intercultural communication.

For supporting mentioned projects was chosen u-Lern LMS. This LMS and its features are described in the third part of the paper. Due to requirements of today web development (portal structured webs) and easy navigation we start using the Moodle LMS, during last year for fulfil the project the **"Preparation and realization of the distance learning bachelor study in the field of Finance, Banking and Investment"** is carried out [1]. The project objectives are:

- *Introduction of the distance on-line bachelor study at the Faculty of Economics*

- *Enhancement of quality of existing external study at the Faculty of Economics*

Project is divided into three phases:

- Introduction of the distance learning materials and dual type of study (face to face and distance learning) from first to third grade of existing external study (May 2003 – February 2005)
- Introduction of the distance learning study for all bachelor subject (May 2003 – February 2007)
- Transition to fully on-line distance learning bachelor study (May 2003 – September 2007)

The project includes all necessary part of distance learning – creation of right distance materials, training the stuff (authors of materials, lectors, tutors, etc.), marketing strategy, and system of the administrative support, quality management system, student support system, and study rules for the students, etc. In September 2005 first students of distance learning are attended the bachelor study program. First outcomes will be known in November 2005 after graduating first course – Informatics.

2.2 Novitech Partner Ltd., Teledom, Educational centre, Košice

The main objective of the e-learning centre of host organization NTP is to provide services in education of adults using progressive information and communication technologies for the sphere managerial, IT skills, languages and also academic. In general the system supports the communication among persons (members of the working group), as a rule geographically remote (mobile) but working on joint activities and project.

The aim of this program is to provide managers in various positions with a specialized complementary education through presentation courses in order to be able to apply the acquired knowledge and intellectual skills in their daily work.

NTP offers Internet on-line English courses as the first one in Slovakia. Based on experiences with this sort of on-line education we co-operate with Trinity System A.S. in the Czech Republic, operating www.anglictina.com portal and also with the company EF Englishtown from USA by the education portal www.englishtown.com.

With the E-learning centre NTP has experiences with in the realization of educational programs by e-learning oriented on the sphere managerial skills and knowledge like e-business, e-banking, marketing, management. These e-learning products have been created with an active partnership with the regions prestige economy and technical universities.

The e-learning centre of the NTP created the e-learning products for the universities, the SME clients and also for large companies like U.S.Steel in the region.

The NTP active used the several learning management systems: The Blackboard, The Intralearn, The eDoceo.

NPT has the richest experiences with the platform eDoceo which has been used from January 2003 till June 2005. Currently we use for e-learning education activities LMS Moodle.

3. LEARNING MANAGEMENT SYSTEMS

The LMS can be defined as tool package assigning management and providing e-learning. Its basic functions, divided into two parts, are:

- **Educational part:**
  - delivering and providing study materials to students
  - monitoring the student's work and testing them
  - evaluation of students
  - communication support for students
- **Management part:**
  - providing information about student's evaluation and problems to teacher
  - registration of all e-learning participants (students, teachers, administrative staff, ...)
  - access control
  - activity monitoring

3.1 eDoceo

Learning Management System eDoceo has been the Web Learning Environment developed by Trask Solution Ltd. LMS eDoceo is appointed for the operation with electronic education programs within company intranet or on internet includes the testing procedure, the evaluation, the monitoring of study results, and the certification of graduates.. LMS eDoceo can be run on the personal database or other ERP systems (HR modul).

The system has been developed for Czech language surround but concurrently works at English and Slovak version (the possibility a next language mutation). It has been built on the e-learning standards and norms for course development (IMS, AICC, SCORM) and open internet technologies like Java, XML. The system has been supported with native application Author. The software Author has appointed for development of scenarios, quizzes and course syllabus.

The eDoceo support IMS, AICC, SCORM standard for communication with realized e-learning courses. This process requires the observance of standards from providers of courses (especially a part CBT for CMI communication) for supported courses.

