DELOS
Network of Excellence on Digital Libraries

DELOS Research Activities 2005

Editor: Costantino Thanos (ISTI-CNR)

July 2005

DELOS: a Network of Excellence on Digital Libraries
www.delos.info

Information Society
Technologies

Sixth Framework Programme

DELOS

Network of Excellence on Digital Libraries

Research Activities 2005

Information Society Technologies

Table of Contents

Introduction

Digital Libraries (DL) support the specialized needs of very diverse technologies and applications, from cultural heritage to general science, health, government, and education. After approximately ten years of development, they have moved far beyond any connotations of the term "Library", to also encompass Digital Archives and Museums and now have functionality to deal with multimedia objects often with embedded general knowledge, semantics, and behaviour. The potential exists for Digital Libraries to become the universal knowledge repositories and communication conduits of the future, a common vehicle through which information in all forms can be accessed, discussed, and enhanced. Digital Libraries may thus become the strongest shield of humanity protecting its historic, cultural and scientific artifacts from time, natural disasters and human malice. To fulfill their new roles as universal knowledge infrastructures, Digital Libraries require research in several new key areas pointing to the development of:

- user-centred system design methodology
- pro-active systems with functionality that facilitates collaboration, communication, and information creation
- generic Digital Library Management Systems that provide basic system infrastructures that can be used to implement application specific digital libraries incorporating context-specific services.

The DELOS Network of Excellence in Digital Libraries intends to advance the field in these new and exciting directions, with the aim of progressing to the development of the next-generation Digital Library system. To this end, DELOS coordinates a joint programme of activities of the major European teams working in digital library related areas. The objective is to develop dynamic ubiquitous knowledge environments, which will transform research, and education at all levels by collecting, organizing and making publicly accessible on-line vast quantities of information. The ultimate goal is to provide access to human knowledge from anywhere and any time and in an efficient and user-friendly fashion. DELOS also aims at disseminating knowledge of digital library technologies to many diverse application domains, by providing access to technological know-how, services, test-beds, and the necessary expertise to facilitate their take-up.

The research activities of DELOS have been organized in seven clusters:

- Digital Library Architecture
- Information Access and Personalization
- Audio/Visual and Non-traditional Objects
- User Interfaces and Visualization
- Knowledge Extraction and Semantic Interoperability
- Preservation
- Evaluation

In the following pages the major projects being carried on in each cluster are briefly described. They have also beeen shown in a poster sesssion held at ECDL 2005 in Vienna. For additional information about DELOS please visit www.delos.info. For further information about specific research projects, please get in contact with the project coordinators.

Digital Library Architecture

Cluster objectives

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.

Peer-to-peer architectures allow for loosely coupled integration of information services and sharing of information such as recommendations and annotations. Different aspects of peer-to-peer systems (e.g. indexes, and P2P application platforms) must be combined. Grid computing middleware is needed because certain services within digital libraries are complex and computationally intensive (e.g., extraction of features in multimedia documents to support content-based similarity search or for information mining in bio-medical data). The service-oriented architecture provides mechanisms to describe the semantics and usage of information services. Moreover, it supports mechanisms to combine services into workflow processes for sophisticated search and maintenance of dependencies.

The main objective of this cluster therefore is the conceptual and experimental evaluation of the impact of these three main directions to a digital library architecture. A thorough evaluation of existing approaches shall reveal the advantages and disadvantages of either approach. For a quantification of the evaluation, benchmarks have to be developed jointly. In particular, this work package is focussing on the following issues:

- Exploring new approaches to the architecture for an intelligent management of digital libraries
- Enabling the coordinated development of information architectures by an adoption of a set of common standards and protocols
- Managing information dynamics and mobility

Cluster activities

To achieve the goals sketched above, the following activities are being carried out:

- Organisation of workshops on digital library architectures
- Developments of surveys that collect most significant contributions of peer-to-peer data management, grid computing, and service-orientation for digital library architectures
- Perform a comparison and feasibility study on the adoption of a set of common standards and protocols
- Evaluation of approaches to connection management and information synchronisation
- Development of a benchmark for the evaluation of digital library architectures
- Running experiments to find out the strengths and weaknesses of different architectures

Cluster coordinator

Hans-Jörg Schek, UMIT, University for Health Informatics and Technology Tyrol

A Reference Model for Digital Library Management Systems

Donatella Castelli, donatella.castelli@isti.cnr.it
Istituto di Scienza e Tecnologie della Informazione (ISTI-CNR)
Yannis Ioannidis
University of Athens
Seamus Ross
HATII- University of Glasgow
Hans Joerg Schek, Heiko Schuldt
University for Health Informatics and Technology Tyrol

Keywords
H.3.7: Digital Libraries, H.1.1: System and Information Theory

Objectives

Despite the large number of software tools named "digital library systems", there is no agreement yet on what Digital Library [Management] Systems (DLMSs) are and on which functionality they must provide. Existing systems are heterogeneous in scope and focus on very different aspects and functionality. These systems range from digital objects/metadata repositories, reference linking systems, archives and commercial systems that provide administration functions, to complex systems (mainly developed in research environments) that integrate advanced digital library services.

Research on DLMSs concerns several different areas. It is often difficult to compare or combine the results achieved in these areas since it is not always clear how they are related, and how they can impact on or constrain one another. This fragmentation of results hinders the embedding of new research achievements into real world systems.

These problems have a common origin: *the lack of any agreement on the foundations for Digital Library Management Systems*.

The objective of this DELOS activity is to define a DLMS reference model, i.e. a formal and conceptual framework describing the characteristics of this particular type of information system. The model will exploit the understanding of the architecture and functionality expected from an operational DLMS which has been acquired by DELOS research groups over the years. This model will identify and characterize key concepts of a DLMS, such as the information space, documents handled, user profile, services, architecture, etc.

The reference model activity, begun by a core group of DELOS members, will soon be taken up by other DELOS groups, working from the perspectives of different domains. Liaisons with similar activities carried out by other research groups and initiatives at international level will be established in order to achieve a global and stable level of consensus on the model.

Perspectives

The problem of defining a DLMS can be approached from at least three perspectives, according to the different classes of actors involved:

1. *DL end-users* – employ the DL functionality managed by the DLMS. These users include authors submitting the content, information seekers and library administrators.
2. *System administrators* – set-up the DL by installing and configuring the DLMS and maintain the DLMS created instance according to the functional needs of the DL users.
3. *Software engineers* - enrich the system by adding other services that satisfy the specific needs of particular classes of applications.

Each class of actors is looking for a DLMS reference model that satisfies its own particular requirements:

1. DL users: *DL description* – What functionalities are provided by the DL? What information space can be accessed? What query language is used? Which metadata formats does the search service return? Does the DL ensure persistent availability?
2. System administrators: *DLMS selection* – What functionalities does the DLMS support? What type of information space can it manage? How does it allow the content to be ingested? What are the configuration parameters? Which architecture and distribution are permitted? What are the deployment constraints among the services that implement the required functionalities? How does it scale down**?**
3. Software engineers: *DLMS extension* – What services are offered? What functions do they implement? What rules must be satisfied by any new service? Which communication protocol is used?

Highest level concepts

The DELOS DLMS Reference Model aims at providing a representation which characterizes existing and future DLMSs from at least the three perspectives listed above. It will introduce the DLMS concepts, the relationships between these concepts, and the constraints that hold among them. It will also prescribe aspects that are mandatory for this type of information system. Figure 1 represents the highest level concepts of this model: *Information Space* is the entry point for all the concepts related to the content that is managed and disseminated by the DLMS, e.g. collections, document model, metadata, ontologies; *User* is the root for concepts like roles, communities, profiles, rights, etc., that represent aspects of the DL users; and, finally, *Functionality* is the entrance to that part of the model which concerns key concepts like mandatory services, architecture organization and quality of services.

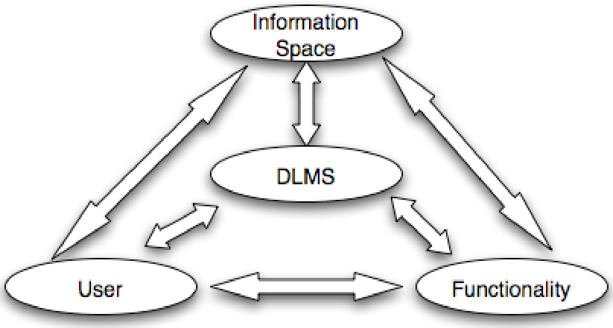


Figure 1 – The highest level DELOS DLMS Reference Model concepts

The first step in the definition of the DELOS Reference model will be to identify the most general representation elements belonging to the three perspectives illustrated and, then, to focus on each of them separately in order to produce three consistent DLMS models that satisfy the representation needs discussed above.

Management of and Access to Virtual Electronic Health Records

Robert Penz, Raimund Vogl
Health Information Technologies Tyrol (HITT), Innsbruck, Austria
Wilhelm Hasselbring, Ulrike Steffens
Kuratorium OFFIS, Oldenburg, Germany
Charalampos Dimitropoulos, Yannis Ioannidis
National and Kapodistrian University of Athens, Greece
Jens-Peter Dittrich, Peter Fischer, Donald Kossmann
Swiss Federal Institute of Technology (ETH) Zürich, Switzerland
Christoph Langguth, Thomas Schabetsberger, Hans-Jörg Schek, Heiko Schuldt, Michael Springmann
University for Health Sciences, Medical Informatics and Technology (UMIT), Hall in Tyrol, Austria

Introduction

eHealth digital libraries contain electronic artifacts that are generated by different healthcare providers (family doctors, laboratories, hospitals, etc.). An important observation is that this information is not stored at one central instance but rather under the control of the organization where data has been produced. The electronic health record of patients therefore consists of a set of distributed artifacts and cannot be materialized for organizational reasons. Rather, the electronic patient record is a virtual entity and has to be generated by composing the required artifacts each time it is accessed (Figure 1). The virtual integration of an electronic patient record is done by encompassing services provided by specialized application systems (e.g., CIS – clinical information systems, or PACS – picture archiving and communication system) into processes. A process to access a virtual electronic health record encompasses all the services needed to locate the different artifacts, to make data from the different healthcare providers available –given appropriate authorization and authentication–, to perform the format conversions needed, and to present the (possibly anonymized) result to a user (i.e., patient X accesses his virtual health record via web; it contains collected medical documents from different health care institutions. Or doctor W, a specialist for internal medicine, retrieves the automatically integrated information related to the coronary heart disease of his patient Y from several health care institutions). In addition to the patient-centric access to virtual health records, also disease-specific applications can be supported by means of processes. These disease-specific applications allow for epidemiological studies across a set of patients, comparisons, identification of similar diagnoses, etc. (i.e., medical scientist M would like to identify all patients that have similar pathological deviations in the X-ray of their lung than patient Z, for whom SARS has been diagnosed. Or the Ministry of Health needs a statistical overview about diagnoses and treatments during the last three years, in order to control the health system and –for civil protection– to prevent from dangerous sanitary situations).

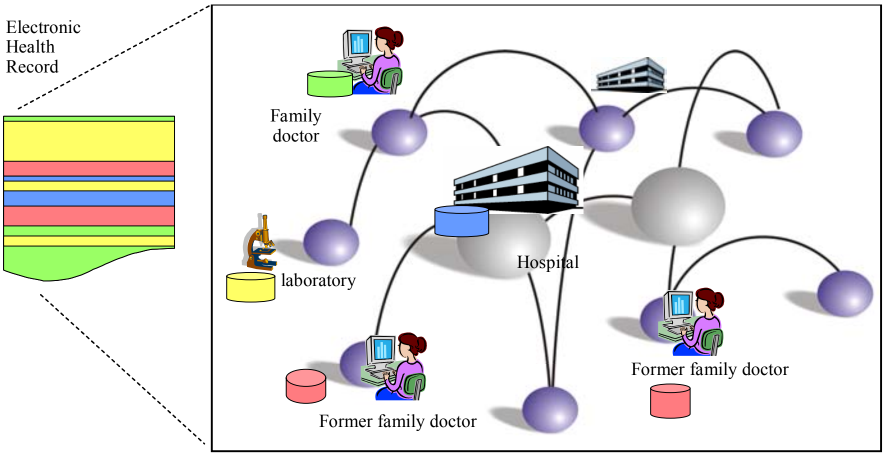


Figure 1: Electronic Health Record

Research Objectives

The realization of these goals first requires the availability of appropriate (web) services in order to access relevant data managed by specialized applications which are hosted by different healthcare organizations. In addition, common standards have to be supported to integrate these legacy applications (e.g., the PACS application where the X-rays of Y and Z are stored) or, alternatively, dedicated services are needed to transform the format of the data retrieved from one application so as to make is available for other, subsequent services. Second, an infrastructure to combine these services into processes is needed that is highly dependable and reliable. Physicians must be given the guarantee that the system and data is always available (i.e., by means of replication) and that processes come to a well-defined end (e.g., by collecting all pieces of information that are of interest), even in case of failures. Third, the infrastructure has to provide a high degree of scalability, and to efficiently schedule the access to computationally intensive services by applying sophisticated load balancing strategies using grid technology. Physicians need information immediately, especially for patient-centric queries (but also for certain disease-specific queries), in order to make vital decisions. Hence, long response times due to a high system load cannot be tolerated. Consider the first disease-specific scenario above (the lung X-ray of patient Z) where similarity search across a potentially large set of documents is needed. In order to support this search, feature extraction has to take place for all documents/images, nearest neighbors have to be determined, etc. All these steps require significant computing power and should not be limited to the organization where the images are stored. Rather, additional feature extraction services should be installed automatically at hosts which currently feature a low load. Finally, the infrastructure has to allow for the transparent access to distributed data by means of appropriate (peer-to-peer) indexing techniques that avoid single points of failures as well as censorship and that, at the same time, preserve the privacy of data.

Expected Results

The main goal of this DELOS task is to identify, design, and build demonstrators for the basic building blocks needed to access virtual electronic health records, i.e., locate the different artifacts, make data from the different healthcare providers available, perform the format conversations needed, and present the result to a user. In particular, the basic building blocks to access distributed artifacts and to intelligently search within a set of these artifacts will be identified. Moreover, a dependable platform that supports the integration of these building blocks into processes will be provided, thereby realizing a system to manage virtual electronic patient records. Finally, sample building blocks and processes in combination with specialized application systems made available by HITT will be implemented. Examples of these building blocks are algorithms for format conversions (e.g., specialized formats that can be found in healthcare applications like DICOM or HL7), services for similarity search, relevance feedback, or replication.

Project Participants

The authors of this extended abstract are the members of this project. The partner institutions are:

- Health Information Technologies Tyrol (HITT), Innsbruck, Austria
- Kuratorium OFFIS, Oldenburg, Germany
- National and Kapodistrian University of Athens, Greece
- Swiss Federal Institute of Technology (ETH) Zürich, Switzerland
- University for Health Sciences, Medical Informatics and Technology (UMIT), Hall in Tyrol, Austria

References

[BHN 04] L. Bischofs, W. Hasselbring, H. Niemann, H. Schuldt, M. Wurz: *Verteilte Architekturen zur intra- und inter-institutionellen Integration von Patientendaten*. In: Proceedings of GMDS 2004, Innsbruck, Austria, September 2004. In German.

[FGK 03] D. Florescu, A. Grünhagen, D. Kossmann: *XL: a Platform for Web Services*. CIDR 2003

[GGH 00] J. Grimson and W. Grimson and W. Hasselbring: *The SI challenge in health care*. In: CACM, 43(6), June 2000, pp. 48-55.

[Has 97] W. Hasselbring: *Federated Integration of Replicated Information within Hospitals*. In: International Journal on Digital Libraries 3(1), Nov. 1997, pp192—208.

[Has 00] W. Hasselbring: *Information Technology Standards and Standardization: A Global Perspective*. In: K.Jakobs (Ed.) The Role of Standards for Interoperating Information Systems, Idea Group Publishing, 2000.

[PH 04] S. Pedersen and W. Hasselbring: *Interoperabilität für Informationssysteme im Gesundheitswesen*. In: Informatik Forschung und Entwicklung 18(3), 2004. In German

[SAG 05] T. Schabetsberger, E. Ammenwerth, G. Goebel, G. Lechleitner, R. Penz, R. Vogl, F. Wozak: *What are Functional Requirements of Future Shared Electronic Health Records?* MIE 2005.

[SSS 04] M. Springmann, H.-J. Schek, H. Schuldt: *Kombination von Bausteinen zur ähnlichkeitsbasierten Suche in elektronischen Multimedia-Patientenakten*. In: Proceedings of GMDS 2004, Innsbruck, Austria, September 2004. In German.

Integration of Data Stream Management into an eHealth Digital Library

Charalampos Dimitropoulos, Yannis Ioannidis
National and Kapodistrian University of Athens, Greece
Jens-Peter Dittrich, Peter Fischer, Donald Kossmann
Swiss Federal Institute of Technology (ETH) Zürich, Switzerland
Gert Brettlecker, Hans-J. Schek, Heiko Schuldt
University for Health Sciences, Medical Informatics and Technology (UMIT), Hall in Tyrol, Austria

Introduction

Recent trends in ubiquitous and pervasive computing, together with new sensor technologies, wireless communication standards, powerful mobile devices and wearable computers strongly support novel types of applications. Especially in healthcare, tele-monitoring applications will make use of this new technology in order to improve the quality of treatment and care for patients and the elderly. In particular, if we consider our aging society, the amount of elderly people suffering from one or more chronic diseases is increasing. Tele-monitoring applications enable healthcare institutions to take care of their patients while they are out of hospital, which is especially useful for managing various chronic diseases as well as for measuring the effects of treatments under real-life conditions. A similar but even more comprehensive application is the support for the elderly and for people with cognitive disabilities living at home. Here, not only physiological sensors and data are relevant for tele-monitoring but also context information, e.g., information on the current activities of a person, where she or he is located, etc. The term eInclusion addresses the extension of tele-monitoring to non-physiological information.

Research Objectives

Continuous data streams generated by (wearable) sensors have to be processed online in order to detect critical situations. For this purpose, usually different streams (generated by different types of sensors) have to be combined (e.g., jointly consider oxygen saturation, ECG signals and blood pressure). This is done by making use of specialized operators (e.g., signal pre-processing, filtering, joining different streams, aggregating stream data, etc.) – see Figure 1. An infrastructure for tele-monitoring has to be able to combine these operators in an application-specific way. However, in addition to the stream operators, also traditional discrete (web) services, e.g., services that do not operate on continuous input data, have to be integrated. These services are needed, for instance, in order to notify emergency physicians in very critical cases, to notify neighbors about abnormal situations, but also to write information aggregated from stream data back to the electronic health record of a patient. The latter is very important to integrate data stream management into an eHealth digital library.

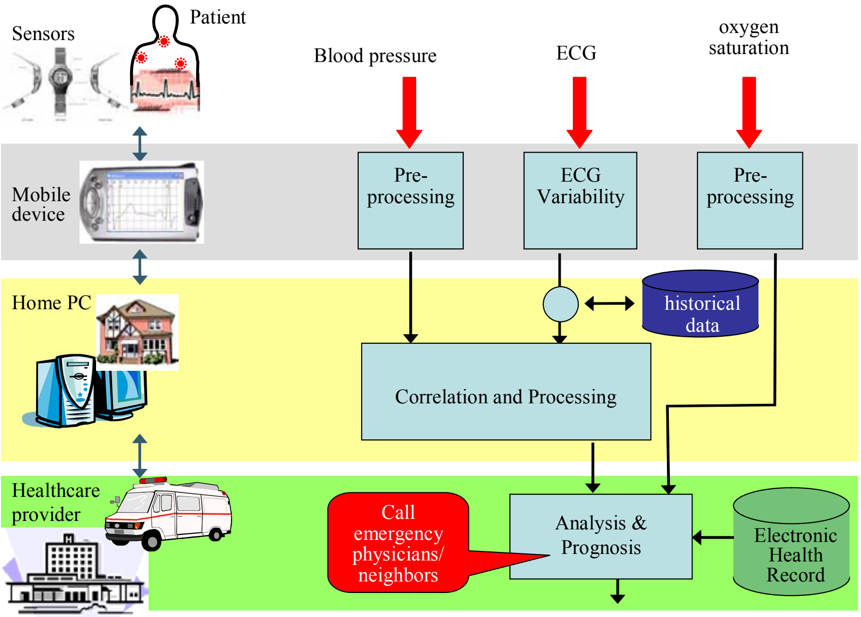


Figure 1: A Sample Data Stream Process for Health Monitoring

Expected Results

The main goal of this task is to identify, design, and build demonstrators for data stream operators, i.e., the continuous processing of combined streaming data generated by different types of sensors. Appropriate web services are also needed in order to process and store the results and aggregates of stream operators. Finally, appropriate notification mechanisms are needed in order to inform a patient as well as healthcare providers on critical deviations of her/his health state. These operators and services will be combined in an infrastructure for stream processing in healthcare.

In particular, this task started with an evaluation of applications where the combination of data stream operators (continuous processing of streaming data) and web services (discrete invocation of application functionality) is an inherent requirement. Based on this evaluation, selected streaming operators (e.g., for searching outliers in a data stream or for joining parallel data streams) will be implemented and notification mechanisms will be provided. In addition, a prototype implementation of an infrastructure for workflow processes including stream processing supporting the integration of stream operators and web services will be provided. This allows for the insertion of the results of data stream processing in electronic health records, thereby tightly integrating data stream processing and eHealth digital libraries.

Project participants

The authors of this extended abstract are the members of this project. The partner institutions are:

- National and Kapodistrian University of Athens, Greece
- Swiss Federal Institute of Technology (ETH) Zürich, Switzerland
- University for Health Sciences, Medical Informatics and Technology (UMIT), Hall in Tyrol, Austria

References

[BSS 05] G. Brettlecker, H. Schuldt, H.-J. Schek: *Towards Reliable Data Stream Processing with OSIRIS-SE*. In: Proceedings of the 11th German Database Conference (BTW'2005), pages 405-414, Karlsruhe, Germany, March 2005.