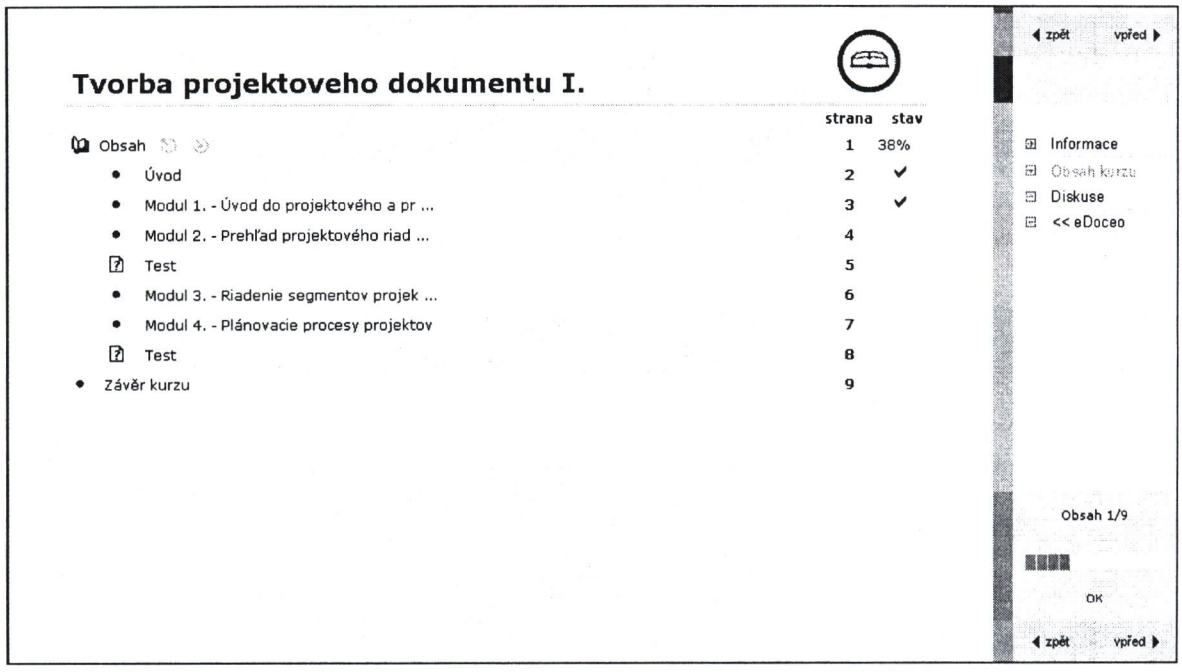


Fig.1 sample of eDoceo course window

Instalation and special requirements:

LMS eDoceo is a multiplatform system. It means the full compatibility with Microsoft as well as Unix (Linux) oriented environments using modern open technologies like J2EE. LMS eDoceo (JAVA BASED application) uses for operation the four basic functions IT server:

- Operation System (MS Windows, Linux, AIX, Unix ... AS 400, Sun Solaris)
- Application Server (MS IIS, IBM WAS, ...)
- Database(MS SQL server, Oracle, DB2, ...)
- SMTP server (Websphere, Microsoft Internet Information Services)

Features:

- Administration of students, teachers, tutors, administrators and courses and training modules in multimedia storage.
- Tools for development of sequential audio/video lessons, tests, tasks, announcements, course info.
- Asynchronous communication using internal email, file exchange and discussions.
- Searching, bookmarks, grade-book, management tools (personal tasks, system announcements), etc.

3.2 uLern

The uLern [3] platform is the Web Learning Environment that was developed by I.T.C., Ltd. It is widely used on Technical University of Košice as a support environment for daily students at faculties of the University. The uLern LMS is developed using DHTML technology on client side including JavaScript, DOM, CSS, PHP and MySQL database on server side and JAVA technology for Virtual Classroom and File Transfer.