[BSS 04a] G. Brettlecker, H.-J. Schek, H. Schuldt: *Information Management Infrastructure for Telemonitoring in Healthcare*. In: Proceedings of GMDS 2004, Innsbruck, Austria, September 2004.

[BSS 04b] G. Brettlecker, H. Schuldt, R. Schatz: *Hyperdatabases for Peer-to-Peer Data Stream Processing*. In: Proceedings of the 2nd International Conference on Web Services (ICWS'2004), pages 358-366, San Diego, CA, USA, July 2004. IEEE Computer Society.

[DFK 05] J.-P. Dittrich, P. M. Fischer, D. Kossmann: *AGILE: Adaptive Indexing for Context-Aware Information Filters*. In: Proceedings of SIGMOD'05, Baltimore, Maryland.

[FGK 03] D. Florescu, A. Grünhagen, D. Kossmann: *XL: a Platform for Web Services*. CIDR 2003

[FK 03] P. M. Fischer, D. Kossmann: *Batched Processing for Information Filters*. In: Proceedings of ICDE'05, pp 902-913, Tokyo, Japan, April 2005.

[WBS 04] M. Wurz, G. Brettlecker, H. Schuldt: *Data Stream Management and Digital Library Processes on Top of a Hyperdatabase and Grid Infrastructure*. In: Pre-Proceedings of the 6th Thematic Workshop of the EU Network of Excellence DELOS: Digital Library Architectures - Peer-to-Peer, Grid, and Service-Orientation (DLA 2004), pages 37-48, Cagliari, Italy, June 2004, Edizioni Progetto Padova.

Information Access and Personalization

Cluster objectives

Information stored in digital libraries needs to be accessed, integrated and individualized for any user anytime and anywhere in possibly multiple comprehensive and efficient ways. Within Delos, Information Access in Digital Libraries is studied from three different aspects:

Information Access: interaction with a single information provider. Information stored in a source comes in different types and formats, each one with its own characteristics and particularities. Organization of data within an individual source and efficient and effective search are the key issues and are actually highly interrelated to each other. Different approaches exist but there is a general trend towards richer representations and languages both at the structural and at the semantic level.

Information Integration: interaction with multiple information providers. Integrated access of different sources presents specialized problems due to information heterogeneity, redundancy etc. Issues such as source selection and results fusion must be considered under different possible settings. Data provenance is often crucial to the trust that is placed in data, hence it should be managed based on sound formulation.

Personalization: customization of information and interaction to user. Different users have different characteristics and preferences concerning the information they are interested in seeing when accessing a digital library. Even users sharing a common information need may expect different results, different functionality or different interface. Moreover, the relevant contents and interface of a digital library may be dependent on other factors as well, e.g. device or network-specific.

The cluster's objectives with respect to the aforementioned aspects are the following:

- Promotion of knowledge about available practices in the fields of information access and personalization in digital libraries is the first goal pursued. This will lead to a uniform understanding of problems among researchers.
- Construction of a common, comprehensive framework for information access and personalization approaches is essential. This framework is intended to serve as a reference point for the DL area and to stimulate research.
- Promotion of research on new information access and personalization models and methodologies.

Cluster activities

The cluster's activities with respect to information access, integration and personalization are very coarsely organized into the following categories:

- Collection, study, and comparison of models, languages and algorithms for data, metadata, and queries with respect to information access and integration
- Collection, study, and comparison of user-profile models and various forms of content and interaction personalization
- Integration of the most effective approaches to information access, integration, and personalization and derivation of new ones
- Development of toolkits and systems for purposes of re-use and demonstration of proposed methods and models

Cluster coordinator

Yannis Ioannidis, University of Athens

Advanced Access Structures for Complex Similarity Measures

Sören Balko, UMIT Innsbruck, Soeren.Balko@umit.at
Giuseppe Amato, ISTI-CNR, Giuseppe.Amato@isti.cnr.it
Pavel Zezula, MUNI, Zezula@fi.muni.cz

Research Problem and Objectives

Query processing in non-alphanumeric data stocks, like similarity search in large collections of media objects is often conducted by means of some distance that computes a score which quantifies two object's similarity. This distance may either operate on the objects themselves or on pre-computed feature vectors that express certain characteristics through numeric values. Many indexing proposals exist [5] to support feature-based query processing that rests upon plain distance measures, like L_p-norms. Unfortunately, most of them fail to answer queries efficiently as soon as the dimensionality of the feature vectors exceeds a certain "usability threshold". Therefore, filter-and-refinement approaches, like the VA-file [1], the AV-method [2], the LPC-file [3], and the VA^+-file [4] have been proposed to overcome this problem, frequently referred to as "curse of dimensionality". Filter-and-refinement approaches basically compute some compressed representation (referred to as 'signature') of the feature data. In terms of query processing, a quick-and-dirty scan through the signatures leaves a few candidates remaining for further inspection. Filter-and-refinement approaches have proven to be feasible for arbitrary distributed, high-dimensional data and simple distance measures. However, (1) some similarity measures employ more complex metrics, that are difficult to tackle in a filtering approach, like Earth Movers Distance on Color Signatures [7] and Smith-Waterman local alignment [6] on protein sequence data. Moreover, (2) some distances directly operate on non-numeric media objects, providing no feature extraction facilities. Some proposals are available to generically deal with indexing in metric spaces [8]. In many application scenarios, however, those approaches either (1) suffer from poor pruning power in many data workloads and/or (2) must frequently invoke the distance function which may be computationally expensive.

Quantization Techniques in Metric Indexing

The objective of this proposal is to analyze metric indexing approaches that avoid those drawbacks. In other words, we seek to broaden the filter-and-refinement approach towards metric indexing. The outcome shall be given by an indexing proposal for metric spaces that (1) efficiently supports nearest neighbor, range, and ranking queries, (2) operates on any kind of multimedia data, (3) is generic in the metric distance measure to be employed, and (4) allows for straightforward parallelization of query processing that comes with good scalability. Our approach investigates ways to integrate (1) existing metric indexing approaches, and (2) VA-file like quantization approaches into a novel indexing approach. We intend to tackle two problems of metrics-based similarity retrieval simultaneously. On the one hand, we address the I/O issue which is to reduce the amount of hard disk accesses. On the other hand, we save computational expenses by rarely invoking a possibly time-consuming distance measure between media objects.

We implemented an indexing (Java) and retrieval (mixed Java/C++) prototype framework focusing on the Earth Movers Distance between colour signatures of images. We are working at a refined quantization scheme that yields even smaller approximation errors, permits retrieval parallelization by disjoint index partitioning, obeys a strictly constrained main memory consumption, incorporates various clustering/pivot finding techniques, and considers other data domains and similarity measures.

Similarity search approach in digital library applications exploiting XML encoded metadata

XML is becoming one of the primarily used formats for the representation of heterogeneous information in many and diverse application sectors, such as multimedia digital libraries, public administration, EDI, insurances, etc. This widespread use has posed a significant number of technical requirements to systems used for storage and content-based retrieval of XML data, and many others is posing today. In particular, retrieval of XML data based on content and structure has been widely studied and it has been solved with the definition of query languages such as XPath and XQuery and with the development of systems able to execute queries expressed in these languages. However, many other research issues are still open.

There are many cases where users may have a vague idea of the XML structure, either because it is unknown, or because is too complex, or because many different structures - with similar semantics - are used across the database [9]. In addition there are cases where the content of elements of XML documents cannot be exactly matched against constants expressed in query, as for instance in case of large text context or low-level feature descriptors, as in MPEG-7 visual or audio descriptors. The standardization effort carried-out by MPEG-7, intending to provide a normative framework for multimedia content description, has permitted several features for images to be represented as visual descriptors to be encoded in XML.

We are investigating the architecture of a native XML search engine that allows both structure search and approximate content match to be combined with traditional exact match search operations. Our XML database can store and retrieve any valid XML document without need of specifying or defining their schema. Our system store XML documents natively and uses special indexes for efficient path expression execution, exact content match search, and approximate match search. For instance, in case of an MPEG-7 visual descriptor, the system administrator can associate an approximate match search index to a specific XML element so that it can be efficiently searched by similarity. The XQuery language has been extended with new operators that deal with approximate match and ranking, in order to deal with these new search functionality.

Scalable and Distributed Similarity Search Structures

Centralized metric indexes achieve a significant speedup (both in terms of distance computations and disk-page reads) when compared to a baseline approach, the sequential scan. However, experience with centralized methods reveals a strong correlation between the dataset size and search costs. More specifically, costs increase linearly with the growth of the dataset, i.e., it is practically twice as expensive to compute a similarity query in a dataset of a given size as it would be with a dataset of half that size. Thus, the ability of centralized indexes to maintain a reasonable query response time when the dataset multiplies in size, its *scalability*, is limited.

An important activity of this cluster will be research of scalable and distributed similarity search structures. The work will capitalize on our previous research on centralized structures as well as the GHT* , i.e. the first attempt to make such structures scalable through distribution [10]. It supports similarity search in generic metric spaces, and it is based on the idea of the *Generalized Hyperplane Tree*. The structure allows storing datasets from any metric space and has many essential properties of the distributed and P2P approaches. It is scalable, because every peer can perform an autonomous split and distribute the data over more peers at any time. It has no hotspot, and all peers use an addressing schema as precise as possible, while learning from misaddressing. Updates are performed locally and splitting never requires sending multiple messages to many peers. Finally, every peer can store data and perform similarity queries simultaneously. Though the first results are encouraging, much more research in this direction is necessary.

Conclusions

Our work aims at providing approaches that substantially differs from existing ones to index metric spaces and nicely be integrated into XML search engines. In particular, we seek to exploit the merits of the filter-and-refinement principles, query workload partitioning and distribution, and approximate matches of XML elements to yield fast access structures that permit simple parallelization with good scalability and little communication effort in distributed architectures.

References

[1] R. Weber, H.-J. Schek, S. Blott: *A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces.* 24th Int. Conf. on Very Large Databases (VLDB), New York (1998)

[2] S. Balko, I. Schmitt, G. Saake: *The Active Vertice Method: a Performant Filtering Approach to High-Dimensional Indexing.* Data & Knowledge Engineering 51 (2004), 369-397, Elsevier.

[3] G.-H. Cha, X. Zhu, D. Petkovic, C.-W. Chung: *An Efficient Indexing Method for Nearest Neighbor Searches in High-Dimensional Image Databases.* IEEE Transactions on Multimedia 4(1), 76-87

[4] H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, A. E. Abbadi: *Vector Approximation Based Indexing for Non-Uniform High Dimensional Data Sets.* 9th Int. Conf. on Information and Knowledge Management (CIKM), 2000

[5] C. Böhm, S. Berchtold, D. A. Keim: *Searching in High-Dimensional Spaces – Index Structures for Improving the Performance of Multimedia Databases.* ACM Computing Surveys 33(3), 322-373 (2001)

[6] W. Xu, D. P. Miranker: *A metric model of amino acid substitution.* Bioinformatics 20(8), 1214-1221 (2004)

[7] Y. Rubner, C. Tomasi, L. J. Guibas: *The Earth Mover's Distance as a Metric for Image Retrieval.* International Journal of Computer Vision 40(2): 99-121 (2000)

[8] E. Chavez, G. Navarro, R. Baeza-Yates, J. L. Marroquin: *Searching in Metric Spaces.* ACM Computing Surveys 33(3): 273-321 (2001)

[9] Giuseppe Amato, Claudio Gennaor, Fausto Rabitti, Pasquale Savino, *Milos: A Multimedia Content Management System for Digital Library Applications*, ECDL 2004, Bath, UK, September 12-17 2004

[10] M. Batko, C. Gennaro, P. Zezula: *Similarity Grid for Searching in Metric Spaces.* Digital Library Architectures: Peer-to-Peer, Grid, and Service-Orientation, Post-Proceedings of the 6th Thematic Workshop of the EU Network of Excellence DELOS, LNCS, Springer , to appear.

Application of the P2P Paradigm in Digital Libraries

Vassilis Christophides, FORTH, christop@ics.forth.gr
Timos Sellis, ICCS, timos@dbnet.ece.ntua.gr
Stratis Viglas, UEDIN, sviglas@inf.ed.ac.uk

Keywords
Query Processing(H.2.4.h), Web Search(H.3.0.a), Query Formulation(H.3.3.e), Distributed Systems(H.3.4.b), XML/XSL/RDF(H.3.5.f), Digital Libraries(H.3.7)

Research Problem and Objectives

DELOS envisions the availability of digital content on a global scale through Digital Libraries (DL) that "can be accessed, integrated and individualized for any user anytime and anywhere in possibly multiple comprehensive and efficient ways". A key point in such a vision is the interaction with multiple DL nodes to support integrated access despite their heterogeneity.

This project is exploring the application of the *peer-to-peer (P2P)* paradigm in DL technologies as a means to deal with (a) DL nodes autonomy, (b) decentralized sharing and management of DL data, (c) heterogeneity and (d) highly dynamic DL networks. The advanced structuring and retrieval functionality of *peers* poses new challenges for view integration, query routing and processing over autonomous, distributed and dynamic networks of DLs. Our study aims to address foundational aspects of P2P DLs, namelly: *query reformulation* and *query roaming* in P2P DLs.

Query reformulation in P2P DLs

The framework of traditional data integration for query reformulation based on mapping rules should be re-visited to meet the requirements of P2P DLs, where a DL node can serve both as local data source and as mediator with a global schema. Our proposal will set up the foundations of query reformulation using mappings in schema-based P2P DLs, where the query language will be an XML (Xquery) or an RDF/S (RQL, RVL) one.

Query roaming in P2P DLs

In our P2P DL setting, both content providers and consumers will be treated in the same manner to allow for a common abstraction. Each peer is only aware of a possibly small set of neighbouring peers. Thus, the peer DL resources can be localized by using adequate routing indexes hosted by the peers. Query roaming involves *planning* (i.e., data localization) and *query execution*. Our proposal will build a roaming query engine able to route, plan, and process declarative queries, and adapt to changes in the P2P DL network. Planning and execution will be interleaved, since the precise evaluation steps of the computation and the participating peers cannot be foreseen in advance.

Objectives

The project aims to investigate the peer-to-peer (P2P) resource-sharing paradigm for large-scale distributed Digital Libraries (DL). In particular, we are interested in applying schema-based P2P technology to study the foundations of P2P DLs. The objective is to support decentralized sharing of data and services among numerous autonomous DL nodes, which may employ heterogeneous metadata schemas for describing their information and computational resources. The objectives are:

1. establish logical foundations of query reformulation in P2P DLs,
2. design routing indexes for XML or RDF/S queries and views in P2P DLs,
3. develop methods for P2P DL query processing and optimization.

Expected results

- A simulation prototype of the P2P DL network that meets the aforementioned requirements
- Deliverable: Query Reformulation, Processing and Optimization in P2P DLs.

References

[AH02] K. Aberer, M. Hauswirth. Peer-to-Peer Information Systems: Concepts and Models, State-of-the-art and Future Scenario. ICDE'02 Tutorial.

[CGLR04] D. Calvanese, G. De Giacomo, M. Lenzerini and R. Rosati. Logical Foundations of Peer-to-Peer Data Integration. In Proceedings of the PODS'04 Symposium, Paris, France, 2004.

[CKK+03] V. Christophides, G. Karvounarakis, I. Koffina, G. Kokkinidis, A. Magkanaraki, D. Plexousakis, G. Serfiotis and V. Tannen. The ICS-FORTH SWIM: A Powerful Semantic Web Integration Middleware. In Proceedings of the SWDB'03 International Workshop, Berlin, Germany, 2003.

[HIST03] A. Halevy, Z. G. Ives, D. Suciu and I. Tatarinov, Piazza: Data Management Infrastructure for Semantic Web Applications. In Proceedings of the WWW'03 International Conference, Budapest, Hungary, 2003.

[KAC+02] G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis and M. Scholl RQL: A Declarative Query Language for RDF In Proceedings of the 11th International World Wide Web Conference (WWW 2002), Honolulu, Hawaii, USA, 2002.

[KAM03] A. Kementsietsidis, M. Arenas and R. J. Miller Mapping Data in Peer-to-Peer Systems: Semantics and Algorithmic Issues In Proceedings of the ACM SIGMOD'03 International Conference, San Diego, California, USA, 2003.

[KC04] G. Kokkinidis and V. Christophides, Semantic Query Routing and Processing in P2P Database Systems: ICS-FORTH SQPeer Middleware. In Proceedings of the P2P\&DB'04 International Workshop (in conjuction with EDBT'04), Heraklion, Crete, Greece, 2004. (pdf, bib)

[KCD+05] G. Kokkinidis, V. Christophides, T. Dalamagas and S. Viglas. Query Processing in P2P Database Management Systems: A State-of-the-Art. Technical Report (deliverable) for IST-FP6, DELOS Network of Excellence, NoE G038-507618, Apr 2005.

[Len02] M. Lenzerini. Data Integation: a Theoretical Perspective. In Proceedings of the ACM PODS’02 Symposium, 2002.

[Len04] M. Lenzerini. Principles of Peer-to-Peer Data Integration. In Proceedings of the DiWeb'04 International Workshop, Riga, Latvia, 2004.

[MH03] J. Madhavan, A. Halevy. Composing Mappings among Data Sources. In Proceedings of the VLDB’03 Conference, Berlin, Germany, 2003.

[NWQ+02] W. Nejdl, B. Wolf, C. Qu, S. Decker, M. Sintek, A. Naeve, M. Nilsson, M. Palmer and T. Risch, EDUTELLA: A P2P Networking Infrastructure Based on RDF. In Proceedings of the WWW'02 International Conference, Honolulu, Hawaii, USA, 2002.

[SKD+05] L. Sidirourgos, G. Kokkinidis, T. Dalamagas and V. Christophides. Indexing Views for Query Routing in Schema-based P2P systems. Submitted to the International Conference on Data Engineering (ICDE’06), Jun 2005.

[TH04] I. Tatarinov, A. Halevy. Efficient Query Reformulation in Peer-Data Management Systems. In Proccedings of the ACM SIGMOD Conference, 2004.

Personalized Query Routing inPeer-to-Peer Federations of Digital Libraries

Matthias Bender, Yannis Ioannidis, Henrik Nottelmann, Hans-Jörg Schek, Gerhard Weikum, Pavel Zezula, Christian Zimmer

Contact Author: Gerhard Weikum, Max-Planck Institute for Informatics, weikum@mpi-sb.mpg.de

Research Objectives

The peer-to-peer (P2P) paradigm is an intriguing approach for coping with dynamically evolving federations of loosely coupled digital libraries. In addition to the libraries, user agents with powerful personalized tools may participate as peers, too. Such a P2P system promises unlimited scalability, robustness to failures, fluctuation, and load dynamics, and also much reduced vulnerability to attacks and information manipulation. Peers remain autonomous and participate in the P2P system and collaborate on behalf of user requests at their discretion. In a typical scenario, a user peer would utilize its local profile to compile an information demand and issue queries to the library peers that most suitable for the given demand. The decision about where a query should be sent is known as *query routing* or *database selection* or *resource selection*.

Database selection has been intensively studied in the literature on distributed IR and metasearch engines [Ca00, LC03, NF03, MYL03]. However, this prior work assumed a static architecture with a fairly small number of data sources that do not change over time. The large scale and high dynamics of a P2P system poses much more challenging issues. For example, the dissemination, maintenance, and effective use of statistical summary information about the peers' information contents entails difficult questions. Moreover and most crucial to this proposal, prior work has paid little attention to personalized strategies for query routing. In P2P systems every peer can share (some explicitly disclosed fraction of) the users' access patterns, query logs, and further user-behavior information with other peers; this provides great opportunities for advanced personalization of information selection and query execution, leading to better search result and information quality [Be04a, NZ95]. The objective of this proposal is to investigate how this potential can be effectively exploited.

Recent work has started studying how user bookmarks can be utilized for better query routing in P2P search engines [Be04b]. The work proposed here will go one step further and consider also query logs, click streams, and further long-term information about user behavior that reflects the user's (or user community's) information needs and biases [KI05, LW04, PI04, TK04]. In this context, information requests posed by the user can vary from keyword queries to structured queries in SQL or XQuery and may also include entire sessions of searching, browsing, annotation, classification, and refinement steps.

The work will specifically focus on how to exploit the above kind of user-behavior information for the purpose of query routing in the distributed P2P federation. Personalization is, of course, also an issue for the query execution on a single digital library; this will not be considered here. The following technical problems will be addressed:

1. How can and should user-behavior information be represented in a compact form, and where and how should it be made available - on demand or by proactive dissemination in a "gossiping" style?
2. Once a compact representation is found, how is it used for deciding to which peers an information request should be sent? How do we choose the most suitable peers for a given query - in terms of similarity to the combination of query and user profile, and at the same time avoid that queries are redundantly sent to many peers that provide more or less the same information?
3. What kinds of statistical models are most appropriate for matching user profiles and queries with the contents of peers? How expensive is their computation? How do we cope with the tradeoffs in expressiveness vs. complexity?

Approach and Expected Outcome

Most of the research work is conceptual in nature: designing models and strategies for personalized query routing. In addition, it is planned to implement the most promising strategies in a demonstrator testbed. The Minerva prototype system [Be05], developed at Max-Planck Institute for Informatics, can be used for this purpose; it can be easily installed at the sites of the partners that participate in this work. A critical issue to be addressed is also the evaluation of strategies. A comprehensive experimental evaluation is beyond the scope of this work, but the conceptual work should already consider this later stage by defining an appropriate experimental setup. The decisions about appropriate datasets, data placement onto peers, workload generation, data and system perturbation, and quality measures of interest will lead us to a first definition of a benchmark for query routing in P2P digital library systems. In addition, it is planned to build an application-oriented demonstrator for personalized search on e-health information [MST04, SSS04], e.g., for individuals that suffer from specific health problems, such as allergies or infrequent diseases, and are interested in finding relevant information in digital libraries and on the Web or exchanging information with other individuals suffering from similar diseases.