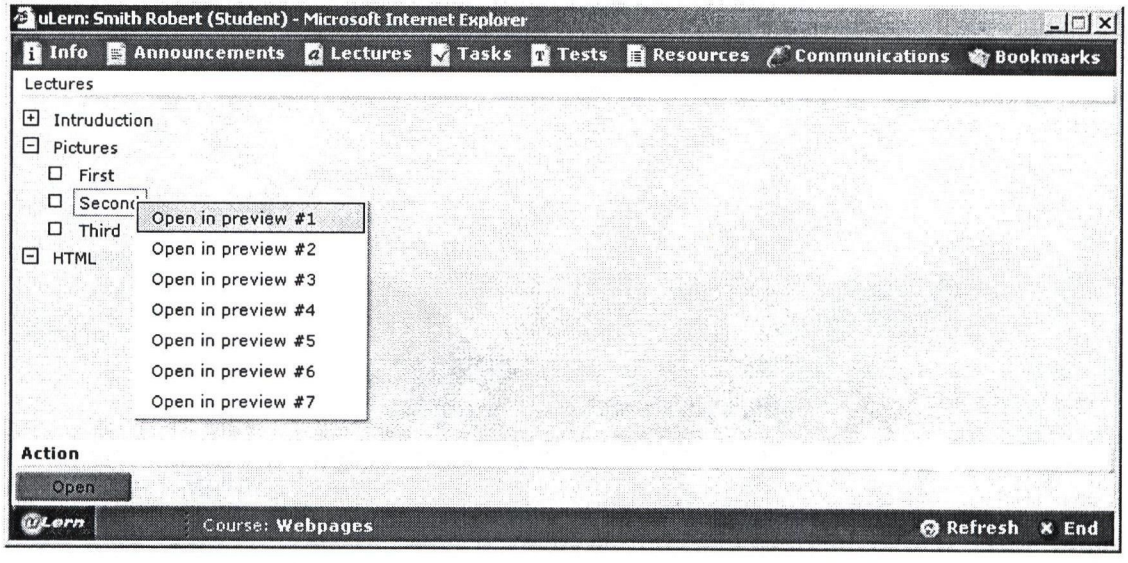


Fig.2 sample of uLern course window

Instalation and special requirements:

For installing uLern LMS there are server side requirements as operating system Windows 2000/2003/XP or Linux, web server Apache including PHP support, MySQL database and Java support (J2RE 1.4.0) for Virtual Classroom and File Transfer and browser (client side) requirements: Microsoft Internet Explorer ver. 5.5 or 6.0, Java support (J2RE 1.4.0), too. Except the requirement on the web browser, on the client side you have to install the Java support, set special java policy file and create special uLern folder used for file transfer.

Features:

- Administration of instructors, students, courses and training modules in multimedia storage.
- Tools for development of sequential audio/video lessons, tests, tasks, announcements, course info, available course calendar and file transfer, wysiwyg html editor.
- Asynchronous communication using internal email, file exchange and discussions.

- Synchronous communication in virtual classroom using chat, streaming of audio/video lectures, whiteboard, synchronous browsing, timed Q&A, and screen sharing.
- Searching, bookmarks, grade-book, management tools (personal tasks, personal calendar, system announcements), etc.

3.3 Moodle

Moodle [4] is a VLE developed and written by a PhD student called Martin Dougiamas. Moodle is written with PHP and MySQL and is based on the open source (GPL) licence agreement. Basically this means Moodle is copyrighted, but that you have additional freedoms. User is allowed to copy, use and modify Moodle provided that you agree: to provide the source to others; to not modify or remove the original license, and apply this same license to any derivative work.

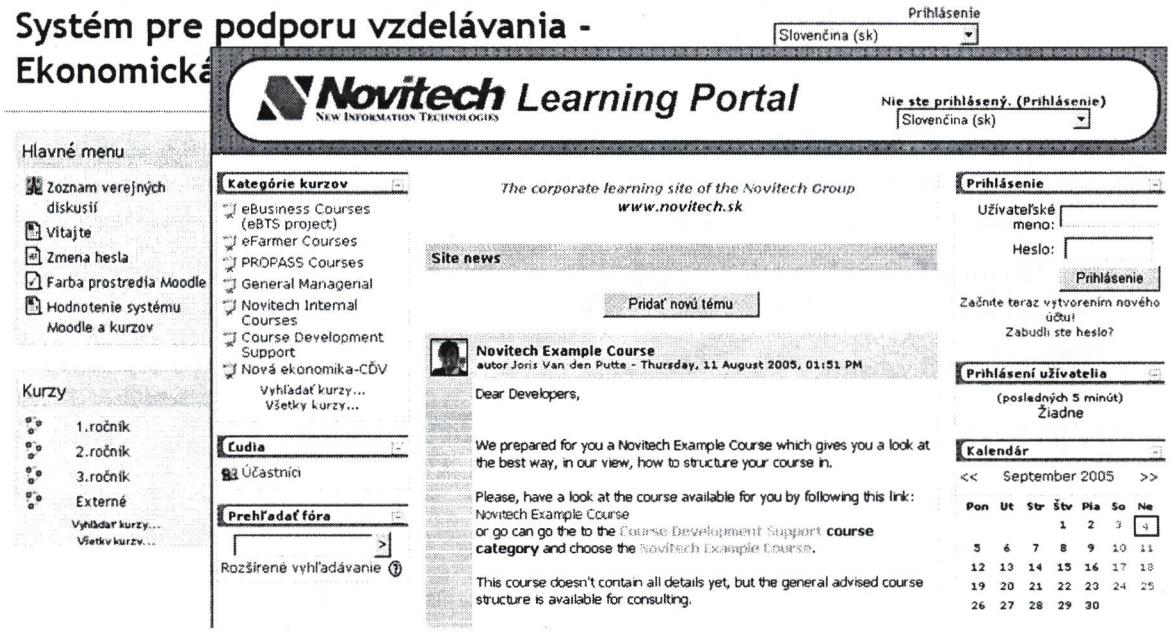


Fig.3 sample of Moodle environment window

Instalation and special requirements:

Installing Moodle was relatively simple. It required the source files to be downloaded from the Moodle website, and then decompressed onto the local hard disk. Once it was successfully saved, the required files were transferred to the web server and then the settings were changed in the configuration files to match our settings at Progress through Training. The installation of Moodle was aided by several automated pages, which speeded up the installation process. Moodle as server will run on any computer that can run PHP, and can support many types of database (particularly MySQL) and as client will run on any computer with web browser.

Features:

- Administration of instructors, students, courses and files in multimedia storage.
- Tools for development of tests, forums, assignments (Ability for trainers to set students any number of assignments, with targeted completion dates), announcements (Users are reminded of forthcoming announcements/ assignments when they first log into the system), course info, available course calendar and file transfer, html editor.
- Asynchronous communication using email, discussions (themed discussion forums), messages.
- Synchronous communication chat.
- Searching, grade-book, management tools (personal tasks, personal calendar, system announcements), etc.
- Customization of layout of the site and theme feature to allow administrators/students to change the look & feel of the VLE without requiring a new style sheet.
- Basic security features to limit customer access to particular courses (teacher can choose the students form list of student or students obtain special course access code).
- 'Journal feature' to allow students to post questions, maintain a course diary, or aid revision.

3.4 Comparison

The presented LMSs were compared from various points of view:

- The environment design
- Education and learning possibilities
- Management and administration

Environment design

It is important to propose study materials through the user-friendly study environment. It means that it must be simple to work with; it needs to have intuitive navigation. From this point of view the uLern has one big disadvantage, each new part of course (lectures, dictionary, students, ...) is opened in new window. The result is disorientation of student, because he has opened too much windows. The second is, that if you want to close opened window correctly, you must do it in down right corner - it is very uncommon.

On the other side, the Moodle LMS has portal-like environment. It is big advantage, because today's information sites have portal structure, that's why users have no problem with navigation and orientation within the system. Basic and key features of system are situated into blocks and their visibility and position within the page can be customized.

When compare the eDoceo with both previous LMS, the study environment of this is very intuitive. The each final user has the own study box with nominate courses. The orientation at the navigation menu of course and portal is very sophisticated but not complicated. The system has write the study progress at a course by %, that's mean the student knows exactly where can continue the study process a next visit LMS.

Educational and learning possibilities

Educational and learning possibilities contain the comparison of delivering and providing materials to students, monitoring the student's work and testing them, evaluation of student, etc.

Once again, the Moodle LMS has an easy way to add new material or activity to course – user choose the activity or source which want to create, fulfill all part of source like name, brief description, etc. and the link is automatically created. After it, there is possibility to move it to another part of course. On the other hand, uLern LMS needs to transfer the files of source to server firstly and only then it is possible to define the position in the course. The pro of this LMS is presentation of course content. All the time is there displayed the list of content of the whole course. It is easy to switch to another lesson.

The Edoceo has used for making the structure of courses the special software Author. The application Author has appointed for development of scenarios, quizzes and course syllabus and content. The final products have accomplished by the standards SCORM, AICC and IMS.