Project participants

The authors of this extended abstract are the principal investigators of the participating institutions:

- the Max-Planck Institute for Informatics in Saarbrücken, Germany
- the National University of Athens, Greece
- the University for Health Sciences, Medical Informatics and Technology in Innsbruck, Austria
- the University of Duisburg-Essen, Germany
- the Masaryk University in Brno, Czech Republic.

References

[Be04a] Matthias Bender, Sebastian Michel, Gerhard Weikum, Christian Zimmer: Towards Collaborative Search in Digital Libraries Using Peer-to-Peer Technology, DELOS Workshop on Digital Library Architectures, 2004.

[Be04b] Matthias Bender, Sebastian Michel, Gerhard Weikum, Christian Zimmer: Bookmark-driven Query Routing in Peer-to-Peer Web Search. ACM SIGIR Workshop on Peer-to-Peer Information Retrieval, 2004.

[Be05] Matthias Bender, Sebastian Michel, Peter Triantafillou, Gerhard Weikum, Christian Zimmer: Improving Collection Selection with Overlap Awareness in P2P Search Engines, ACM SIGIR Conference on Research and Development in Information Retrieval, 2005.

[Ca00] James P. Callan: Distributed Information Retrieval. In: W. Bruce Croft, Editor, Advances in Information Retrieval. Kluwer Academic Publishers, 2000.

[KI05] Georgia Koutrika, Yannis E. Ioannidis: Personalized Queries under a Generalized Preference Model. Int. Conf. on Data Engineering (ICDE), 2005.

[LC03] Jie Lu, James P. Callan: Content-based retrieval in hybrid peer-to-peer networks. Int. Conf. on Information and Knowledge Management (CIKM), 2003

[LW04] Julia Luxenburger, Gerhard Weikum: Query-log Based Authority Analysis for Web Information Search, International Conference on Web Information System Engineering (WISE), 2004.

[MYL02] Weiyi Meng, Clement T. Yu, King-Lup Liu: Building efficient and effective metasearch engines. ACM Computing Surveys 34(1), 2002.

[MST04] Michael Mlivoncic, Christoph Schuler, Can Türker: Hyperdatabase Infrastructure for Management and Search of Multimedia Collections. DELOS Workshop on Digital Library Architectures, 2004.

[NF03] Henrik Nottelmann, Norbert Fuhr. Evaluating different methods of estimating retrieval quality for resource selection. ACM SIGIR Conference on Research and Development in Information Retrieval, 2003.

[NZ05] David Nowak, Pavel Zezula: Indexing the Distance Using Chord: A Distributed Similarity Search Structure, DELOS Workshop on Future Digital Library Management Systems, 2005.

[PI04] Fragkiskos Pentaris, Yannis E. Ioannidis: Query Trading in Digital Libraries. DELOS Workshop on Digital Library Architectures, 2004.

[SSS04] M. Springmann, H.-J. Schek, H. Schuldt: Combination of Building Blocks for Similarity Search in Multimedia Electronic Patient Records (in German). Proceedings of 49th Conf. of the German Society for Medical Informatics, Biometry, and Epidemiology (GMDS), 2004.

[TK04] Martin Theobald, Claus-Peter Klas: BINGO! and DAFFODIL: Personalized Exploration of Digital Libraries and Web Sources, 7th International Conference on Computer-Assisted Information Retrieval (RIAO), 2004.

Context-dependent Access to Digital Libraries

Timos Sellis, ICCS, timos@dbnet.ece.ntua.gr
Yannis Ioannidis, UOA, yannis@di.uoa.gr
Nicolas Spyratos, UPS-XI, spyratos@lri.fr

Keywords
context, relational databases, personalization.

Research problem

DELOS envisages Digital Libraries that would allow information to be accessed and used in a global environment, where implicit assumptions about data become less and less evident. Users with different backgrounds or viewpoints may interpret the same data in a different way. Moreover, the *interpretation* and *suitability* of data may depend on changing conditions, e.g. the current position of the user or the media he is using (laptop, mobile, PDA, etc.). To cater for such ambiguous situations the information provider needs to specify the *context* under which information becomes relevant. Conversely, information users can specify their own current context when requesting for data in order to denote the part that is relevant to their specific situation. It is therefore interesting to explore how the notion of context can be directly supported by and incorporated in relational databases, which form the data management backbone of Digital Libraries. As we have shown in previous work [1,2] incorporating context at the level of databases leads to several benefits:

- Management of data according to the interpretation frame of prospective users.
- New types of queries (*cross-world*) that have no counterpart in context-unaware systems [5].
- Direct support for managing and querying data histories and schema histories [3,6], where context is used to express the valid time under which different variants of an information entity hold.
- A uniform and flexible mechanism [4,1] that adapts the structure, value, and presentation of information to the context of the end user.

In particular, delivering information in a personalized way is essential for future Digital Library systems. Modelling user preferences is a crucial prerequisite for this process, and it has attracted interest in both IR and Database research communities [7, 8, 9, 10]. There are various types of preferences. Most efforts tackle context-free preferences, i.e., preferences that hold under any circumstances. However, it is very common for people to express preferences that are valid within a certain context. For example, "I like W. Allen as a director only if the movie is comedy". The support of such preferences is therefore very important for any personalization service. On the other hand, currently the task of personalization burdens the application and is implemented ad-hoc for every different situation. However, recent work [7, 8] has shown that personalization can be supported in a uniform way as part of a relational database system. We believe that adding context as a data management mechanism will, among other things, significantly enhance the personalization capabilities of database management systems.

Objectives

The goal of the project is to investigate how context-dependent data support can be integrated into Digital Libraries. In particular, we will consider how a notion of context can be incorporated into relational databases. A number of issues arise, including context-aware models, operations and query languages, etc. Our aim is to extend the relational model so as to accommodate information entities that assume different structure or value under different conditions. An interesting application direction in Digital Libraries lies in using context to better support personalized access to information. Context-aware services in this domain should also exploit contextualized ontologies in order to provide generic support for context-dependent access.

Work done

Foundations of a logical model for context-aware relational data: In traditional databases and information systems the number of users is more or less known and their background is to a great extent homogeneous. In distributed and heterogeneous environments however such as Digital Libraries accessible through the Web, users do not apply the same conventions when interpreting data due to different backgrounds, knowledge or culture. Interpreting and managing data according to the context is a topic not explored in its full potential in these new environments. The focus of our work has been to define the Context Relational model (CR model) [9], a model that extends the relational model to argue also about context. The interesting part of this new approach is that context is treated as first-class citizen at the level of database models and query languages. This is due to the fact that an attribute may not exist under some contexts or that the attribute may have different values under different contexts.

Expected results

We expect to implement a prototype system that will demonstrate a substantial part of the functionality of a context-aware database system, including context-dependent data definition capabilities and context-aware querying. Furthermore, we plan to address the following issues:

- Extending the relational database foundations for incorporating context
- Context-aware operations and query language
- Using context to enhance the personalization capabilities of database systems

Project participants

National Technical University of Athens
University of Athens
University of Paris-South XI.

References

[1] Yannis Stavrakas. Multidimensional Semistructured Data: Representing and Querying Context-Dependent Multifaceted Information on the Web. PhD Thesis, Department of Computer Engineering and Informatics, National Technical University of Athens, Greece, 2003.

[2] Yannis Stavrakas, and Manolis Gergatsoulis. Multidimensional Semistructured Data: Representing Context-Dependent Information on the Web. In Proceedings of the 14th International Conference on Advanced Information Systems Engineering (CAiSE 2002), Toronto, Canada, May 2002.

[3] Yannis Stavrakas, Manolis Gergatsoulis, Christos Doulkeridis, and Vassilis Zafeiris. Representing and Querying Histories of Semistructured Databases Using Multidimensional OEM. In Information Systems Journal (IS), Vol. 29, Issue 6, pp 461-482, September 2004.

[4] Manolis Gergatsoulis, Yannis Stavrakas, Dimitris Karteris, Athina Mouzaki, and Dimitris Sterpis. A Web-based System for Handling Multidimensional Information through MXML. In Proceedings of the 5th East-European Conference on Advances in Databases and Information Systems (ADBIS 2001), Vilnius, Lithuania, September 2001.

[5] Yannis Stavrakas, Kostis Pristouris, Antonis Efandis, and Timos Sellis. Implementing a Query Language for Context-dependent Semistructured Data. In Proceedings of the East-European Conference on Advances in Databases and Information Systems (ADBIS 2004), Budapest, Hungary, September 2004.

[6] Manolis Gergatsoulis, and Yannis Stavrakas. Representing Changes in XML Documents Using Dimensions. In XML Database Symposium (XSym 2003) in Conjunction with VLDB 2003, Berlin, Germany, September 2003.

[7] G. Koutrika, Y. Ioannidis. Personalized Queries under a Generalized Preference Model. In Proceedings of 21st Intl. Conf. On Data Engineering (ICDE), Tokyo, 2005 (to appear).

[8] G. Koutrika, Y. Ioannidis. Personalization of Queries in Database systems. In Proceedings of 20th Intl. Conf. On Data Engineering (ICDE), Boston, 2004.

[9] Yannis Roussos, Yannis Stavrakas, and Vassia Pavlaki. Towards a Context-Aware Relational Model. To appear in the Workshop "Contextual Representation and Reasoning" held in conjunction with CONTEXT'05, the Fifth International and Interdisciplinary Conference on Modeling and Using Context, in Paris, France, July 5-8, 2005.

Modeling of User Preferences in Digital Libraries

Nicolas Spyratos, UPS-XI
Carlo Meghini, CNR-ISTI
Vassilis Christophides, FORTH
Yannis Ioannidis, UoA

Keywords
Information Search and Retrieval (H.3.3), Online Information Services (H3.5), Library Automation (H3.6), Digital Libraries (H3.7)

Research description
As information becomes available in increasing volumes, and to growing numbers of users, the shift towards a more user-centered access to information is becoming an important issue. As a consequence, support of personalized user interaction is an important concern in the design of advanced information systems, in general, and of digital libraries, in particular.

Personalization in a digital library is about building a meaningful one-to-one relationship between digital library services and users, by understanding the needs of each individual user; it can involve either adapting the user interface or adapting the content to the needs or preferences of a specific user. The aim of this task is to study an important aspect of content adaptation, namely, the provision of services for the management of user preferences, focusing on *query personalization*, *user notification*, and *document customization*.

Query personalization
The objective is to enhance the effectiveness of user information access via querying, by incorporating into queries user-specific preferences that are either stored in a profile or given explicitly by the user during query formulation time; the expected results are improved information filtering (decrease in the volume of data presented to the user), and a ranking of the presented data that reflects user preferences.

Presently, two different approaches of query personalization are being pursued, namely, a qualitative approach and a quantitative approach.

- In the qualitative approach, user preferences are captured via statements (obtained vie elicitation or mining) that induce an ordering on the document space--a preference relation in fact, which can be used for several tasks related to information access, such as filtering or ordering query results, refining or enlarging a previously stated query.
- In the quantitative approach, user preferences are captured via numbers indicating the degree of user preference on information schema elements. These numbers are combined to define scoring functions, which are used for modifying the document ranking resulting from a query evaluation.

This task studies both these approaches, as one can be preferred to the other depending on the application context.

User notification
The objective is to notify a user when an “event” of interest happens (i.e., a document of interest is registered, removed or modified at the library). In this context, the modelling of events and of their effect on user/system interactions are studied, with the objective of defining appropriate mechanisms which automatically modify the behaviour of the system depending on the history of event occurrences.

Document customization
The objective is to allow the user to browse a composite document and select desired components and their order of appearance, so that to compose a sub-document that will then be “materialized” by the system (i.e., augmented with a table of contents and an index).

Objectives
- to study a formal framework for specifying user preferences;
- to enrich the digital library information organization and retrieval services with preference capabilities for supporting query personalization, user notification, and document customization;
- to design algorithms for supporting preferences.

Expected results

- a report describing the formal framework for the definition of quantitative/qualitative preferences and their incorporation in the query language
- a demonstrator toolkit

References

[1] R. Agrawal, E.L. Wimmers, A Framework for Expressing and Combining Preferences, SIGMOD 2000, pp 297—306

[2] H.Andreka, M.Ryan, P-Y.Schlobbens, Operators and Laws for Combining Preferential Relations, Journal of Logic and Computation, 12.1, 2002, pp 13-53

[3] J.Chomicki, Preference formulas in relational queries, ACM Trans. Database Syst. 28(4): 427-466 (2003)

[4] J.Chomicki, Querying with Intrinsic Preferences, EDBT 2002: 34-51

[5] J. Chomicki, Semantic Optimization of Preference Queries, Intl Symposium on Constraint Databases, June 2004, Paris, France

[6] B.A.Davey, H.A.Priestly, Introduction to Lattices and Order, Cambridge University Press, 2002 (2nd Edition)

[7] K.Govindarajan, B.Jayaraman, S.Mantha, Preference Queries in Deductive Databases, New Generation Computing 2000, Vol. 19, No1, pp57-86

[8] S.Holland, W.Kießling: Situated Preferences and Preference Repositories for Personalized Database Applications. ER 2004: 511-523

[9] W.Kießling: Foundations of Preferences in Database Systems. VLDB 2002: 311-322

[10] W.Kießling, Gerhard Köstler: Preference SQL - Design, Implementation, Experiences. VLDB 2002: 990-1001

[11] G.Koutrika, Y.Ioannidis, Constrained Optimalities in Query Personalization, ACM-Sigmod 2005

[12] G.Koutrika, Y.Ioannidis, Personalization of Queries in Database Systems, ICDE, 2004, pp 597-608

[13] M.Lacroix, P.Lavency, Preferences: Putting More Knowledge Into Queries, VLDB 1987, pp 217-225

[14] U. Mamber, A. Patel, J.Robison, Experience with personalization on Yahoo!, Communications of the ACM, August 2000, Volume 43, Number 8

[15] D. Riecken, Personalized views of personalization, Communications of the ACM, August 2000, Volume 43, Number 8

[16] P. Rigaux, N. Spyratos, Metadata Inference for Document Retrieval in a Distributed Repository (Invited Paper), 9th Asian Computing Science Conference (ASIAN’04), Chiang-Mai, Thailand, 8-10 December 2004, LNCS 3321/2004

Audio/Visual and Non-traditional Objects

Cluster objectives
Digital libraries will capture, organize, store and manage the access to large amounts of digital information regarding human knowledge, culture, and history in various, possibly interconnected, presentation forms like video, audio, images, etc. The objectives of this cluster are to establish a common ground of knowledge for European researchers about the state of the art, the research directions and important new applications for digital libraries with audio-visual and non-traditional objects, as well as to advance the state of the art in these areas.

Cluster activities
The most important activities foreseen in the workpackage include:

- Establishing common functionalities and advancing the state of the art in the area of metadata capturing from audiovisual content, including the investigation of issues related to multimodal information extraction, and the use of domain specific, context specific, and historical information in the extraction process.
- Establishing common foundations and advancing the state of the art in the area of information access and interactions with audio-visual digital libraries exploring multimedia content standards, domain and context specific knowledge, and investigating advanced interactions and interfaces to multimedia content.
- Establishing common foundations and advancing the state of the art in the area of management of audiovisual content, including new database models and data structures for storage, retrieval, and dissemination of multimedia data in emerging architectures and applications.

Cluster coordinators
Alberto Del Bimbo, Università degli Studi di Firenze
Stavros Christodoulakis, Technical University of Crete

Video Annotation with Pictorially Enriched Ontologies

Alberto Del Bimbo, Marco Bertini, Carlo Torniai {delbimbo, bertini, torniai}@micc.unifi.it)
D.S.I - MICC Università degli Studi di Firenze
Stavros Christodoulakis, Chrisa Tsinaraki {stavros, chrisa }@ced.tuc.gr),
Laboratory of Distributed Multimedia Information Systems & Applications,
Tech. University of Crete (TUC/MUSIC), Chania, Greece.
Rita Cucchiara, Costantino Grana {rita.cucchiara, grana.costantino}@unimore.it
D.I.I. Università degli Studi di Modena e Reggio Emilia

Keywords

H.3.7 [Information Storage and Retrieval]: Digital Libraries; H.2.4 [Systems]:Multimedia databases

Introduction

Classifying video elements according to some pre-defined ontology of the video content domain is a typical way to perform video annotation. Ontologies are defined by establishing relationships between linguistic terms that specify domain concepts at different abstraction levels. However, although linguistic terms are appropriate to distinguish event and object categories, they are inadequate when they must describe specific patterns of events or video entities. Instead, in these cases, pattern specifications can be better expressed through visual prototypes that capture the essence of the event or entity. Therefore *pictorially enriched ontologies*, that include both visual and linguistic concepts, can be useful to support video annotation up to the level of detail of pattern specification.

Objectives

This project aims at defining methodologies and techniques to describe concepts and their specializations by *augmenting an ontology of linguistic terms with "visual concepts" that represent these instances in a visual form*. The visual concepts should be learned from occurrences of the highlights through analysis of their similarity (in the spatio-temporal domain) and automatically extracted from both raw and edited videos and integrated into the ontology. *Visual concepts*, once added to the ontology, will integrate the semantics described through linguistic terms up to a more detailed representation of the context domain. Visual concepts will be defined by means of global features, meaningful spatial segments (such as regions of frames or key-frames) and temporal segments (as highlights or representative shots). The end result is a *pictorially enriched ontology* (PE-Ontology) that fully supports video annotation, allowing classification and annotation of events up to very specialized levels. The domain of interest is the sport video domain in particular soccer and Formula one videos.

Preliminary results

The following activities have been carried on so far in order to achieve the task objectives:

- **Preliminary definition of PE Ontologies structure**
  PE Ontologies contain a linguistic part and a visual part. The linguistic part is composed by the video and clip classes, the actions class and its highlights subclasses and an object class with its related subclasses describing different objects within the clips. The visual part is created adding to the linguistic part of the ontology the visual concepts as specializations of the linguistic concepts that describe the highlights. Preliminary ontologies, containing highlights, patterns for important complex soccer actions and visual concepts have been defined for both soccer and Formula one domains.
- **Definition of a preliminary set of low level visual features for visual concept implementation**
  We have defined a set of visual features in order to define visual concepts for PE Ontologies. For soccer domain, for instance, we are employing the playfield area, the number of players in the upper part of the playfield, the number of players lower part of the playfield the motion intensity, the motion direction and the motion acceleration.
- **Definition of the creation process of PE Ontologies**
  The creation process of the pictorially enriched ontology is performed by selecting a representative set of sequences containing highlights described in the linguistic ontology, extracting the visual features and performing an unsupervised clustering. The clustering process, based on visual features, generates clusters of sequences representing specific pattern of the same highlight that are regarded as specialization of the highlight. Visual concepts for each highlight specialization are automatically obtained as the centers of these clusters. We have employed Fuzzy C-Mean (FCM) clustering with normalized Needleman-Wunch distances

in order to define visual prototypes for highlights. For PE Ontology creation a shot detection algorithm and a sub shot identifier were developed.

- **Definition of metrics for evaluation of similarity**
  The distance between videos and visual concepts can be computed as the sum of all the normalized Needleman-Wunch distances between the visual features, to take into account the differences in the duration and the temporal changes of the features values. This distance is a generalization of the Levenshtein edit distance and has been used since the cost of character substitutions is an arbitrary distance function. In our case the cost can be used to weight differently the differences in the motion intensity. The normalization is useful in order to better discriminate differences between short and long sequences and is performed dividing the Needleman-Wunch distance by the length of the shorter sequence.
- **Preliminary definition of complex actions patterns**
  A definition of sequences of simple highlights or events that compose more complex actions and events have been identified both for soccer and formula 1. These patterns have been added to the ontology in order to refine the annotation system.
- **Preliminary definition of an Annotation Algorithm based on PE Ontologies**
  The pictorially enriched ontology can be used effectively to perform automatic video annotation with higher level concepts that describe what is occurring in the video clips. This is made by checking the similarity of the clip content with the visual prototypes included in the ontology. If similarity is assessed with a particular visual concept then also higher level concepts in the ontology hierarchy, that are linked to the visual concept, are associated with the clip, resulting in a more complete annotation of the video content. An annotation algorithm for some highlights of soccer video clips has been defined.

Expected results

Expected results are the formalization of PE Ontologies structure, the definition of metrics and visual features for PE Ontologies automatic creation. Furthermore we will be able to implement a framework that could perform automatic annotation of video clips based on PE ontologies. The framework will be used for experiments and tests in video annotation and query evaluation on ground truth dataset videos. The preliminary framework structure is shown in Figure. The annotator, which automatically recognizes specific actions, takes into account **(a)** ontologies associated with pictorial information, which will be (image or video) prototypes of the ontology class instances as well as visual concepts describing them; **(b)** a-priori knowledge that will assist the recognition process (e.g. the players participating in a soccer game or the drivers in an F1 race); and **(c)** action patterns, that allow inferring complex actions from the association of the simple actions recognized by the annotator with their temporal order. The output of the annotator will be the simple and the complex actions recognized in OWL/RDF format, which will be then transformed into MPEG-7 metadata and stored into the VAPEON Semantic Base.

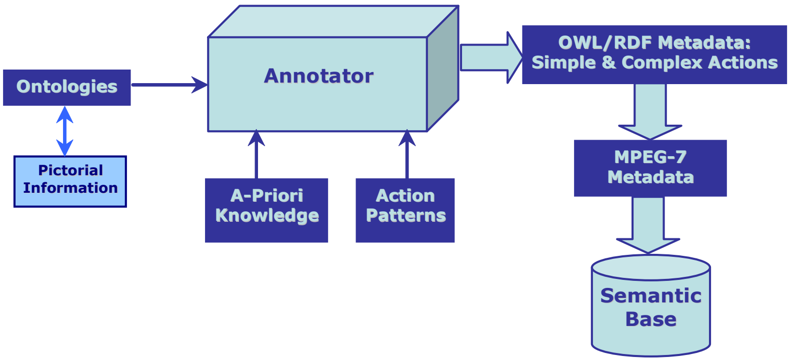


Figure 1: Overview of the VAPEON Architecture

MIMA: Multimedia Interfaces for Mobile Applications

Alberto Del Bimbo (UNIFI-MICC)
Rita Cucchiara (UNIMO)
Giuseppe Santucci (ROMA1)
Margherita Antona (FORTH)

Keywords
H.3.7 [Information Storage and Retrieval]: Digital Libraries; H.2.4 [Systems]:Multimedia databases

Problem Specification

Besides living in a "ubiquitous age", we are also living in a "digital age". Besides the talk about digital libraries, digital cameras, digital TVs (and virtually digital *anything*), and the experience of the same, the reality of the physical is still with us, is necessary, and is arguably here to stay. In the context of digital libraries, there is therefore a dire need to support the library user to fluidly, continuously or seamlessly: switch between the digital and physical fronts, interact with the digital and physical entities, and operate within the digital and physical settings. Among (or together with) possibly other choices, considering the adoption of mobile computing for that cause would not be a miscalculation.

In the above framework, the MIMA project stems from the opportunity coming from the large availability of reliable connections through handheld mobile devices such as PDAs and smart phones and the vast amount of digital video streams daily produced. The trend suggests that one of the main challenges of the very next times will be the availability of tools and solutions that support effective real-time presentation of specialized live video content, like news or sport videos.

The project intends to address the main problems associated with the customizable transmission of video streams on portable devices. To this aim, the project will investigate several strictly interrelated subproblems, producing results in the framework of multimedia access and video presentation on mobile devices. The topics addressed by the project will range from video stream handling to adaptable user interfaces on mobile devices.

In particular, the project will concentrate on:

- Automatic video annotation (see, e.g., [1], [2]), including several sub topics, like the extraction of the basic video features, the modeling of interesting events through suitable knowledge models, and the construction of video summaries.
- through a flexible model, based on Description Logic, able to minimize the interaction with the user, so avoiding the filling in of long and tedious forms.
- User-centred design of flexible small screen device interface, able to minimize the user interaction and adapt to user profiles and devices characteristics. This topic covers two main issues. The first one is user and device profiling (that deals with the modeling of user preferences and device characteristics); the second is to design adaptive interfaces, i.e. interfaces that automatically adapt according to user profiles and preferences as well as device characteristics. The third one stems from the idea of adopting information visualization techniques (see, e.g., [3]) for managing the video annotations, presenting the user with rich and effective visual representations.
- Define and test suitable indexes able to measure the quality of different representations, choosing the solution that optimizes usability, user experience and required connection time.

Based on the theoretical and practical results stemming from the investigation of the above topics, the research units involved in the MIMA project will set up a comprehensive system prototype that will be used as a proof of concept to demonstrate the feasibility of the adopted approach.

Project description and objectives

The project intends to investigate several strictly interrelated subproblems, producing results in the framework of multimedia access for video presentation on mobile devices. The main subjects of investigation will be:

- Automatic video extraction of meaningful objects and events according to user's interests.
- User profiling and design of flexible small screen device interface, able to minimize the user interaction and adapt to devices characteristics.
- Performance measures and quantitative/qualitative indexes of user experience and satisfaction.

The overall project scenario is: The user is equipped with a PDA in order to receive multimedia information, i.e., video, images, graphic objects, text, and audio. The foreseen field of application is transmission of sports and news video, enriched with video summaries. The overall architecture is composed of three subsystems: *Video annotation*; *Video summarization*; *User Interface*.

Video annotation Off-line annotation takes place on uncompressed video, producing a more precise annotation, extracting highlights and significant objects/events. Highlights must be represented with appropriate knowledge models based on the a-priori knowledge of the spatial-temporal structure of events and recognized by a model checking engine, based on statistical or model-based classification frameworks.
Image processing and analysis is used to extract the salient features of the video such as motion vectors (that quantify the activity), color patterns (that distinguish background zones), lines, corners and shapes (that identify objects). Text appearing on the video can be extracted and recognized. Players' position in the playground can be detected in order to build statistics of the field occupancy. Construction of video summaries upon user's request are managed by the *Video summarization* subsystem. Summaries are obtained dynamically, combining the user request with the annotations obtained from the off-line annotation process.

The *User interface subsystem* is in charge of handling the interaction with the user, and is faced with two main problems/objectives:
1- it should nicely fit the device characteristics and the user preferences.
2- it should include new interaction and visualization techniques to effectively convey the information produced by the annotation and video summarization systems.

Towards achieving the above, the system will include the following models (a) user preferences, (b) device characteristics, and (c) the characteristics of the objects/services available to the user, in terms of their informative contents. Additionally, the designed interface should be capable of automatic adaptation according to device characteristics and user preferences in order to ensure usability. For example, menus should be automatically organised and displayed, possibly through different modalities, according to the user's individual interaction preferences and type of device used.
Furthermore, content-based adaptation will be addressed. For example, non interesting events should be transmitted and visualized in textual or audio form (or non transmitted if definitely uninteresting), while interesting events should be displayed at full screen resolution or given in audio. The user should be able to switch among these modalities, request summaries (either textual or visual), and provide feedback that modify display options and his/her profile preferences.

To assess the above aspects, the project will explore suitable metrics able to measure the quality of different representations and choosing the solution that optimizes usability, user experience and required connection time. The involved partners will collaborate at the project completion according to their skills:

- UNIFI: video analysis and annotation
- UNIROMA 1 user interfaces for mobile systems and information visualization
- UNIMORE: video segmentation and multimedia information extraction
- FORTH: quality probing.

Project state
According to the project work plan, a comprehensive analysis of state-of-the-art has been performed, covering aspects such as video automatic annotation, multimedia interfaces and visualization for mobile devices, usability of mobile devices interfaces for video content access. Moreover, an analysis of the user requirements is in progress, in order to produce the specifications for the user interface and to design the overall system architecture.

References

[1] G. Tardini, C. Grana, R. Marchi, R. Cucchiara, "Shot Detection and Motion Analysis for Automatic MPEG-7 Annotation of Sports Videos" in press on *13th International Conference on Image Analysis and Processing (ICIAP 2005)*, Cagliari, Italy, 6-8 Sept, 2005

[2] Bertini M., Cucchiara R., Del Bimbo A., Prati A, "Content-based video adaptation with user's preferences", International conference on multimedia and expo icme, Taiwan, 2004

[3] E. Bertini, G.Santucci - Improving 2D scatterplots effectiveness through sampling, displacement, and user Perception - nei Proceedings della 9th International Conference on Information Visualisation IV05 - July 2005, London.

Description, Matching and Retrieval by Content of 3D Objects

Stefano Berretti, Alberto Del Bimbo, Pietro Pala (UNIFI-MICC)
Nozha Boujeama, Michel Crucianu (INRIA)
Rita Cucchiara, Costantino Grana (UNIMORE)
Arnold Smeulders, Marcel Worring (UVA)

Keywords
H.3.7 [Information Storage and Retrieval]: Digital Libraries; H.2.4 [Systems]:Multimedia databases

Introduction

Beside image and video databases, archives of 3D models have recently gained increasing attention for a number of reasons: advancements in 3D hardware and software technologies – in particular for acquisition, authoring and display – their ever increasing availability at affordable costs, and the establishment of open standards for 3D data interchange (e.g. VRML, X3D).

Thanks to the availability of technologies for their acquisition, 3D models are being employed in a wide range of application domains, including medicine, computer aided design and engineering, and cultural heritage. In this framework the development of techniques to enable retrieval by content of 3D models assumes an ever increasing relevance. This is particularly the case in the fields of cultural heritage and historical relics, where there is a growing interest in solutions enabling preservation of relevant artworks (e.g. vases, sculptures, and handicrafts) as well as cataloguing and retrieval by content. In these contexts, retrieval by content can be employed to detect commonalities between 3D objects (e.g. the "signature" of the artist), to monitor the temporal evolution of a defect (e.g. the amount of bending for wooden tables), to support services for tourists and visitors of historical sites (e.g. assist tourists in finding information related to an object of interest given a sample photograph of the object).

A major difficulty for the development of a system for retrieval by content of 3D objects relates to the need to capture the twofold nature by which 3D objects are experienced by humans: view based and structural. On the one hand, 3D objects can be perceived through multiple (or single) 2D views. On the other hand, 3D objects can also be examined so as to analyze their 3D structure. This twofold nature is not separable and indeed, our perception of 3D object similarity is purely view based in some cases, purely structural in other cases and a combination of both in the general case.

Objectives

The ultimate goal of this project is to develop a system to support structural as well as view based retrieval by content of 3D objects. In this context, the project aims at the investigation of models for extraction of view based and structural based descriptors, models for indexing and similarity matching of structural and view based descriptors, models and metaphors for querying archives of 3D objects. The theoretical investigation of these models will end up with the design and development of a prototype system supporting structural and view based retrieval of 3D objects. In particular, project activities will address the following issues:

- *Content description*. Models will be investigated and experimented to extract descriptors of 3D object content from multiple viewpoints. These descriptors should capture prominent features of object views so as to enable retrieval by similarity based on a single photograph of an object, taken from a generic viewpoint. Models will also be investigated and experimented to extract descriptors of 3D object structure. For this purpose, 3D object segmentation techniques will be developed so as to allow decomposition of a 3D object into its structural components. Each component will be described separately so as to enable description and retrieval based on characteristics of object parts in addition to global object features.
- *Indexing and similarity matching*. For both descriptors of object views and object structure a distance measure should be defined to allow computation – on a perceptual basis – of the similarity between a generic 3D object and a template, this latter being represented either as the image of an object from a particular viewpoint or as a compound set of 3D parts.
- *Querying and presentation*. Despite its large use to support access by content to image libraries the query by example paradigms, in its original form (pick one item from the archive and retrieve similar items), exhibits some limitations when applied to libraries of 3D objects. This is particularly true in the context of this project where retrieval based on an object photograph (image) as well as retrieval based on object parts are addressed. The former demands for the definition of models to manage specification of the query through an external image (representing one view of the object of interest). The latter relies on the possibility for the user to select a subset of the structural parts of an archived object and use these parts only to retrieve objects with similar parts in a similar arrangement.

Preliminary Results and Future Work

The following activities have been developed so far in order to achieve the task objectives:

- **Objects partitioning**

  Different approaches (including watershed, object skeletonization, oscillator networks, etc.) have been experimented in order to decompose a 3D model into its constituent salient parts. In the perspective to identify perceptually relevant protrusions of a model, to be used in the construction of a graph based representation, techniques based on topological properties of 3D objects appear as the most attractive. In fact, this kind of approaches tend to do not detect patches originated by slight curvature changes which can instead occur using curvature based methods and result in the identification of perceptually irrelevant object parts. In particular, we developed a segmentation approach based on the construction of a Reeb-graph and its successive reduction based on graph topological information.
- **Objects representation**

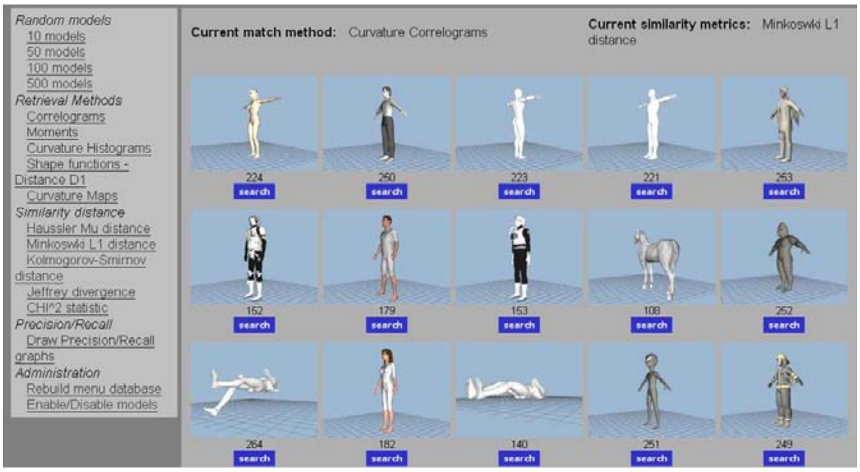
  An original solution based on curvature correlograms has been developed and experimented in order to represent the characteristics of objects surface. In fact, while descriptions based on curvature histograms or statistics of surface features (like curvature, distance distribution between salient vertexes, etc.) usually fail to capture local properties, correlograms have the advantage to also encoding information about the relative position of local features. In particular, correlograms can encode information about curvature values and their localization on the object surface. This is obtained by first discretizing the object curvature into a set of reference classes, then, given a distance δ, evaluating a matrix whose elements account for how many mesh vertexes at distance δ exist for every possible combination of two curvature classes. For its peculiarities, description of 3D objects based on correlograms of curvature proves to be very effective for the purpose of content based retrieval of 3D objects.

  We are also studying alternative representations based on the Hough transform and spin images.
- **Comparative evaluation**

  In the perspective to compare the approach/es developed in this task, against state of the art solutions, a set of reference methods have been implemented. In particular, the description techniques presently implemented include *3D geometric moments*, *curvature histograms* and *shape functions*. For the purpose of comparison, we have also planned to implement methods based on *spherical harmonics*, MPEG-7 s*hape spectrum*, and *Zernike moments*. In addition, several distance functions can be used to evaluate the similarity between 3D objects descriptions during retrieval. Currently, these distance measures include the *Minkowski distance*, *Histogram Intersection, Kullbach-Leibler divergence* and χ^2 *distance*.
- **Prelimary retrieval system**

  A preliminar system has been designed and implemented to demonstrate retrieval by content of 3D objects using different approaches. The system features a Web accessible interface (available at http://viplab.dsi.unifi.it:8080/CV) that allows users to browse a database of 3D objects. Browsing can be accomplished either by random sampling or by retrieval by visual similarity.

  Figure shows the system interface with the retrieval results obtained by using the curvature correlograms representation and the Euclidean (Minkowski L1) distance.



Current match method: Curvature Correlograms

Current similarity metrics: Minkoswki L1 distance

Figure 1: Retrieval results for the query model in the upper left corner using curvature correlograms as matching method, and Minkowski L1 distance as similarity metric.

Automatic, Context-of-Capture Based, Categorization, Structure Detection and Segmentation of News Telecasts

Arne Jacobs, George T. Ioannidis {jarne, george.ioannidis}@tzi.de
Center for Computing Technologies (TZI), University of Bremen, Germany
Martha Larson matha.larson@imk.fraunhofer.de
Fraunhofer Institute for Media Communication, Sankt Augustin, Germany
Stavros Christodoulakis, Nektarios Moumoutzis {stavros, nektar}@ced.tuc.gr
Laboratory of Distributed Multimedia Information Systems and Applications
Technical University of Crete (TUC/MUSIC), Chania, Greece

Keywords
Information search and retrieval, modeling multimedia data, ontology design, speech recognition, structural models, video analysis

Research problem

The context of an element within a time series can be conceptualized as a state. In the case of digital libraries, the time series of interest are multimedia streams. If a multimedia stream is a recording of a discussion, the context of any single utterance within can be understood to be the state of the discussion at the point the utterance was spoken. This state can be understood as encoding the subject being discussed and the propositions that have been agreed upon by the parties participating in the discussion. For the purpose of multimedia analysis, it is useful to model a context state with an abstract set of hierarchies. In such a model, which we call a Context-of-Capture (CoC), a particular context is characterized by a set of interrelated concepts described in an ontology. It is possible to infer a CoC using the set of words that have occurred in a discussion, or in any general discourse. Knowing the CoC associated with each captured segment of a discourse makes possible discourse interpretation on a semantic level, previously unattained. In the work presented here we focus our research efforts on a particular type of discourse, namely, news broadcasts.

Automatic audiovisual content segmentation is performed mainly at the syntactic level in several systems today, but only few systems exist that take into account the semantics of the audiovisual content. Furthermore, the CoC concept, which represents the context regarding the information captured in an audiovisual segment (e.g. persons, places, events etc.), is either ignored or superficially utilized.

The CoC may be of great importance, especially for programs like news and telecasts, consisting of totally independent video segments and many topic changes. In these cases, the sudden changes in the CoC denote the end of the current context segment and the beginning of a new one.

In addition, the CoC allows for the automatic assignment of the audiovisual segments detected into appropriate thematic categories, as the CoC of a segment contains adequate information for thematic category determination. Equally important with the recognition of a specific context for segmentation and indexing purposes is also the possibility to associate all the knowledge in the knowledge base that is associated with a context.

The above discussion makes clear the need for generic models for describing CoC and scenarios of context appearance, and their use in recognition, segmentation and structuring of the knowledge bases so that complex queries can be answered.

Objectives

The objective of this project is to develop a demonstrator for automatic categorization, structure detection and segmentation of news telecasts that utilizes advanced structural models. Segment boundary detection will be assisted by a powerful CoC model, which will be used by the appropriate context detection and context change evaluation mechanisms. The segmentation/structural metadata will be finally exported in MPEG-7 format. A query API and user interface will be provided in order to evaluate the results.

Expected results

In particular, task activities will address the following issues:

Definition of the Context-of-Capture (CoC) model. A powerful model for the CoC and of CoC scenarios will be developed. A CoC will be represented as an ontology based on MPEG7 and OWL interoperability will be pursued [1][2]. CoC scenarios will be represented as graphs representing structures of possible transitions among CoC during a news telecast. Furthermore, the model will support algorithms for utilizing CoC for identification and knowledge management including its use for recognition and inference. The CoC model will be used to classify news stories into relevant contexts based on their content as specified by the recognition process (see below). Moreover, the identified context may be further used to guide the recognition process and resolve ambiguities that may be present after the first recognition phase.

Development of CoC recognition mechanisms. Appropriate algorithms will be developed for CoC recognition, utilizing image, speech and video text processing for audiovisual feature extraction. Simple audiovisual cues (characteristic color, texture, loudness), extraction of text inserts [3] and higher-level visual features (e.g. faces, indoor/outdoor [4]) will be taken into account. In addition, the spoken audio track will be processed using a speech recognizer that identifies the occurrences of keywords, using syllable-based methods [5][6]. The speech recognizer is trained to recognize exactly those keywords whose presence makes it possible to enrich the CoC models. We assume that the presence of individual keywords will allow us to infer the relevance of higher semantic concepts for particular context states.

Development of mechanisms for CoC–based segmentation and categorization of telecasts. A syntactic segmentation of a telecast can be obtained using algorithms for both shot detection on the visual signal and speaker and speech/music recognition on the audio signal [7][8]. In addition, a model of the visual syntax of a broadcast [9] may be used to categorize shots into higher-level syntactic segments (like, e.g., report, credits, or presentation). This segmentation can then be further refined by changes in the CoC, denoting the segment boundaries. This refinement mainly concerns the merging of adjacent syntactic segments that have very similar CoC.

Development of a query API and user interface for evaluation. A query API and user interface will be provided in order to evaluate the results. The query API will be used to search a set of classified news stories and the user interface will provide a user-friendly way to browse the results and evaluate the efficiency of the followed approach.

References

[1] Tsinaraki C., Polydoros P., Christodoulakis S.: "Interoperability support for Ontology-based Video Retrieval Applications", In the Proceedings of CIVR 2004, Dublin/Ireland, July 2004

[2] Tsinaraki C., Polydoros P., Moumoutzis N., Christodoulakis S.: "Coupling OWL with MPEG-7 and TV-Anytime for Domain-specific Multimedia Information Integration and Retrieval", In the Proceedings of RIAO 2004, April 2004, Avignon/France

[3] A. Miene, Th. Hermes, G.T. Ioannidis: "Extracting Textual Inserts from Digital Videos", in Proc. of the Sixth International Conference on Document Analysis and Recognition (IDCAR'01), pp. 1079-1083, Seattle, Washington, USA, IEEE Computer Society, September 10-13, 2001

[4] A. Miene, Th. Hermes, G.T. Ioannidis, R. Fathi, and O. Herzog: "Automatic Shot Boundary Detection and Classification of Indoor and Outdoor Scenes", In Voorhees, E. M. and Buckland, L. P. (eds.), Information Technology: The 11th Text Retrieval Conference, *TREC 2002*, NIST, pp. 615-620, 2003

[5] S. Osang: "Entwicklung eines Schlüsselworterkennungssystems zur Medienbeobachtung", Diploma thesis, Fachhochschule Bonn-Rhein-Sieg and Fraunhofer IMK

[6] M. Larson, S. Eickeler: "Using Syllable-based Indexing Features and Language Models to Improve German Spoken Document Retrieval", In *Proceedings of Eurospeech*, 8th European Conference on Speech Communication and Technology, Geneva, September 2003

[7] K. Biatov, J. Köhler: "An Audio Stream Classification and Optimal Segmentation for Multimedia Applications". In *ACM Multimedia*, San Francisco, November 2003

[8] J. Löffler, J. Köhler, H. Blohmer, K.-U. Kaup: "Archiving of Radio Broadcast Data Using Automatic Metadata Generation Methods within MediaFabric Framework", In *Proceedings of the 116th AES Convention*, Berlin, May 2004

[9] L. Bankert, A. Jacobs, A. Miene, Th. Hermes, G.T. Ioannidis, O. Herzog: "An Environment for Modelling Telecast Structures". In T. Catarci, S. Christodoulakis, A. Del Bimbo (eds.), *AVIVDiLib'05 Proceedings*, pp. 176--179, May 2005.

CoCoMA: Content and Context Aware Multimedia Content Retrieval, Delivery and Presentation

Stavros Christodoulakis, Chrisa Tsinaraki {stavros, chrisa}@ced.tuc.gr.
Laboratory of Distributed Multimedia Information Systems & Applications,
Tech. University of Crete (TUC/MUSIC), Chania, Greece
Christian Breiteneder, Horst Eidenberger, Doris Divotkey
{breiteneder, eidenberger, doris.divotkey}@ims.tuwien.ac.at
Technical University of Vienna (TUV), Vienna, Austria
Susanne Boll, Ansgar Scherp {boll, scherp}@offis.de
Oldenburg Research and Development Institute for Information Technology Tools and Systems (OFFIS), Germany
Elisa Bertino, Andrea Perego {bertino, perego}@dico.unimi.it
University of Milano (UNIMI), Milan, Italy

Keywords
[H.2.4 Systems] Multimedia Databases, [H.5.1 Multimedia Information Systems] Video, [H.3.7 Digital Libraries] User issues, [H.3.3 Information Search and Retrieval] Information Filtering, Relevance Feedback, [H.5.1 Multimedia Information Systems] Video.