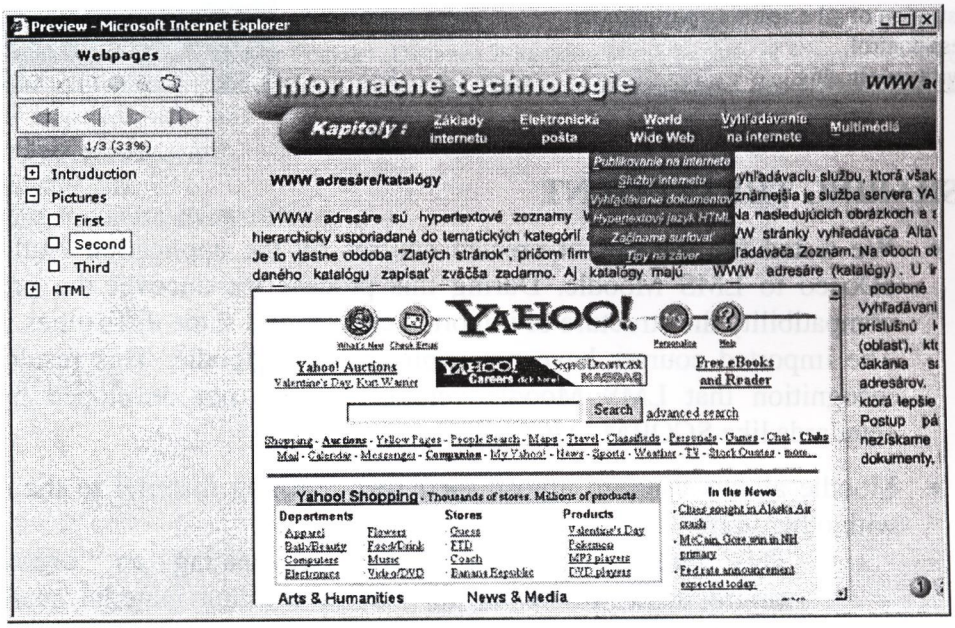


Fig.4 Sample of uLern lesson window

For evaluation of students graduation, the best way is to create some activities like tests, assignments. To other elementary features of LMS belongs the evaluation of these activities and their monitoring. The compared LMSs open up these features. Moreover, Moodle tracks assignments completed and grades allocated by trainers, however, this

information has to be manually entered – the system does not mark, or allocate grades automatically. To pros can be included possibility to restrict the connection to test only for special range of IP addresses, or using access code. Using Moodle teacher can monitor whole activity of students through the course. It is displayed as list of activities of chosen student.

The Moodle, uLern and eDoceo support the communication in the system. The Moodle supports message communication and chatting in the course. The uLern supports the virtual classroom too. It is feature like chat, but there can be used videoconferencing, desktop sharing and so on. Both of systems support the widest way of asynchronous communication – email exchange. The eDoceo doesn't support on-line communication only off-line like e-mail exchange, discussion forum, FAQ etc.

Table 1. Summary of comparison of three LMS – eDoceo, Moodle and uLern

	eDoceo	Moodle	uLern
Instalation	+	+	-
Environment design	+	+	-
Educational part:			
delivering and providing study materials to students	+	+	+/-
monitoring the student's work and testing them	+	+	+/-
evaluation of students	+	+	+
communication support for student's	+	+	+
Management part:			
providing information about student's evaluation and problems to teacher	+	+	+
registration of all e-learning participants	+	+/-	+
access control	+	+	+
activity monitoring	+	+	-

4. SHARING THE CONTENT

- We tested the transfer of created courses by the application Author for LMS eDoceo to LMS Moodle. During this process we uncover the problems with compatibility the structure of imported courses and same difficulties with content. The imported courses have not applicable for Moodle. This result support the recognition that LMS Moodle version 1.5 has not developed by e-learning standards like SCORM, AICC, IMS.
- Moodle allows users to upload their own learning material to the site to share with other users
    - This feature could be useful in creating an 'organic learning environment' where the amount of learning material available on the system grows without administrator intervention
    - However, there could be several copyright implications associated with the type of material that users would upload, so regular monitoring would be required by PTT administration

5. CONCLUSION

As can be seen from the features list made above for each of the three systems there are various benefits/disadvantages incurred when selecting any one VLE over another. Creating a new course can be as simple as designing several different 'pages' (Moodle) of content and then organizing them so that they flow in a logical manner.

In addition to these 'course slides' consultants can then assign specific resources to accompany each slide. In Moodle and uLern there would not be a requirement for HTML knowledge when authoring content. The collaborative features included in the learning environments vary in both complexity and usability. For communication, each system requires a valid e-mail address to be collected from each user so communication via external e-mail systems would be entirely possible.

On an authoring note, Moodle's interface for creating course content was the most intuitive, and featured a very simple to use What You See Is What You Get (WYSIWYG) editor. The use of this editor would remove the need for the consultants to any Internet display languages such as HyperText Mark-Up Language (HTML), which would significantly reduce the development time required to start creating course content.

Knowing our staff possibilities, it was necessary to choose the LMS which is the easiest for content creation and course administration – the Moodle is winner. Today we train our teachers to this LMS and start to use it for distance learning. The project outputs will by soon presented.

REFERENCES

[1] Nataša Urbančíková, Vladimír Penjak: „Distance Education at the Faculty of Economics, Technical University of Košice", Proceedings of ICETA 2003 conference, 11-13 September 2003, Kosice, Slovak Republic, 485 – 487

[2] http://www.edoceo.cz/en/

[3] http://ulern.com/

[4] http://www.econtent.lu/moodle/

EU FP6 SSA Project INCO-CT-2003-003401

Generic Issues of Knowledge Technologies

ISBN 954 - 91700 -2 0