Introduction

The increasing availability of high-speed wired and wireless networks as well as the development of a new generation of powerful (mobile) end-user devices like PDAs or cell phones leads to new ways of multimedia resource consumption. At the same time, new standards like MPEG-7/21 have become available, allowing us the enrichment of media content with semantic content annotations, which in turn facilitates new forms of multimedia experience, like search on specific topics or semantic-based content selection, filtering, and retrieval. *CoCoMA (Content and Context Aware Multimedia Content Retrieval, Delivery and Presentation)* focuses on the *integration* of *content and context-based multimedia retrieval from digital libraries* with the *personalized delivery and consumption* of the retrieved multimedia data. We aim at providing users of digital library systems with a solution for intelligent personalized retrieval from large media collections where media transport and presentation of the retrieval results are based on adaptation according to the user preferences.

The major objectives of CoCoMA are listed below:

1. *Support for Semantic User Preferences in the MPEG-7 Multimedia Description Scheme*. User preferences can be encapsulated in the MPEG-7 MDS *UserPreferences* descriptor. As the MPEG-7 MDS user preferences, although structured, do not allow taking semantic entities into account, the MPEG-7 MDS user profiles will be enriched, in order to allow for the expression of semantic user preferences.
2. *Personalization of the content (or semantic)-based flow and duration of the presentation, with respect to the interests and skills of the end-users*. Multiple execution flows, with possibly different duration, for the same multimedia presentation will be provided, by specifying (MPEG-7) semantic relationships among multimedia objects and the relevant portions of each object. This information will then be used to dynamically generate a personalized 'view' of the multimedia presentation satisfying the end user's preferences.
3. *Audiovisual content adaptation based on the user's individual context, specific background, interests, and knowledge, as well as the heterogeneous infrastructure of end-user devices*. The multimedia content will be selected based on the user profile, adapted to the user's context and assembled into a multimedia composition. Multi-channel multimedia presentation generation will also be provided so that all the different users can get and use the retrieved content in their individual device configuration.
4. *Novel interaction paradigms during retrieval*. Novel interaction paradigms will be introduced that make use of user preference information, employ semantically enriched versions of the low-level MPEG-7 descriptors, and introduce novel distance measures for similarity modeling. Furthermore, we will make use of relevance feedback cycles to improve content adaptation iteratively.

Architecture

The CoCoMA architecture, which has already been agreed among the task members, will utilize and extend the components listed below:

- The MM4U framework [12][4], developed by OFFIS, which is a generic and modular framework that supports multimedia content personalization applications. The framework components are generic and support the steps for composing and delivering semantically rich, personalized multimedia content.

- The VizIR framework for content-based multimedia retrieval [5][6][7][8][9], developed by TUV, allows for content-based metadata extraction and modeling, media annotation (e.g. the entire MPEG-7 MDS), query formulation and refinement, media access and user interface design. In addition, tools for media and media metadata visualization have been developed, which are used in the VizIR user interface framework.
- The KoMMa [18] framework, developed by Klagenfurt University, provides an open, extensible, and intelligent adaptation framework for multimedia data which can be used to build powerful multimedia adaptation servers or proxies. The framework makes use of MPEG-7 and MPEG-21 metadata and features a Prolog unit which is responsible for the adaptation decision taking process.
- The multimedia authoring system developed by UNIMI [1][2][3], which supports constraints for personalizing the presentation of multimedia objects according to users' preferences and skill levels. The execution flow is built dynamically on the basis of the semantic correlations existing among multimedia objects. Personalization features are provided concerning, both duration requirements and/or preferences, and the end user's device and network connectivity.
- The DS-MIRF framework developed by TUC/MUSIC [10][11][13][14][15][16][17]. It allows for the interoperability of OWL with the complete MPEG-7 MDS so that domain ontologies described in OWL can be transparently integrated with MPEG-7 metadata. Thus, applications that recognize and use the MPEG-7 MDS constructs to make use of domain ontologies, resulting in more effective user retrieval and interaction with the audiovisual material.

The above components will be integrated in a uniform architecture and extended in order to allow the utilization of semantic metadata from the MPEG-7 MDS user profiles, and the combination of the filtering and search preferences of the end-users with their summary preferences, in order to support the dynamic generation and delivery of personalized summaries. In addition, relevance feedback cycles will be used to improve content adaptation iteratively and multiple execution flows, with possibly different duration, for the same multimedia presentation will be provided. With the architecture we achieve an integrated support for content and context aware from the retrieval in digital libraries to the delivery and consumption of the retrieved multimedia data.

References

[1] Bertino E., Ferrari E., Perego A., Santi D., "Constraint-based Techniques for Personalized Multimedia Presentation Authoring". In Proc. of the DELOS AVIVDiLib'05 Workshop, Cortona, Italy, May 2005, pp. 91-94, 2005.

[2] Bertino E., Ferrari E., Stolf M., "MPGS: An interactive tool for the specification and generation of multimedia presentations". In IEEE Trans. on Knowledge and Data Engineering, 12(1):102–125, 2000.

[3] Bertino E., Ferrari E., Perego A., Santi D., "A Constraint-based Approach for the Authoring of Multi-topic Multimedia Presentations". In Proc. of the IEEE ICME'05, Amsterdam, The Netherlands, July 2005.

[4] Boll S., "MM4U - A framework for creating personalised multimedia content". In the Proc. of International Conference on Distributed Multimedia Systems (DMS' 2003), Miami, Florida, USA, September 2003.

[5] Eidenberger H., Breiteneder C., "VizIR – A Framework for Visual Information Retrieval". In Journal of Visual Languages and Computing (2003) 14 443-469.

[6] Eidenberger H., "Media Handling for Visual Information Retrieval in VizIR". In SPIE volume 5150, 1078-1088 (2003).

[7] Eidenberger, H., "A Video Browsing Application based on visual MPEG-7 Descriptors and Self-Organising Maps". In International Journal of Fuzzy Systems, vol. 6, no. 3, pp. 124-137, 2004.

[8] Eidenberger, H., "Statistical analysis of MPEG-7 image descriptions". In ACM Multimedia Systems Journal, Springer, vol. 10, no. 2, 2004.

[9] Eidenberger, H., Divotkey, R., "A Data Management Layer for Visual Information Retrieval". In Proc. of ACM SIGKDD Multimedia Data Mining Workshop, pp. 48-51, Seattle, USA, 22.8.2004.

[10] Kazasis, F.G., Moumoutzis, N., Pappas, N., Karanastasi, A., Christodoulakis, S., "Designing Ubiquitous Personalized TV-Anytime Services". In Proc. of UMICS, Velden, Austria, June 2003.

[11] Pappas N., Kazasis F., Moumoutzis N., Tsinaraki C., Christodoulakis S., "Personalized and Ubiquitous Information Services for TV Programs". In Proc. of the DELOS Workshop on Multimedia Contents in Digital Libraries, Chania, Crete, June 2003.

[12] Scherp A., Boll S., "MM4U - A framework for creating personalised multimedia content". In: Managing Multimedia Semantics. Surya Nepal, Uma Srinivasan (Hrsg.). Idea Group, Inc., 2005.

[13] Tsinaraki C., Fatourou E., Christodoulakis S., "An Ontology-Driven Framework for the Management of Semantic Metadata describing Audiovisual Information". In Proc. of CAiSE 2003, Velden, Austria, pp. 340-356, June 2003.

[14] Tsinaraki C., Papadomanolakis S., Christodoulakis S., "Towards a two - layered Video Metadata Model". In Proc. of the DEXA Workshop - DLib, 2001, Munich, Germany, pp. 937-941

[15] Tsinaraki C., Polydoros P., Christodoulakis S., "Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing". In Proc. of CAiSE 2004, Riga, Latvia, pp. 398-413, June 2004.

[16] Tsinaraki C., Polydoros P., Christodoulakis S., "Interoperability support for Ontology-based Video Retrieval Applications". In Proc. of CIVR 2004, Dublin, Ireland, pp. 582-591, July 2004.

[17] Tsinaraki C., Polydoros P., Kazasis F., Christodoulakis S., "Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content". In Special issue of the Multimedia Tools and Application Journal on Video Segmentation for Semantic Annotation and Transcoding, August 2005 (to appear).

[18] Leopold K., Jannach D., Hellwagner H., "A Knowledge and Component Based Multimedia Adaptation Framework". In Proc. of the IEEE Sixth International Symposium on Multimedia Software Engineering (MSE), December 2004.

Natural Language and Speech Interfaces to Knowledge Repositories

Anastasia Karanastasi, Stavros Christodoulakis, {allegra, stavros}@ced.tuc.gr
Laboratory of Distributed Multimedia Information Systems and Applications
Technical University of Crete (TUC/MUSIC), Chania, Greece
Silvia Gabrielli, Tiziana Catarci, {gabrielli, catarci}@dis.uniroma1.it
University of Rome 'La Sapienza',Rome, Italy,
Joachim Köhler, joachim.koehler@imk.fraunhofer.de
Fraunhofer Institute for Media Communication, Sankt Augustin, Germany,

Keywords
[1.2.1 Applications and Expert Systems] Natural language interfaces; [I.2.7 Natural Language Processing] Speech recognition and synthesis, [H.2.7 Database Administration] Data warehouse and repository, [H.5.2 User Interfaces] Evaluation/Methodology; [H.1.2 User/Machine Systems] Human factors

Research Problem

In the Digital Libraries of the future, knowledge management will be of major importance. However, traditional interfaces are particularly inflexible and difficult to use because of the highly structured and complex structures of knowledge. Natural Language Interfaces (NLI) and speech interfaces become particularly attractive for such environments. This is especially true when the knowledge repositories are accessed using mobile devices. In addition to having high importance on their own, coupling the NLI user interface style with the traditional GUI style increases the accessibility of the overall system both from the point of view of the users with disabilities and from the point of view of users working in difficult contexts such as environments of scarce lighting, or when they carry out contemporarily another task such as driving a car.

Objectives and Expected Results

The objective of this project is to provide principles, methodologies and software for the automation of the construction of natural language and speech interfaces to knowledge repositories. These interfaces include capabilities for declaration and manipulation of new knowledge, as well as querying, filtering and ontology driven interaction formulation. We will also provide a specific application demonstrator of natural language and speech interfaces to knowledge repositories and we will evaluate the approach with human subjects.

The overall technical objective is to automate as much as possible the construction of natural language interfaces to knowledge bases. It has been shown that the overhead of developing natural language interfaces to information systems from scratch is a major obstacle for the deployment of such interfaces [6]. In this design we do not specify what the storage structure for the metadata is. The metadata could be stored in a knowledge repository (such as an RDF repository) or they could be stored in relational systems provided that the inference mechanisms that support the knowledge manipulation language have been built on top of them. The natural language system will also have to take into account in addition to the concept (domain) ontologies, word ontologies (like WordNet) and the interface between the two [7].

The project will investigate a theoretical basis of the proposed approach that utilizes the domain ontologies to find how a user query in natural language can be converted to an (expanded) query in the knowledge manipulation language using the user profile and context, and allowing for the ranking of the results instead of disambiguation dialogues A preliminary restricted approach to this framework was first explored by TUC/MUSIC within DELOS and is published in [1], [2], [3].

A speech recognizer takes as input a vocabulary produced by the natural language interface subsystem that includes words representing the concepts of the domain ontologies and their relationships with the word ontologies. It uses this input to convert the speech input or a user interaction to possible phrases in natural language. The natural language phrase is processed using the user context and profile as described above for disambiguation and ranking of the results from the knowledge base.

Demonstrator Architecture

We have designed the architecture of a prototype system named OntoNL. The architecture of OntoNL has the following components:

- The *Natural Language Processing Parser*. The *NLP Parser* extracts a list of feature structures with information about the subject, the object, the verb, the prepositions, the adjectives and any and/or's, by using the Penn Treebank Tag Set [4].
- The *Dialogue Manager*. The list of stem of nouns and verbs are refined from a thesaurus (WordNet [5]) to provide the system with semantics, like senses, hypernyms and synonyms of the basic concepts within the utterance.

- The *Ambiguities Resolver*, which is the module that is responsible for the semantically disambiguation of the words in an utterance. It contains the modules: (1) The *Ontology Semantic Resolver*, which communicates with the ontologies to find similarities between the concepts in the utterance and the ontology. Based on the semantic enhancement of the words in the utterance, the system concludes in a small number of specific combinations of senses that are listed based on a weight value according to the relativeness with the domain. (2) The *User Profile Resolver*. The system checks if the user has declared any of the ambiguous words as a value in his User Profile and in what sense. The classification of the questions that apply the system has as a scope to prioritize the user preferences.(3) The *Date/Time Resolver* is responsible of the translation of subphrases about date and time in a format described by the Upper Ontology of the system.
- *Ontologies* providing knowledge to OntoNL for the semantic enhancement. Three kinds of ontologies are used: (1) an Upper Ontology (like MPEG-7), (2) a Domain Ontology, that provide vocabularies about concepts within a domain and their relationships and (3) a Functional Ontology that provides a vocabulary for the possible functions, which can be applied in a specific application.
- The *OWL Repository* that contains both the OWL class definitions and the individuals belonging to them, for the retrieval and any other functionality, like the storage or the deletion of information.
- The *Response Manager*, which is responsible for the construction of queries in a formal language to communicate with the repository, the communication with the user, the classification of the results and the presentation of them to the user, by the use of proper messages.
- The *Speech* component communicates with the natural language component to get vocabulary of the domain ontology and/or standards used. It delivers to the language component recognized tokens.

Project participants

TUC/MUSIC (Technical University of Crete, Greece) is working on building an application concerning a system that includes higher level (MPEG-7) and domain specific ontologies (soccer), and the goal is to build a Natural Language Interface for this particular application environment (for example management of a knowledge repository of MPEG-7 metadata in the domain of soccer).

Fraunhofer/IMK (Fraunhofer Institute for Media Communication, Germany) has to investigate the interplay of the speech recognition with the natural language interfaces and the ontology/ profile of users approach in a realistic application device (like handheld device). Note that the speech recognition system should be able to exploit the existence of ontologies as well as the metadata in order to improve its performance

The implemented system will be evaluated by UniRoma1 (University of Rome 'La Sapienza', Italy) through the deployment of inspection techniques (based on usability engineering principles/heuristics) and user testing in relevant application scenarios, to feed further improvements of the interface features designed.

References

1. Karanastasi, A., Kazasis, F., Christodoulakis, S., A Natural Language Model for Managing TV-Anytime Information from Mobile Devices, in Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems (NLDB), 2004
2. Karanastasi, A., Kazasis, F., Christodoulakis, S., A Natural Language Model and a System for Managing TV-Anytime Information in Mobile Environments, in Proceedings of UMICS 2004
3. Karanastasi, A., Kazasis, F., Christodoulakis, S., A Natural Language Model and a System for Managing TV-Anytime Information in Mobile Environments, to appear in the ACM/Verlag Personal and Ubiquitous Computing Journal, Volume 8, 2004
4. Marcus, M., Santorini, B. and Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. In Computational Linguistics, volume 19, number 2, pp313-330
5. Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.J., `Introduction to WordNet: an on-line lexical database.', International Journal of Lexicography, 3(4), 235 -- 244, (1990)
6. Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pfleger, N., Poller, P., Streit, M., Tschernomas, V., "Smartkom – Adaptive and Flexible Multimodal Access to Multiple Applications", in Proceedings of the 5th international conference on multimodal interfaces. Vancouver BC. 5-7 Nov 2003, Canada
7. Vargas-Vera, M., Motta, E., AQUA – Ontology-based Question Answering System. Third International Mexican Conference on Artificial Intelligence (MICAI-2004), Lecture Notes in Computer Science 2972 Springer Verlag, 2004

User-Interface and Visualization

Cluster objectives
The notion of a "Digital Library" (DL) is currently associated with technological and scientific efforts to build, maintain, and use large collections of electronic documents. However, it can also be regarded as a cornerstone in the construction of an information-enriched environment. Once this broader perspective is adopted, a variety of problems arise which will have to be solved in order to ensure the usability and accessibility of this environment to different users with varying needs and capabilities for both professional and recreational purposes. The ultimate goal of the User-Interface and Visualization cluster is to develop methodologies, techniques and tools to enable future DL designers and developers to meet not only the technological, but also the user-oriented requirements in a balanced way.

Cluster activities
User Requirement related Activities

- *Systematic study of user requirements*. The different perspectives on a digital library are being analysed to relate them to the requirements and technical implementation options that emerge from the ongoing development projects of the NoE partners.
- *Analysis of user-related aspects in the development and usage of a DL system*. The analysis will not focus only on the DL end user but will also take into account other DL stakeholders such as librarians, content providers and maintainers. The DL life cycle will be related to functional and non-functional requirements.
- *Characterization of DL users.* The characterization will take into account that the user interface accords accessibility for all categories of users, including users with special needs. In addition, this cluster will also explore how users can exploit a multi-modal DL user interface to meet their particular needs.

User Interface and Visualization Design Activities:

- Development of *a taxonomy of relevant context models.* A language specification is being investigated, which shall encompass the pertinent characteristics and requirements of context models that were identified during the development of the taxonomy.
- Development of a *comprehensive model for relevance criteria*. The consequence of taking the usage situation/context into account results in rethinking the basic assumptions underlying most contemporary approaches to information filtering and retrieval. This should lead to more realistic definitions of "relevance".
- Development of a theoretical *framework.* from which user interface designers/developers can design DL user interfaces. The designer/developer gathers various resources provided by the theoretical framework (e.g. methodologies and tools) and designs a DL user interface (e.g. tailored for some particular application domain). Moreover, taking into account that future DL solutions will have to provide integrated customizable components that cover the appropriate functionality needed in a given context, the ultimate goal is to *develop a design methodology and guidelines* that, starting from a generic user interface, will allow to define tailored technical solutions that can be implemented in a given scenario, starting from the users needs.

Cluster coordinator
Tiziana Catarci, University of Roma 1

Visualization in DL Systems (Relevance feedback)

Claudia Plant (UMIT), **Enrico Bertini** (Roma1), **Stefano Berretti** (UNIFI-MICC)

Problem Specification

The major challenge in content-based multimedia retrieval is the so-called "semantic gap" between machine computed similarity and human perception. Multimedia objects, e. g. images are represented in the system by high-dimensional feature vectors consisting of plain characteristics of the objects such as color histograms, wavelet coefficients, Fourier descriptors. Objects are represented as points in a high-dimensional vector space. In principle similarity between objects can be easily expressed using a metric distance function. But not only the curse of dimensionality complicates the retrieval process. The main problem is that users identify similar objects on a semantically higher level, e.g. a user might be looking for images portraying the same person. To bridge this semantic gap, relevance feedback mechanisms have been successfully applied. This task will investigate novel methods of combining relevance feedback mechanisms with advanced visualization concepts of query results.

Relevance Feedback Mechanisms

In content-based multimedia retrieval, e.g. image retrieval, the most common paradigm is *query-by-example*. Starting with a query image, the goal of the relevance feedback process is to find similar images. The query image can be brought from outside to the system but is more frequently one of the database objects. Retrieval accuracy and the number of required relevance feedback cycles strongly depend on an appropriate choice of the starting image. We want to investigate how data mining techniques, such as clustering methods, and advanced visualization techniques can be used to give the user an overview on the database content and facilitate the choice of the starting image. In the subsequent iterations, in most of the existing systems the top *k* of the retrieved images are presented to the user in an ordered list and relevance feedback can be given by marking objects as positive or positive and negative examples [1]. We go beyond this binary classification allowing continuous degrees of relevance. Currently we are experimenting with giving relevance feedback by moving objects and interactive clustering in a 2D user-interface. Similar as in [2], objects can be moved towards or away form the query object. In addition, the displayed result set is clustered giving the user an impression of the current similarity measure. In an interactive process the user groups similar objects together, thus changes the cluster structure of the displayed objects. In this interactive process the system adapts the similarity measure to the users need (cf. figure 1). Another interesting aspect is implicit feedback since we assume that the user considers the objects he moves to be important in some sense.

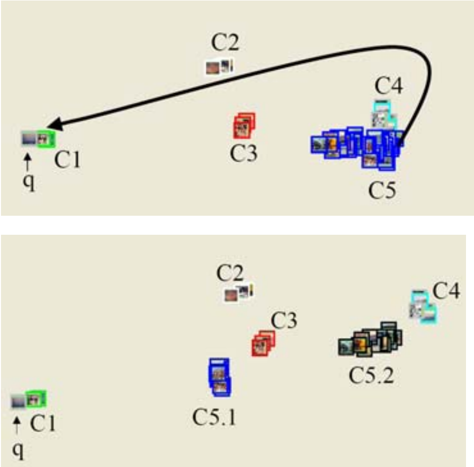


Figure 1: Interactive Clustering

Visualization of Multimedia Objects for Relevance Feedback

Information Visualization can provide extensive support to the visualization of multimedia objects for relevance feedback, being its techniques aimed at providing compact meaningful representations of data objects and having a special focus on interaction. We have found five major potential interventions. (i) Apply existing visualization techniques, but exploring new ways of giving relevance feedback: searching for suitable visual techniques for representing multimedia objects drawing from the large set of existing infovis techniques to see how suitable they are for the representation of multimedia objects; (ii) Enhance interaction capabilities for relevance feedback: the existing interaction methods for relevance feedback are quite limited, basically implemented as simple selections, therefore more complex and semantically rich interactions can be employed like, e.g., dynamic filtering as a way to focus on relevant subsets of data objects; (iii) Exploring new ways for obtaining overviews for browsing and exploration tasks: a large number of visualization techniques are aimed at providing overviews and many of them can be employed for multimedia objects for providing meaningful arrangements of visual items that permits to easily detect interesting subsets of data and trends; (iv) Representing interaction data: since relevance feedback is based on the idea of taking into account past user interaction as a way to personalize data output, it is worth to investigate ways to represent already manipulated objects visually; (v) Apply sampling-based techniques for providing accurate visualization: in order to deal with visualizations with to many data items (i.e., cluttered visualizations) sampling based techniques can be employed to reduce data density an make visual representations more intelligible.

Mathematical Foundations to Incorporate Relevance Feedback

Relevance feedback information can be used to adjust a weighted distance function order to emphasize those features separating positive and negative examples. Heuristic methods for global independent axis weighting have been proposed [1]. In [2] new weights are determined by solving an optimization problem minimizing the deviation of the coordinates from the 2D display space and the high dimensional feature space. Many approaches also regard incorporating relevance feedback as a classification problem, e.g. [3]. These approaches face the problem that too few training instances are available since only a very limited amount of objects can be examined by the user. Some of the classifiers, such as Support Vector Machines and Linear Discriminant Analysis are originally designed for two-class problems, i.e. they can only support binary relevance feedback information. Recently there has been some interesting work on semi-supervised clustering. Semi-supervised clustering algorithms can deal with the small amount of examples examined by the user. In [4] a semi-supervised EM-algorithm for image retrieval has been proposed. Due to the drawbacks of the underlying clustering algorithm this method assumes Gaussian distributed data and the number of clusters has to be specified in advance. We are experimenting with density-based clustering and local feature weighting currently done using Linear Discriminant Analysis [5]. We determine different weights between the query and the different clusters of the displayed result. This approach needs to be further elaborated and can be extended to non-linear feature space transformation using kernel functions. We are also investigating methods to directly incorporate relevance feedback information into density-based clustering. Another interesting direction to elaborate is to replace the distance function by a shortest path algorithm and modify the adjacency matrix of the objects after relevance feedback.

Extensibility to Distributed Environments

Relevance feedback techniques have been usually studied and applied in centralized scenarios, where users and databases are located at an individual, local site. Actually, the increasing availability of multiple repositories accessible in a distributed environment over the Network, poses additional issues to relevance feedback. In particular, searching information through the Internet often requires users to separately contact several digital libraries, use each library interface to author the query, analyze retrieval results and merge them with results returned by other libraries. Such a solution could be simplified by using a centralized server that acts as a gateway between the user and several distributed repositories: The centralized server receives the user query, forwards the user query to federated repositories - possibly translating the query in the specific format required by each repository - and fuses retrieved documents for presentation to the user. To accomplish these tasks efficiently, the centralized server should perform some major operations, such as: *resource selection*, *query transformation* and *data fusion*. Resource selection is required to forward the user query only to the repositories that are candidate to contain relevant documents. In this operation, relevance feedback could be used to adaptively change the selection process. Also in the data fusion phase, relevance feedback could be used to modify the fusion process according to different relevances associated to the repositories in successive queries. In this scenario, we will investigate if current relevance feedback techniques, targeted for centralized applications, can be easily and effectively expanded in order to account for distributed contexts or if specifically tailored relevance feedback techniques should be developed.

Project participants

Universita' di Roma "La Sapienza", Italy (Roma1)
National and Capodistrian University of Athens, Greece (UOA)
University for Health Sciences, Medical Informatics and Technology, Hall in Tyrol, Austria (UMIT)
Institut National de Recherche en Informatique et en Automatique, France (INRIA)
Universita' degli Studi di Firence, Italy (UNIFI-MICC)

References

[1] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., Jain, R.: “Content-Based Image Retrieval at the End of the Early Years.” IEEE Trans. Pattern Analysis Machine Int. 22:12, 2000.

[2] Santini, S., Gupta, A., Jain R.: “Emergent Semantics through Interaction in Image Databases”. IEEE Trans. Knowl. Data Eng. 13:3, 2001.

[3] Hong, P., Qi, T., Huang, T. S.: “Incorporate Support Vector Machines To Content-Based Image Retrieval With Relevance Feedback.” Proc. of Image Processing, 2000.

[4] Dong, A., Bhanu, B.: “A New Semi-Supervised EM Algorithm for Image Retrieval”. Proc. of CVPR 2003

[5] Damjanovic, D., Plant, C., Balko, S., Schek, H. J.: ‟User-Adaptable Browsing and Relevance Feedback in Image Databases.” In: Proc. DELOS Workshop on Future Digital Library Management 2005.

User Requirements-driven Support for a DL Design Framework

Davide Bolchini – University of Lugano
Tiziana Catarci – UniROMA1
Norbert Fuhr, Saadia Malik - University of Duisburg
Margherita Antona – FORTH-ICS
Annelise Pejtersen - RISOE

Keywords
H.4.3 [Information Systems Applications]: Communications Applications – Internet; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia – user issues.

Research Problem

Despite the rapid ongoing technological evolution in recent years (the success of the World Wide Web, the diffusion of various kinds of interactive applications, and the availability of different end-user devices), DL interfaces are still based mainly upon "search" and "search refinement" mechanisms. During the first 18 months of the DELOS NoE, the work of WP4 "User Interfaces and Visualization" has focussed on identifying functional and non-functional requirements of Digital Libraries (DLs), with the aim to establish an empirical basis for user interface design for DLs and to define a taxonomy of functional DL infrastructure and visualization paradigms. Additionally, a preliminary DL usage lifecycle model has been elaborated, targeted to facilitate, in the light of the different usage phases that characterise the long-term life and evolution of DLs, further analysis of user requirements. The results of such an empirical study calls for the investigation of the potential effectiveness and benefits to the user stemming from a full adoption of alternative interaction paradigms, and especially of novel techniques for navigation such as browsing by catalogues, semantic linking, information visualization, interactive maps, social navigation, etc., which are seldom and occasionally employed in current DLs.

Objectives

This proposal aims at extending and enhancing the results of previous efforts in DELOS towards the systematic investigation of non-conventional interaction paradigms, and the correlation of such paradigms with different usage phases of DLs.

In particular, the research objectives to be pursued are the following:

1. Further extend the empirical analysis of user functional and non-functional requirements, and further analyse, refine and enhance the preliminary life-cycle model in the context of the continued collection and analysis of empirical data.
2. Specify the user requirements regarding novel interaction paradigms in addition to the ones traditionally employed, such as "search", "querying" and their variations.
3. Define advanced interaction paradigms for DL and thereby build a "theoretical" framework for the design of new DL interfaces.
4. Develop a prototype (on top of already existing DL's, or for new ones) demonstrating the new concepts and mechanisms.
5. Test the effectiveness and usability of the prototype against the needs of selected user communities.

The work done

- Definition of the operative proposal, agreement and discussion of the objectives and workplan (ALL)
- Investigation and review of existing literature on non-conventional access paradigms (UNISI)
- Development of "accessible" access paradigms, specifically tailored for visually-impaired users (UNISI)
- Specification of user requirements for non-conventional interaction paradigms (UNIROMA1)
- Empirically motivated model of the user experience lifecyle in DLs (FORTH-ICS)
- Surveying the existing literature on non-conventional access paradigms such as browsing, navigation to target open-ended exploration in context of structured documents. So far user studies have been carried out for the conventional search paradigm in the INEX 2004 interactive track. In INEX 2005, the user studies will be continued in same direction.(UNIDUE)

Expected Results

D1a: Report on DL usage lifecycle, with a particular emphasis on user requirements for effective access paradigms (preliminary version): Month 10

Some examples of non-conventional interaction paradigms to explore:

• *Accessibility strategies*: reading strategies for visually-impaired.

• *Catalogues browsing*: Use of alternative taxonomies to organize information
• *Open-ended navigation*: learn/discovers by free exploration, supporting complex / ill-defined user goals
• *Semantic linking*: strategies for navigation based on the exploitation of semantically related pieces of content, supporting "associative" or "serendipitous" thinking.

Some paradigms will be selected for their feasibility and relevance to the DL domain. Each paradigm may include:

– User requirements addressed (Why should I care?)
– Lifecycle context (When do I need it?)
– Key concepts or metaphors used ("in a nutshell")
– Typical user scenarios and profiles met (what for?)
– High-level design solutions (how does it work?)
– Prototype example (At the end?)
– Additional Resources: e.g. for implementation, user validation, literature, examples in other domains. (To know more?)

D2: Demonstrative Prototype: Month 14

D1b: Access paradigm refined (final): Month 17

Dissemination of the results: Journal papers (e.g. IJDL), conference papers (e.g. ECDL, …), and proactive presence in leading workshops (e.g. CHI, HCII, HT, ER, RE, WISE, …).

Project participants

UNISI (University of Lugano) - Davide Bolchini. UNISI matured expertise in user requirements analysis and methodologies for user-centered design and usability evaluation. UNISI is partner of WP4.

UNI ROMA 1 (Università di Roma 1) - Tiziana Catarci. UNIROMA 1 has a long research tradition in user interface design, human-computer interaction, information visualization and is the cluster coordinator of WP4 (User Interfaces and Visualization) in DELOS.

UNIDU (University of Duisburg-Essen) – Saadia Malik. UNIDU is leader of WP7, co-leader of INEX, coordination of INEX interactive track.

FORTH-ICS - Margherita Antona. The Human Computer Interaction Laboratory of FORTH-ICS has recognised research expertise in user interface software technologies, design methodologies, and software tools.

The Laboratory carries out leading research activities focused on developing user interfaces for interactive applications and services that are accessible, usable, and ultimately acceptable for all users in the Information Society..

RISOE - Annelise Pejtersen. Task leader in WP4, responsible for provision of an empirical basis. Risoe has a long-standing research tradition and background in Human Computer Interaction, interfaces for semantic navigation and Computer-Supported Collaborative Work.

References

- Bolchini, D., Paolini, P., Interactive Dialogue Model: a Design Technique for Multi-Channel Applications, accepted for publication on IEEE Transactions on Multimedia, 2005 (to appear).
- Bolchini, D., Paolini, P., Goal-Driven Requirements Analysis for Hypermedia-intensive Web Applications, Requirements Engineering Journal, Springer, RE03 Special Issue (9) 2004: 85-103.
- Bertini, E., Catarci, T., Di Bello, L., & Kimani, S., Visualization in Digital Libraries. From Integrated Publication and Information Systems to Information and Knowledge Environments, Hemmje, M., Niederee, C., & Risse, T. (Ed.s), Springer-Verlag Berlin, 2005: 183-196.
- Mirabella, V., Kimani, S., Gabrielli, S., & Catarci, T., Accessible e-Learning Material: A No-Frills Avenue for Didactical Experts. NRHM (New Review of Hypermedia and Multimedia) journal, Special Issue on Accessible Hypermedia and Multimedia. Taylor & Francis Group, 10(2) 2004: 165-180.
- Antona, M., Mourouzis, A., Kartakis, G., Stephanidis. C. (2005, in print).User requirements and usage life-cycle for digital libraries. Proceedings of the 11th International Conference on Human-Computer Interaction (HCI International 2005), 22-27 July 2005, Las Vegas, Nevada, USA.
- Anastosios Tombros, Saadia Malik; Birger Larsen (2005). Report on the INEX 2004 interactive track. SIGIR Forum 39(1).
- Anastasios Tombros; Birger Larsen; Saadia Malik (2005). The Interactive Track at INEX 2004. In Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Lecture Notes in Computer Science 3493, Springer-Verlag GmbH.

Task-centered Information Management

Tiziana Catarci – Università di Roma "La Sapienza"
Alan Dix – University of Lancaster
Yannis Ioannidis - University of Athens

Keywords
H.4.3 [Information Systems Applications]: Communications Applications – Internet; H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia – user issues.

Research Problem

In order to effectively translate digital information into human knowledge it is critical to envision, develop, and deploy applicable human-digital library interfaces. These must be interfaces that not only mimic established learned human behavior in regards to reading and comprehending printed material, but interfaces that also endeavor to meld the human perceptual, cognitive, and social states to the applied digital library technologies in order to synergistically enable future human learning.

Despite the need for such advanced interfaces, digital library (DL) interaction is still completely WIMP-oriented. The WIMP (windows, icons, menus, pointers) interface, based on the desktop metaphor, first appeared in the commercial marketplace in April 1981, when Xerox introduced the 8010 Star Information System. However, many of the underlying windowing techniques were used some years before in Engelbart's group in NLS and at Xerox Parc in the experimental precursor to Star, the Alto.

So, approximately thirty years have passed since the introduction of windows, menus, folders and pointers as interaction mechanisms. Nothing has changed in these mechanisms since then. Another quite old approach has to do with the modeling of the interaction. Indeed, traditionally the user initiates precise actions and the system reacts to such actions. The interaction is totally operational and the execution of tasks through a sequence of actions is completely in charge of the user.

Objectives

The "Task-centered Information Management" (TIM) joint plan of activity focuses on the definition of an innovative approach to the management of DL content and services that builds on two key issues:

1. ontology-based information classification, where the basic idea is to substitute the tree-based classification of documents with a DAG-based classification or other rich structure supported by an ontology;
2. task-oriented interaction, where the basic idea is to have the system learn the way in which users perform typical tasks (e.g. filling out a reimbursement form) and then providing a task-based environment in which the necessary operations are almost automatically performed by the system..

The work done

We decided to focus on the ontology-based information classification during the first year and defined the workplan (ALL).

Analysis of the different proposals of ontology definition languages and choice of the most appropriate for our specific case, taking into account that the ontology will contain, among other things, information at different granularity level, temporal information, representation of the users' tasks (ROMA 1).

Modification of automatic classification algorithms currently classifying into simple hierarchies to work with more complex ontologies. This will be used to either automatically classify filed information or to suggest classifications to users at the time of filing. (UoLanc).

Literature review and empirical studies of the way people 're-find' previously seen material in particular web bookmarks. This will help uncover existing classification schemes . (UoLanc).

Literature survey on the definition, representation, and evolution of "personal" ontologies. Analysis of different existing approaches, adaptation to the TIM goals, and investigation of new ones for their effective and efficient use.

Expected Results

- "Personal" ontology definition and manipulation language.
- Design and prototype implementation of a visual interface to support the user in the manual modification/extension of a predefined ontology that s/he is supposed to get as a starting point.
- Design and prototype implementation of set of different tools to be exploited in the semi-automatic modality of ontology construction. In particular, the following tools are foreseen:

    - tool for the automatic document classification and concept extraction;
    - tool for instance reconciliation;
    - personalization tool based on user profiling;
- Definition of the TIM architecture, in particular with respect to the "semantic save" layer.
- Formal task-definition language

Project participants

ROMA 1 (Università di Roma "La Sapienza") – Tiziana Catarci. ROMA 1 is the cluster coordinator of Delos WP4 "User Interfaces and Visualization". In addition to a long tradition of research on hci, usability, information access and visualization, ROMA 1 has also significant expertise on data integration, data quality, e-services.

UoA – Yannis Ioannidis. UoA is the cluster coordinator of Delos WP2 "Information Access and Personalization". Recently, it is active in work related to various forms of personalization and customized behavior, whereas in the past it has been involved in visual modeling and information visualization issues. Other areas of UoA expertise include query processing and optimization, data integration, and scientific systems.

UoLanc (Lancaster University) – Alan Dix. UoLanc also brings expertise in interface modelling and design, in particular relating to mobile devices and web cognition. It also brings experience in using automatic inference in creating user interfaces that embody 'appropriate intelligence' – AI within well-matched interactive frameworks. UoLanc is also a member of the UK Memories for Life Network looking at long-term storage and retrieval of personal memories.

References

- Tiziana Catarci, Paolo Dongilli, Tania Di Mascio, Enrico Franconi, Giuseppe Santucci, Sergio Tessaris, An ontology based visual tool for query formulation support, Proc. of the16th European Conference on Artificial Intelligence (ECAI2004), Spain, 2004.
- Enrico Bertini, Andrea Cali', Tiziana Catarci, Silvia Gabrielli, and Stephen Kimani - Interaction-based Adaptation for Small Screen Devices. – Proc. of the 10th Int. Conf. On User Modeling (UM 2005), UK, 2005.
- Alan Dix, Andrew Howes, Stephen Payne. Post-web cognition: evolving knowledge strategies for global information environments International Journal of Web Engineering Technology, Vol. 1, No. 1, 2003. pp. 112-126.
- Alan Dix. The ultimate interface and the sums of life?. Interfaces, no 50, Spring 2002. pp. 16. http://www.hcibook.com/alan/papers/dust2002/
- Georgia Koutrika and Yannis Ioannidis, "Constrained Optimalities in Query Personalization", Proc. of the 2005 ACM SIGMOD Conference, June 2005, Baltimore, MD, pp. 73-84.
- Georgia Koutrika and Yannis Ioannidis, "A Unified User-Profile Framework for Query Disambiguation and Personalization", Proc. of PIA 2005 - Workshop on New Technologies for Personalized Information Access (in conjunction with UM 2005), July 24-25, 2005, Edinburgh, Scotland, UK.

Design, Implementation, and Evaluation of the Use of Annotations in Interactive and Collaborative DL Access

Maristella Agosti, Nicola Ferro, {agosti, ferro}@dei.unipd.it
Department of Information Engineering – University of Padua – Italy,
Hanne Albrechtsen, hanne.albrechtsen@risoe.dk
Risoe National Library – Denmark
Ingo Frommholz, ingo.frommholz@uni-due.de
University of Duisburg-Essen – Duisburg – Germany
Emanuele Panizzi, panizzi@di.uniroma1.it
University of Rome "La Sapienza" – Roma – Italy,
Ulrich Thiel, thiel@ipsi.fraunhofer.de
Fraunhofer IPSI – Darmstadt – Germany

Project description

In most contemporary digital library (DL) management systems the contents are conveyed to the user as a "collection of information items" which can be searched or browsed. However, this paradigm is often not sufficient to cope with embedded usage, for which access to the contents is not seen as an isolated activity, but as part of a larger work process, where tasks like interaction with other users, extraction of knowledge, and analysis or evaluation of documents need to be integrated. Most of these scholarly activities result in new texts or multimedia objects which refer to the already existing documents as "annotations".

Up to now, annotations have been - in most cases - stored together with the documents they are related to in a central DL repository. With the advent of decentralised DL architectures in Grid or Peer-to-Peer environments, but also in Service-oriented architectures, these design choices need to be revised by solutions which allow us to manage annotations independently from a particular DL management system.

Thus, the main goals of this project are:

- to develop an annotation DL service and to define a set of API to allow the access to this service from different DLs. The annotation service will also provide an infrastructure for advanced annotation-based retrieval functionality;
- to integrate the service into the DAFFODIL system (http://www.daffodil.de) and the BRICKS (http://www.brickscommunity.org) system;
- to evaluate the use of the annotation system as a new way to interact with a DL and to establish collaboration among DL users and stakeholders, performing a study of their behaviour, of the system usability, and of the impact on the DL development and use.

The proponents, listed below, can base their work on experiences gathered in developing a general purpose annotation system, called MadCow [6], and a collaboratory system for the Humanities (COLLATE) [8, 9].

FhG/IPSI	Ulrich Thiel	thiel@ipsi.fraunhofer.de
Risoe	Hanne Albrechtsen	hanne.albrechtsen@risoe.dk
	Annelise Mark Pejtersen	amp@risoe.dk
Roma1	Emanuele Panizzi	panizzi@di.uniroma1.it
SICS	Preben Hansen	preben@sics.se
UniDU	Ingo Frommholz	ingo.frommholz@uni-due.de
	Norbert Fuhr	fuhr@uni-duisburg.de
UniPD	Maristella Agosti	agosti@dei.unipd.it
	Nicola Ferro	ferro@dei.unipd.it

They already collaborated on the definition of a comprehensive annotation model [1, 3] and the architecture for an annotation service [2] which will serve as design guidelines for the specification and implementation of a DL annotation service compliant with contemporary interface standards. Finally, the proponents have cooperated in proposing methods which integrate annotations in the information retrieval process [3, 4, 7, 8, 9].

The most important requirements that should be fulfilled by the annotation service are the following:

- it should support nested annotations, meaning that not only documents or document parts can be annotated, but also other annotations;
- each annotated object must be referable by an handle – for example, the Uniform Resource Identifier (URI) will be one of the schemes to be supported;
- sign (e.g., textual, graphical, referential or a combination of these) and meaning of an annotation should be represented. This way, different annotation types can be supported;
- different scopes of annotations (private, public, shared) must be considered;
- support for annotation indexing – as annotation threads made of nested annotations form a linked structure, some basic functionality to support incremental indexing of annotations and the structural context should be provided. Services indexing the annotation corpus for annotation-based retrieval should be able to access the repository appropriately.

A test of the usage of annotations by DL users is going to be performed in order to assess:

- how and to what extent DL users annotate the DL content pages for their personal use;
- how and to what extent do they co-operate using annotations;
- if DL users search annotations and navigate related annotations in order to discover new content.

References

1. M. Agosti, N. Ferro. Annotations: Enriching a Digital Library. In: T. Koch, I.T. Sølvberg (Eds). *Proc. of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2003)*. Springer LNCS 2769, 2003, 88-100.
2. M. Agosti, N. Ferro. An Information Service Architecture for Annotations. In [5], 115-126.
3. M. Agosti, N. Ferro, I. Frommholz, U. Thiel. Annotations in Digital Libraries and Collaboratories - Facets, Models and Usage. In: R. Heery, L. Lyon (Eds). *Proc. of the 8th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2004)*. Springer LNCS 3232, 2004, 244-255.
4. M. Agosti, N. Ferro. Annotations as Context for Searching Documents. In F. Crestani, I. Ruthven (Eds). *Proc. 5th International Conference on Conceptions of Library and Information Science (Colis5)*. Springer LNCS 3507, 2005, 155-170.
5. M. Agosti, H.-J. Schek, and C. Türker (Eds). *Digital Library Architectures: Peer-to-Peer, Grid, and Service-Orientation, Pre-proceedings of the 6th Thematic Workshop of the EU Network of Excellence DELOS*. Edizioni Libreria Progetto, Padova, Italy, 2004.
6. P. Bottoni, R. Civica, S. Levialdi, L. Orso, E. Panizzi, R. Trinchese. Digital Library Content Annotation with the MADCOW System. In T. Catarci, S. Christodoulakis, and A. Del Bimbo (Eds.). *Proc. Audio-Visual Content and Information Visualization in Digital Libraries (AVIVDiLib'05)*. Centromedia, Viareggio, Italy, 2005, 111-116.
7. P. Knezevic, B. Mehta, C. Niederée, T. Risse, U. Thiel, I. Frommholz (2004). Supporting Information Access in Next Generation Digital Library Architectures. In [5], 49-60.
8. I. Frommholz, U. Thiel, and T. Kamps. Annotation-based Document Retrieval With Four-Valued Probabilistic Datalog. In T. Rölleke and A.P. de Vries (Eds.). *Proc. First SIGIR Workshop on the Integration of Information Retrieval and Databases (WIRD'04)*, 2004, 31-38.
9. U. Thiel, H. Brocks, I. Frommholz, A. Dirsch-Weigand, J. Keiper, A. Stein, E. Neuhold. COLLATE - A collaboratory supporting research on historic European films. *Int. J. on Digital Libraries* 4(1): 8-12 (2004) .

Knowledge Extraction and Semantic Interoperability

Cluster objectives

The thematic area of Semantic Interoperability is growing in importance in digital library (DL) research (taking the interpretation of "digital library" at its broadest). It applies to the application of different vocabularies and terminology used in descriptions of digital objects for both learning and research, collections of those objects, collections of datasets and resources used in the wider cultural heritage sector and in e-research. Indeed, cross-sectoral and cross-domain shared understanding of semantic descriptions is one of the goals of the Semantic Web as envisaged by Tim Berners-Lee in his roadmap published in 1998. This vision has more recently (2001) been applied to "Grid computing" and e-science / e-research initiatives in the Semantic Grid approach. In addition, the application of algorithms for the mining and analysis of digital resources (text, data, complex objects), offers exciting opportunities for the extraction of new knowledge and the re-use of data and information in new ways.

The Knowledge Extraction and Semantic Interoperability research cluster has two key strategic goals:

- To co-ordinate a programme of activities which brings together research excellence from a range of inter-related knowledge engineering and information management areas, and which facilitates the sharing of experience and expertise amongst practitioners from both DL and Grid/computing science backgrounds.
- To explore the potential of new models, algorithms, methodologies and processes in a variety of technical applications, institutional frameworks and cross-sectoral environments, which will lead to the creation of guidelines and recommendations of best practice for dissemination to the widest possible community of interest.

Cluster activities

A Forum has been created to provide a physical and virtual arena for the exchange of experience and research in all the areas/themes of this cluster. It provides an opportunity to integrate systematically other relevant groups into the cluster. This development is being supported by a moderated virtual forum or discussion list for the expansion of discussion on selected topics. It is also intended to maximise opportunities to harmonise with other relevant initiatives such as CIDOC and FRBR.

In the area of Knowledge Extraction, a study has been produced to determine the requirements for and usage of extracted knowledge for biblio-metrics, domain analysis, issue tracking and community modelling

Semantic Interoperability is being addressed by scoping the area with the aim of producing a state-of-the-art overview of DL semantic issues including the application of standards, thesauri, ontologies, Knowledge Organisation Systems and the implementation of metadata schema registries.

Cluster coordinator

Elizabeth Lyon, UKOLN, University of Bath.

Interoperability of eLearning Applications with Audiovisual Digital Libraries

Stavros Christodoulakis, Polyxeni Arapi, Nektarios Moumoutzis, {stavros, xenia, nektar}@ced.tuc.gr
Laboratory of Distributed Multimedia Information Systems and Applications
Technical University of Crete (TUC/MUSIC), Chania, Greece,
Manjula Patel, m.patel@ukoln.ac.uk
University of Bath (UKOLN), Bath, UK
Sarantos Kapidakis, Antonia Arahova, Lina Bountouri, {sarantos, tonia, boudouri}@ionio.gr
Department of Archives and Library Science, Ionian University, Corfu, Greece

Keywords
[H.3.7 Digital Libraries] Standards, [D.2.12 Interoperability], [K.3.1 Computer Uses in Education] Distance Learning

Research Problem

Digital Libraries are an important source for the provision of eLearning resources [McLean, 2004]. However, there is a problem in that digital library metadata standards and eLearning metadata standards have been developing independently, presenting interoperability issues between digital libraries and eLearning applications. This is a critical multi-facet problem that needs to be addressed systematically over the next few years.

In this project, we concentrate on the interoperability of audiovisual digital libraries and eLearning applications, in order to support the modular development of personalized learning experiences. The aim is to develop a robust model that will allow for the use of semantic descriptions of audiovisual content objects and/or segments that reside on a digital library in order to support the creation of reusable learning objects that may be used in the assembly of personalized courses. In addition, the model will allow users that participate in eLearning activities, to browse and retrieve audio-visual objects from digital libraries that match their interests, and use them as learning resources.

Objectives and expected results

The related standards for digital libraries (METS, Dublin Core, etc.), eLearning (SCORM) and audiovisual content description (MPEG7) will be comparatively studied and an appropriate interoperability framework will be developed for their usage. A demonstrator will be implemented supporting the above interoperability framework using state of the art service oriented architectures.

This work will also include the development of a Learner Model for generating learner profiles (IMS LIP, IEEE PAPI) for use in creating personalized learning experiences, as well as an investigation into various options for packaging objects and the related metadata using for example, METS, MPEG-21 DIDL or IMS content packaging, to see which is best suited for the exchange of records within the above interoperability framework..

The project will also measure the reactions of different classes of users and their acceptance of the proposed model. The user groups that will be addressed include digital library users as well as users that participate in distant learning activities taking into account the level of metadata integration with the digital library content. The project will undertake the responsibility of the definition of detailed measurements of the user reactions and acceptance of the proposed approach, the definition of effectiveness criteria. It will also set forth suggestions for improvement and refinement of the proposed approach based on the evaluation result.

Demonstrator architecture

We have designed the architecture of demonstrator named the Personalized Audiovisual Learning Experiences System (PALES). The architecture of PALES conforms to the IMS Digital Repositories Interoperability (IMS DRI) Specification [IMS DRI, 2003]. The DRI specification acknowledges a wide range of content formats and is applicable internationally to both learning object repositories, as well as to other traditional content sources, such as libraries and museum collections.

The components of the PALES architecture are the following:

- The *Digital Library* (provision service), which is logically separated in two parts: (1) the Audiovisual Objects part, where audiovisual content along with metadata descriptions (MPEG7) are stored (2) the Learning Objects part, where learning objects along with metadata descriptions are stored in SCORM format. This part could be used for storing the generated personalized learning experiences for a later use.
- *Applications* (Software Agents in terms of IMS DRI, like Learning Content Management Systems, Learning Management Systems etc.) that discover, access and use the content of the audiovisual content of the digital library through appropriate services (resource utilizers).

- The *Middleware* comprising the essential component of the architecture, which is responsible for the assembly of personalized learning experiences. The middleware consists of the following parts:
    - The *MPEG7/SCORM transformation component*, which is responsible for the transformation of the MPEG7 descriptions of the selected resources to the LOM format and the construction of IMS Content Packages from the final decided hierarchy corresponding to the learning experience, and
    - The *Personalized A/V Learning Experiences Assembler (PALEA)*, which, taking into account the knowledge provided by the *Learning Designs* and the *Learner Profiles* described later, constructs the personalized learning experiences and delivers them in the form of IMS Content Packages.
- *Ontologies* providing knowledge to the PALEA for the automatic construction of personalized learning experiences: (1) *Domain Ontologies* that provide vocabularies about concepts within a domain and their relationships, and (2) the *Instructional Ontology* that provides a vocabulary for the instructional function of a resource (e.g. Example, Motivation, Explanation) and also a vocabulary for the construction of possible pedagogical approaches (instructional strategies/didactical templates), which can be applied to the construction of learning experiences. Audiovisual objects are indexed using both the domain ontologies and the instructional ontology to capture both their semantic meaning and their learning role.
- *Learning Designs* are abstract training scenarios in a certain domain combining information taken from the corresponding domain and the instructional ontology.
- *Learner Profile Ontology* that represents a learner model for the creation of learner profiles.
- *Model for Usage-Oriented Evaluation* of eLearning applications with audiovisual digital libraries, that support the modular development of personalized learning experiences inside a virtual environment according to specific measurement strategy.

Project participants

TUC/MUSIC (Technical University of Crete, Greece) has already elaborated an integration framework between SCORM and TV-Anytime [Frantzi et al., 2004] and has also worked on interoperability issues with respect to the use of SCORM in existing eLearning repositories [Arapi et al., 2003]. In this project it will be engaged in the development of the interoperability framework as well as in the development of the instructional ontology and the the demonstrator that will implement the interoperability framework on top of a service-oriented architecture.

UKOLN (University of Bath, UK) will investigate issues relating to personalization and the development of a Learner Model and Learner Information Profiles, as well as interoperability issues related to the packaging and exchange of resources between digital library systems and learning object repositories.

IU (Ionian University, Greece) will construct a domain ontology to be used in the demonstrator. Moreover, it will undertake the responsibility of the definition of detailed measurements of the user reactions and acceptance of the proposed approach, the definition of effectiveness criteria. IU will also set forth suggestions for improvement and refinement of the proposed approach based on the evaluation results.

References

Arahova, A., Kapidakis, S. (2004). National Libraries and Academic Libraries: Partners in E-educational National and Global Action. IFLA/FAIFE 2004, Bulgaria

Arahova A., Kapidakis S.(2004). National Libraries: A Perspective for a Leading Role in the E-Services Époque. International Conference on National Library Services (ICONLIS 2004), India

Arapi P., Moumoutzis N. Christodoulakis S. (2003). Supporting Interoperability in an Existing e-Learning Platform using SCORM. ICALT 2003. Greece.

Frantzi M., Moumoutzis N., Christodoulakis S. (2004). A Methodology for the Integration of SCORM with TV-Anytime for Achieving Interoperable Digital TV and e-Learning Applications. ICALT 2004 . Finland.

IMS DRI (2003). IMS Digital Repositories Interoperability Specification, URL: http://www.imsglobal.org/digitalrepositories/driv1p0/imsdri_infov1p0.html

McLean N. (2004). The Ecology of Repository Services: A Cosmic View. Keynote Address. ECDL 2004. UK.

Ontology-driven interoperability
A joint research activity in the frame of the DELOS NoE

Martin Doerr, ICS-FORTH

Rationale

Traditional Libraries provide access to documents via general subjects and metadata about the creator and creation of the document. The search paradigm is restricted to the retrieval of like documents with respect to some search criteria. Although advances have been made in search engine technology, Information Retrieval techniques and standard metadata schemes such as Dublin Core, current methods for integrating material from different domains and terminology systems remain poor, compared to the research demands and the results of manual human investigation.

The challenge for the next generation of information access systems is the ability to retrieve complementary objects and deep paths of relevant relationships that cross multiple document and resource boundaries. Intellectual, logical and physical architectures must be found to identify dynamically relevant resources for complementary information, and to reliably link multiple resources with mechanisms to disambiguate referred items, such as persons, places, objects, periods, but also types and other scientific concepts. Resources of general background knowledge such as gazetteers, VIP lists and domain ontologies should be accessible as automated information services providing necessary bits of information to close gaps in queries and information chains, such as placename-to-coordinate translations.

Complementary information can only be identified against an application and domain overarching core-ontology that allows for relating, mediating or translating the elements of the necessarily heterogeneous data and metadata schemata employed in multiple applications and domains. The promoters of the CIDOC CRM core-ontology (ISO21127) could prove that such ontologies can be created, that they can be kept extraordinarily generic supporting a kind of general discourse (macroscopic, discrete historical or retrospective analysis in this case), and that they can be fairly compact. It appears feasible to organize in the sequence the internal and external logical structure of wide research networks of DL of various disciplines and KOS services in a way that will allow for seamless access to relevant data paths across multiple resources.

Objectives

The activity addresses the key aim of achieving semantic interoperability at both data and metadata levels. Knowledge Organization Systems (KOS), such as classifications, gazetteers and thesauri provide a controlled vocabulary and model the underlying semantic structure of a domain for purposes of retrieval. Ontologies provide a higher level conceptualisation with more formal definition of roles and semantic relationships. The objective of this project is the investigation and development of methods for the integration of heterogeneous data types, models, upper level ontologies and domain specific KOS. This effort will be driven by a domain overarching core ontology starting from the CIDOC CRM (ISO 21127) and will be realised via research reports, guidelines, real world case studies and a demonstrator. Tasks selected for investigation will span the spectrum of applied to general focus. The experimental material will be taken from the particularly rich cultural heritage domain and traditional library science.

Description of Work

In more detail, the work will comprise three units:

A) A collaboration with CIDOC CRM-SIG and IFLA-FRBR Review Group on the creation of a core ontology merging the FRBR and CIDOC CRM concepts. The harmonization of FRBR and CIDOC CRM will be a milestone in the semantic interoperability of ALM (Archives, Libraries and Museums) on international level. It will be carried out by a series of meetings of an interdisciplinary group of experts, and elaboration of the results of the meetings in between. In parallel, the progress of the work is communicated to and approved by the involved interest groups. Criticism and proposals from the interest groups form an integral part of the work. A draft model has already been created.

B) Semantic schema mapping prepares the ground for semantically rich and precise infrastructures to multi-purpose clusters of digital libraries. Whereas most semantic Web activities try to recover knowledge from bad and idiosyncratic structures, here we show systematic ways to produce the necessary diversity of ergonomically optimized data structures without losing the meaning with respect to a common ontology. The scenario of use foresees the core-ontology as an intermediate global schema of medium complexity. Work will demonstrate two directions: the derivation of rich data structures for documentation units from the core ontologies on one side, and the derivation of minimal metadata structures for simplified querying on the other side. Work will contain theory and demonstration. Besides others, a mapping of the CRM to Dublin Core (DC) will allow for automatically creating DC access to all CIDOC CRM compatible resources, a major step to increase semantic richness without losing the simplicity of DC.

C) In virtually all data record we find as data categories of things, frequently called “types”. Depending on the degree of detail a schema captures, some of those categories may be more analyzed than in others. E.g.: “type: clay pot” in one schema may appear in another as “material: clay, form: pot”. The demonstrator will show a solution to the dependency of mappings between different schemata on categorical data (“types”) employed in the data records. It will employ a manual mapping of the upper level of a thesaurus, such as the AAT to the CRM ontology and make use of sample data in the cultural heritage domain encoded in schemata of different levels of detail. This will be supported by a review of the state of the art in facet analysis in LIS and Semantic Web/Grid communities, including a study of the relationships that govern formation of valid (and useful) compound terms.

Project participants:

- Foundation for Research and Technology – HELLAS (FORTH) (leader),
- Norges Teknisk-Naturvitenskapelige Universitet (NTNU),
- DSTC,
- Magyar Tudomanyos Akademia Szamitastechnikai es Automatizalasi Kutatointezet, Department of Distributed Systems (MTA SZTAKI, DSD)
- Imperial College London,
- Ionian University, Archive and Library Science Department (IU),
- National and Capodistrian University of Athens (UOA),
- University of Glasgow,
- Lunds Universität (ULUND),
- Technical University of Crete (TUC),
- IFLA - FRBR Review Group,
- CIDOC CRM Special Interest Group.

References

M. Doerr, K. Schaller, M. Theodoridou, “Integration of complementary archaeological source”, Computer Applications and Quantitative Methods in Archaeology Conference, CAA 2004, 13-17 April, 2004, Prato, Italy

M. Doerr, "The CIDOC CRM - An Ontological Approach to Semantic Interoperability of Metadata”, AI Magazine, Volume 24, Number 3, 2003

M. Doerr, J. Hunter, C. Lagoze, “Towards a Core Ontology for Information Integration”, In Journal of Digital information, Volume 4 Issue 1, April 2003

D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Description Logic Framework for Information Integration"; In Proc. of the 6th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR'98), 1998, pages 2-13

Digital Preservation Cluster

Cluster objectives
Research in the area of digital preservation is fragmented and in need of integration. From the array of possible research tasks, the Preservation Cluster is focusing on those designed to initiate collaborative interaction between institutions and individuals, focus and enable digital preservation, and deliver tangible results by bringing together fragmented research results in different laboratories. The Preservation Cluster has four strategic goals:

- To eliminate the duplication of effort between research activities by creating an integrating framework to co-ordinate and promote research and projects and to enable identification, collection, and sharing of knowledge and expertise
- To examine core issues that will deliver essential guidelines, methods, and tools to enable the construction of preservation functionality within digital library activities and deliverables are created.
- The establishing of testbeds and validation metrics. These will provide a framework for testing preservation strategies, for establishing the preservation worthiness of digital library implementations, and create greater w comparability between research and implementation activities.
- To relate the digital preservation research agenda more directly to the development of exploitable product opportunities and to develop links with the industrial sectors.

The major objectives of the Preservation Cluster are to lay the foundation for testbeds and necessary metrics and tools for assessing preservation strategies, to raise the profile of digital preservation issues within the Digital Library Community, to collaborate with other international bodies to ensure consistencies of digital repository standards, to ensure access to file format information and to establish the relationship between a typology of file formats and preservation strategies, to enable the definition of attributes and functionalities that need to be represented, and ensure that system development methodologies reflect preservation analysis and design issues.

Cluster activities
The Cluster has focused its activities on five major topics.

- Establish a framework for a digital preservation testbed environment and produce metrics for testing and validating digital preservation strategies.
- Contribute to the development of digital repository frameworks and mechanisms for validating the suitability of digital repository implementations. Evaluate the current and emerging systems and storage models for digital repositories.
- Contribute to the development of file format registries and the mechanisms for their use through the definition of relationship between file format types and preservation methods and to assess the viability of producing generic metrics to measure the viability of this preservation approach.
- Define framework for documenting behaviour and functionality. Develop an overview of the attributes of functionality and behaviour that need to be represented and mechanisms for representing them.
- Develop the requirements for a preservation functionality-modeling tool and integrate that into design and development technologies.

Cluster coordinator
Dr Seamus Ross, HATII, University of Glasgow

Delos DPC Testbed:
A Framework for Documenting the Behaviour and Functionality of Digital Objects and Preservation Strategies

Andreas Rauber, Stephan Strodl, Carl Rauch (Technische Universität Wien),
Hans Hofman (Nationaal Archief),
Giuseppe Amato (Consiglio Nazionale delle Ricerche),
Max Kaiser (Österreichische Nationalbibliothek),
Heike Neuroth (Staats- und Universitätsbibliothek Göttingen)

Abstract

Preservation projects have the choice between a wide array of different preservation solutions without knowing which of them best fits their requirements. So instead of leaving preservationists to decisions on a gut level, Cluster 6 on Digital Preservation is developing a framework for evaluating and deciding (the combination of) which preservation solution(s) is optimal in which preservation setting.

Motivation and Testbed Overview

An increasing amount of our cultural and scientific heritage is being produced and maintained digitally, providing enormous benefits in terms of access, requiring less storage space and allowing for easier handling. Yet, all these enormous amounts of information are at risk of being lost due to their dependence on both current hardware and software for rendering and interaction. In order to mitigate this risk, a range of approaches for digital preservation, such as migration or emulation, to name the most prominent ones, are being investigated. For each of these approaches, a range of tools are slowly becoming available to assist in the long term preservation endeavour. Yet, each of these tools has different characteristics, performs differently, preserves different aspects of a digital object while loosing others. It is thus of eminent importance to provide assistance in the process of selecting the optimal preservation strategy for a given setting.

To this end, both a testbed methodology as well as a methodology for evaluating preservation approaches is being developed and integrated to create a digital preservation testbed. This testbed will allow institutions to evaluate preservation strategies by enforcing the precise definition of preservation requirements and supporting the documentation of the processes and experiments. It provides a means to make informed and well-documented decisions, establishing a trusted preservation process. During the first Joint Programme of Activities of the DELOS Network of Excellence the Vienna University of Technology developed tools for comparing and evaluating digital preservation solutions [2], while a testbed for performing preservation experiments was created by the Dutch National Archives. Based on the concepts of the two testbeds, a joint testbed is being developed [1], combining the strengths of both, which shall further be evaluated in several preservation settings by cluster members.

The goal of the current project is to integrate, automate, and evaluate a framework for digital entity preservation by combining the testbed framework and the metrics with a specific focus on evaluation of the resulting framework in a set of real-world case studies at preservation institutions. This requires the development of specific tools to automate selected steps of the preservation process, such as ingest validation, preservation experiment set-up and control, and preservation criteria definition to support semi-automatic alternative evaluation.

Testbed Framework

Figure 1 provides an overview of the workflow within the joint digital preservation cluster testbed. The process consists of 14 steps which are grouped into 3 stages. The first steps are the definition of the project's basic characteristics, i.e. the preservation setting. A set of representative objects is then identified, which shall be used for evaluation purposes. Next, all criteria that could influence the selection of one preservation solution over another, so-called objectives, are collected and structured in an objective tree. Such trees can consist of several hundred criteria, which on the top level are usually grouped in three branches, namely 'File Characteristics', Process Characteristics' and 'Costs'. An excerpt from such a tree is depicted in Figure 2. In the next step measurable units, such as EURO for costs or a subjective ranking for complexity, are assigned to the leaves. Finally, importance factors are assigned to explicitly describe and weight, which criteria have a major or minor impact on the final decision.

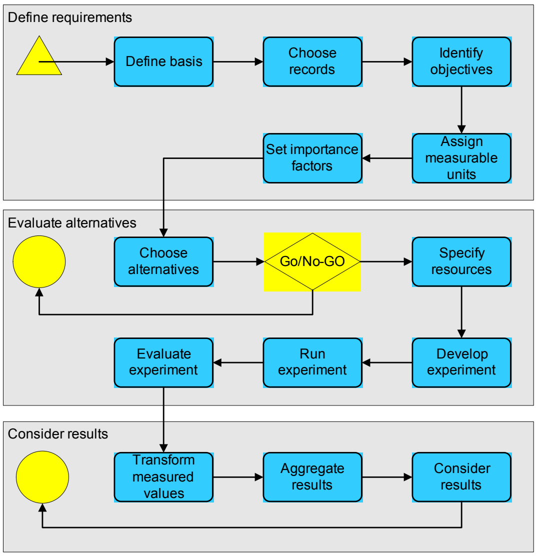


Figure 1 DPC Testbed Workflow

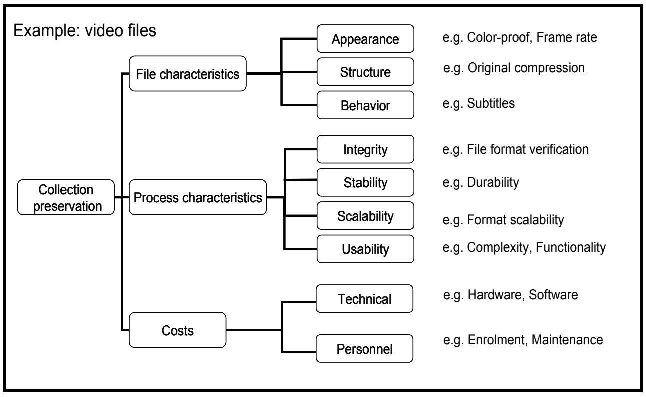


Figure 2 Objective Tree

In the second part of the testbed possible alternative preservation strategies are listed. Alternatives can be from all different preservation strategies, such as specific emulators, or the conversion of digital objects from one specific format to another (version of the same or a different) format, using a specific tool on a given target platform. After that a Go/No-Go-Decision to continue, stop or redefine the process is made, based on the usefulness and cost-effectiveness of the procedure, the required resources and the expected results. If the decision is positive the experiment process is formulated in detail. Afterwards the alternatives are evaluated with respect to the criteria defined in the first part. Here, actual measurements are collected on the objects selected for evaluation.

In the third part the evaluations per alternative are transformed to comparable numbers by using transformations tables, where the translation between the previously measured units (EURO, minutes) into a standardized scale is defined. These comparable values are multiplied with the importance factors and finally aggregated to one final value per alternative. This single value per alternative can be used to rank the alternatives, while the results within the individual branches of the objective tree make their advantages and disadvantages in specific sub-criteria clearly visible, thus assisting in a decision on the combination of different preservation strategies.

A primary goal of the project is to investigate solutions to automate the process of performing preservation experiments. To improve the usability of the existing frameworks and to make them easier applicable, we are developing software support for the evaluation process. A second goal is to deepen our knowledge of the evaluation metric, determining the effect of small changes in the weight of the criteria on the final ranking of the alternatives, measuring, how well an alternative fulfils all criteria in comparison with an optimal preservation solution, and which facets of an alternative to change to receive the highest impact. Finally, a third objective is to evaluate our results in practice on a set of digital object collections. These collections will comprise a specialized document collection of (1) the Austrian National Library and (2) the State and University Library Göttingen, (3) a video collection of the Austrian Phonogrammarchiv, and (4) archival records from the Nationaal Archief.

Bibliography

[1] Hofman, H., Rauber, A. Identifying, Evaluating and Selecting Preservation Methods: An Introduction to the DELOS Testbed and Utility Analysis. Lecture given at the DELOS Summer School on Digital Preservation, June 5-11 2005, Sophia-Antipolis, France.

[2] Rauch, C., Rauber, A. Preserving Digital Media: Towards a Preservation Solution Evaluation Metric. In Proceedings of the International Conference on Asian Digital Libraries (ICADL 2004). Springer LNCS 3334, December 13-17 2004, Shanghai, China.

Digital Preservation Automated Ingest and Appraisal Metadata

Seamus Ross, Yunhyong Kim
Digital Curation Centre (DCC) &Humanities Advanced Technology Information Institute (HATII)
University of Glasgow, {s.ross, y.kim}@hatii.arts.gla.ac.uk

Introduction

Cluster Six of the DELOS NOE focuses on generating synergistic research in the area of digital preservation and curation as applied to digital libraries. Within this large problem space there is a consensus that persistent, cost-contained, manageable, and accessible digital collections depend on the automation of appraisal, selection, and ingest of digital materials [6,12]. ERPANET's Packaged Object Ingest Project (POIP) ([2]) examined the processes of ingesting digital objects (mainly documents) into a digital repository. Even where it proved possible to use auto-extraction tools to acquire the technical metadata (which for many formats was not possible because the necessary tools were not available) the ingest process remained labour intensive. The effort required to extract the necessary descriptive, structural, and semantic information proved to be substantial, as it had to be manually derived. The qualities of digital materials that repositories need to handle makes automation of processes essential. Our research investigates the viability of automating several preservation processes as a mechanism for improving them and demonstrating the viability of automation more generally in the area of digital curation. Specifically, it focuses on automating the extraction of semantically encoded metadata from digital documents. In an effort to make coherent progress we have focused in the first stage on addressing the following problems:

- Summarizing the existing tools that can be integrated to provide the underlying mechanisms for supporting ingest and automated metadata extraction (both technical and semantic);
- Constructing an experimental corpus of documents on which to test our metadata extraction work and to enable other researchers to benchmark similar research against by using our tightly defined and structured corpus;
- Experimenting with metadata extraction from a single document type, in this instance pdf;
- The prototyping of an integrated tool for semantic, structural, and technical metadata extraction for documents represented as pdfs; and,
- The automated population of metadata and document repositories.

Automatic semantic metadata extraction

On the level of automatic semantic metadata extraction, the project aims to integrate suitable existing and implement new information extraction and language processing techniques. Three core activity areas are being pursued concurrently:

- *Establishing a standardized experimental document corpus*. The research will begin with text files in PDF. We are building a corpus of one million pdf documents distributed representatively over five European languages and collected to reflect the different genres and files sizes.
- *What semantic metadata will be extracted*. In the area of non-technical metadata an enriched form of Dublin Core to fit into the framework of Functional Requirement Bibliographic Records (FRBR) [3] has been adopted. However, in the first instance, the work involves extracting: title, author, subject, keywords, genre classification, date, language, identifier (i.e. DOI or persistent handle if present), and a content summary. The extraction of title, author, date, keywords and possibly identifier is a data extraction problem, that is, a matter of finding the appropriate string(s) within the document. Extracting subject, genre classification, language and deriving a content summary lie in the domain of information extraction/creation.
- *Looking at and integrating previous and current related work*. Metadata extraction of partial bibliographic information has been attempted using spatial information of strings (e.g. [4]), or using lexical and capitalisation information as features for a support vector machine (e.g. [5]). Other related work comes scattered across the domain of Natural Language Processing in the form of Text Categorization dealing with subject extraction (e.g. [13]), genre classification using verb and noun counts and length features (as described in [7]), or using lexicon, punctuation and length analysis features ([8]), not to mention document clustering methods [9]. There are also helpful studies which endeavour to automatically detect subjective sentences to aid summarisation, keyword searches and location of key theories([10]).

The Research approach

As a general guideline, the classifier should maximize the independence of modules within the model, ensure that the level of precision exceeds that of an information retrieval task on measuring relevance (reliability in a digital library context is crucial) and include scope for improvement and expansion. The classifier needs to differentiate language dependent and independent features which may also lead to differentiation of lexical, syntactic, and semantic features. We are first building a model for one language by performing genre classification (e.g. distinguishing whether a document is a work of fiction or a research paper), which will narrow down the structural form of the document to facilitate data extraction (e.g. retrieving author, title, date and keywords). By incorporating the extracted metadata in the subsequent analysis stages the performance on subject identification and content summarization tasks can be improved.

Although statistical methods will provide a core component of the tool, a purely statistical approach is bound to limit the performance and may necessitate significant changes at a later point. Accordingly, longer term perspectives on the research will try to look at general frameworks which could serve as a template for handling information extraction problems such as reproducing kernel Hilbert space model for error function space ([1], [14]), or quantum mechanics model for information retrieval ([11]). In the long term, the goal is to discover textual harmonics akin to those for speech to obtain an objective or fairly objective representation of the text from which information or data can be easily extracted.

Conclusion and integration

The long term accessibility and usability of the digital manifestations of our cultural and scientific heritage depends in part on its storage in suitable digital repositories. These repositories require that the information they hold is effectively documented, classified, and summarized if it is to be efficiently managed and users are to be able to pull content from the repository with minimal effort. In the former instance, repositories as they fulfill their preservation and curation responsibilities may well wish regularly to re-appraise their digital holdings to determine whether they should continue holding certain materials. If this is to be done efficiently it must be automated. Users of repositories wish to retrieve ‘the documents relevant to their search’ with minimal effort and with high precision of recall. Richer descriptions and more accurate classifications will help.

While our initial work focuses on only one file type and we would argue that it is a file type that can be tackled using existing technologies. We envisage extending this research to cover other file types (e.g. spreadsheets) and integrating it with tools for extracting technical metadata about representation information. The real success of the work will come from the longer-term development of a framework for designing prototype tools for assisting with appraisal. But definition of a generic rule set for supporting ‘factory-like’ production of tools for enabling semantic metadata extraction from other documentary file formats might be one next step. As we progress this research during 2005 other partners in Cluster Six, and specifically colleagues at the Universities of Urbino and Cologne and UKOLN, will undertake validation and testing of the approaches.

References

[1] Cucker F and Smale S, 2001, ‘On the Mathematical Foundations of Learning’, *Bull. Am. Math. Soc.*, Vol. 39, No 1, pp1-49.

[2] ERPANET’s Packaged Object Ingest Project, http://www.erpanet.org/poip.

[3] International Federations of Library Associations and Institutions (IFLA), 1998, *Functional Requirements for Bibliographic Records*, UBCIM publications, vol 19, www.ifla.org/VII/s13/frbr/frbr.pdf

[4] Giuffrida G, Shek E and Yang J, 2000, “Knowledge-based Metadata Extraction from PostScript File”, *Proc. 5th ACM Intl. conf. Digital Libraries*, 77-84.

[5] Han H, Giles L, Manavoglu E, Zha H, Zhang Z and Fox E A, 2003, ‘Automatic Document Metadata Extraction using Support Vector Machines", *Proc. 3rd ACM/IEEE-CS conf. Digital libraries*, 37-48.

[6] Hedstrom M, Ross S, Ashley K, Christensen-Dalsgaard B, Duff W, Gladney H, Huc C, Kenney A R, Moore R, and Neuhold E, 2003, *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and* Preservation, (http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf Report of the European Union DELOS and US National Science Foundation Workgroup on Digital Preservation and Archiving.

[7] Karlgren J and Cutting D, 1994,‘Recognizing Text Genres with Simple Metric using Discriminant Analysis’, *Proc. 15th conf. Comp. Ling.,* Vol 2, 1071-1075.

[8] Kessler B, Nunberg G and Schuetze H, 1997, ‘Automatic Detection of Text Genre’, *Proc. 35th ann. meeting ACL,* 32-38.

[9] Rauber A and Mueller-Kroegler A, 2001, ‘Integrating Automatic Genre Analysis into Digital Libraries’, *Proc. 1st ACM/IEEE-CS conf. on Digital Libraries,* 1-10.

[10] Riloff, E, Wiebe, J, and Wilson, T, 2003, ‘Learning Subjective Nouns Using Extraction Pattern Bootstrapping’, *Proc. 7th CoNLL,* 25-32, Edmonton, CA.

[11] van Rijsbergen K, 2004, *The Geometry of Information Retrieval*, Cambridge University Press.
[12] Ross S and Hedstrom M, 2005, 'Preservation Research and Sustainable Digital Libraries', *International Journal of Digital Libraries* (Springer), DOI: 10.1007/s00799-004-0099-3.
[13] Sebastiani F, 2005, 'Text categorization', *Text Mining and its Applications*, WIT Press, 109-129.
[14] Smale S. and Zhou D., 2005, 'Learning Theory Estimates via Integral Operators and Their Applications', http://www.tti-c.org/smale_papers/samplll5412.pdf.

Evaluation

Cluster objectives

Digital libraries need to be evaluated as systems and as services to determine how useful, usable, and economical they are and whether they achieve reasonable cost-benefit ratios. Results of evaluation studies can provide strategic guidance for the design and deployment of future systems, can assist in determining whether digital libraries address the appropriate social, cultural, and economic problems, and whether they are as maintainable as possible. Consistent evaluation methods also will enable comparison between systems and services.

The evaluation cluster is working both on evaluation methodologies in general as well as on providing the infrastructure for specific evaluations. Thus, the following objectives are addressed:

- *Development of a comprehensive theoretical framework for DL evaluation,* which can serve as reference point for evaluation studies in the DL area.
- *Research on new methodologies,* in order to overcome the lack of appropriate evaluation approaches and methods.
- *Development of toolkits and test-beds* in order to enable new evaluations and to ease the application of standard evaluation methods.

Cluster activities

In order to reach these goals, the following activities are being carried out:

- *Workshops on DL evaluation,* for collecting existing evaluation approaches and methods.
- *Evaluation support to the DL community,* by creating an evaluation forum for enabling communication between evaluation specialists and DL developers.
- Development of new approaches and methods, in order to overcome the weaknesses of current approaches and the lack of methods for new types of applications.
- Development of evaluation toolkits, e.g. for collecting and analyzing experimental data.
- Creation of test-beds for new content and usage types in DLs, by starting from the existing test-beds for XML and cross-lingual retrieval and extending these towards new media, applications and usage types.
- Creation of test-beds for usage-oriented evaluation, by extending existing test-beds or by creation of test-beds of user interactions

Cluster coordinator

Norbert Fuhr, Universität Duisburg-Essen, Germany

The INEX Initiative for the Evaluation of XML Document Access and Retrieval

Mounia Lalmas, Anastasios Tombros {mounia, tassos}@dcs.qmul.ac.uk
Department of Computer Science, Queen Mary University of London

Keywords
Algorithms; Experimentation; Measurement

Initiative for the Evaluation of XML Retrieval
Digital libraries (DL) need to be evaluated to determine how useful, usable, and economical they are, and whether they achieve reasonable cost-benefit ratios. Results of evaluation studies can provide strategic guidance for the design and deployment of future systems, can assist in determining whether digital libraries address the appropriate social, cultural, and economic problems, and whether they are as maintainable as possible.
One task of the DELOS research cluster on evaluation (please see http://dlib.ionio.gr/wp7/ for further details) is the evaluation of content-oriented access to XML documents [1], where XML stands for “extensible Markup Language”. XML is increasingly being used in digital libraries and similar systems or platforms (e.g. XML is becoming the W3C standard for representing documents). The provision of effective access to XML-based content has become a key research issue, and is the focal point of XML retrieval research. XML retrieval systems aim to exploit the logical structure of documents to retrieve document components, the so-called XML elements, instead of whole documents in response to a user's query. According to this retrieval paradigm, an XML retrieval system needs not only to find relevant information in the XML documents, but also to determine the appropriate level of component granularity to return to the user. Evaluating how good these systems are, hence, requires test-beds where the evaluation paradigms are provided according to criteria that take into account the imposed structural aspects.

In 2002, the Initiative for the Evaluation of XML Retrieval (INEX, please see http://inex.is.informatik.uni-duisburg.de/ for further details) started to address these issues. INEX has a strong international character; participants from over 50 organisations, distributed across Europe, North America, Australia, New Zealand and Asia, have so far registered to participate in this year’s fourth INEX run. The aim of the INEX initiative is to establish an infrastructure and to provide means, in the form of a large XML test collection and appropriate scoring methods, for the evaluation of content-oriented XML retrieval systems.
Evaluating retrieval effectiveness is typically done by using test collections assembled specifically for evaluating particular retrieval tasks. A test collection usually consists of a set of documents, a set of user requests (the so-called topics, or queries) and relevance assessments of the documents with respect to the queries. The characteristics of traditional test collections have been adjusted in order to appropriately evaluate content-oriented XML retrieval effectiveness: the document collection comprises documents marked up in XML, the topics specify requests relating both to the content of the desired XML elements and to their structural properties, and the relevance assessments are made on the XML element level rather than just on the full document level. In addition, relevance is measured in a different way compared to traditional Information Retrieval (IR) research, in order to more effectively quantify the systems’ ability to return the right granularity of XML elements. To date, the INEX test collection has approximately 300 topics in total. More details about the INEX test collection can be found in [2].

Main INEX Activities
Ad-hoc task. The retrieval task to be performed in INEX was defined as the ad-hoc retrieval of XML documents. In IR literature, ad-hoc retrieval is described as a simulation of how a library might be used, and it involves the searching of a static set of documents using a new set of topics. While the principle is the same, the difference for INEX is that the library consists of XML documents, the queries may contain both content and structural conditions and, in response to a query, arbitrary XML elements may be retrieved from the library. Within the main ad-hoc retrieval task, three sub-tasks have been identified depending on how structural constraints are expressed in queries.

In the Content-Only (CO) sub-task, queries ignore the document structure and contain only content-related conditions. Depending on how we assume that a user would want the output of an XML retrieval system to be, three different strategies are defined. In a focussed strategy, we assume that a user prefers a single element that most exhaustively discusses the topic of the query (most exhaustive element), while at the same time it is most specific only to that topic (most specific element). In a thorough strategy, we assume that a user prefers all highly exhaustive and specific elements, and in a fetch and browse strategy we assume that a user is interested in highly exhaustive and specific elements that are contained only within highly relevant articles. An extension of

the CO sub-task that includes structural hints is the +S subtask, where a user may decide to add structural hints to his query to narrow down the number of returned documents resulting from a CO query.
In the Content and Structure (CAS) sub-task, structural constraints are explicitly stated in the query and they can refer both to where to look for the relevant elements (i.e. support elements), and what type of elements to return (i.e. target elements). A structural constraint can also be interpreted as strict (i.e. the structural requirements must be followed strictly) or vague (i.e. the structural constraints are interpreted as hints and the main goal is to satisfy the overall information need). Strict and vague interpretations can be applied to both support and target elements, giving a total of four strategies for the CAS subtask.

Additional tasks. In addition to the evaluation of retrieval effectiveness for the ad-hoc task, further research issues, with direct applications to digital libraries, are being explored in INEX in the form of research tracks. The current run of INEX, which started in April 2005, includes a number of additional tracks.

The interactive track aims to investigate the behaviour of users when interacting with components of XML documents, and to investigate and develop effective user-based approaches for XML retrieval. The track is currently in its second year. In INEX 2004 the behavior of users when interacting with XML elements was investigated. One of the major outcomes was the need to look into methods that can be supportive during the search process based on features extracted from the XML formatting. Problems that might be solved using such methods include that of overlapping elements, i.e., elements from the same document at different ranks in the result list. Such overlaps proved frustrating for users. Another issue to be investigated is how to present XML elements in the result list and when browsing within the structure of individual documents, so that users can get a good indication of whether the element might be worth examining further or not.

The heterogeneous track looks at the case at which an XML document collection comprises documents from different sources, based on multiple document structures (i.e. multiple Document Type Definitions, DTDs). The current INEX collection is based on a single DTD, and as part of the track further document sources were added to the original collection. One of the issues that the track looked into in its first year in 2004, was how to map structural constraints in queries across collections with different DTDs. It was discovered that there does not exist exact information about the, possibly, equivalent tags from different collections, and as a consequence, only a subset of such collections may be used while indexing and retrieving. In its second year, the primary focus of the track will still be on the construction of an appropriate test collection and of appropriate tools for the evaluation of heterogeneous retrieval, and on the investigation of effective ways to retrieve from heterogeneous collections of XML documents.

Two new tracks to be included in INEX 2005 are the multimedia and document mining tracks. The main objective of the multimedia track is to provide an evaluation platform for structured document retrieval systems that include other types of media, apart from text, such as images, speech, and video, in the retrieval process. An additional objective, is to open a discussion forum where the participating groups can exchange their ideas on different aspects of the multimedia XML retrieval task, so as to promote and support research in this area. The document mining track, which is a joint effort between INEX and PASCAL (please see http://www.pascal-network.org/ for further details), aims to look at defining challenges and techniques for effectively mining information from XML documents. The track will have a specific focus on classification and clustering techniques for XML documents in its first year.

Conclusions

With the inclusion of the various research tracks, INEX is expanding in scope and in the number of participating organisations. INEX is also in the process of acquiring new collections of XML documents in an effort to enhance the evaluation environment. INEX has shown that XML retrieval is a challenging field within IR and DL research. In addition to learning more about XML retrieval approaches, INEX is making steps in the evaluation methodology for accessing and retrieving XML documents.

References

[1] Agosti, M. and Fuhr, N. (eds). Working Notes of the DELOS Workshop on the Evaluation of Digital Libraries. Padova, Italy, 2004.
[2] Fuhr, N., Lalmas, M., Malik, S. and Szlavik, Z. (eds). Advances in XML Information Retrieval, Third International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2004). Lecture Notes in Computer Science, volume 3493, 2005.

Multilingual Information Access for Digital Libraries CLEF – a System Testing and Evaluation Framework

Carol Peters, carol.peters@isti.cnr.it
Istituto di Scienza e Tecnologie dell'Informazione (ISTI-CNR), Pisa, Italy

Categories, Subject descriptors and Keywords
H.3.3 [Information Storage and Retrieval]: Search process; H.3.4 [Systems and Software]: Performance evaluation, Multilingual information access, System evaluation, Test collections.

Project description

Multilinguality is a key issue for European digital libraries although frequently neglected in practice. For this reason, the DELOS Network of Excellence for Digital Libraries promotes research into cross-language information retrieval by supporting the activities of the Cross-Language Evaluation Forum (CLEF). The objective of CLEF is to stimulate research into the development of systems for all kinds of multilingual information access through:

- the organization of annual system evaluation campaigns
- the creation of a discussion forum and the holding of annual workshops for the comparison of approaches, ideas and results
- the construction of reusable test-collections for system benchmarking.

The aim is to encourage the development of fully multilingual and multimedia user-friendly systems.
When the first cross-language information retrieval (CLIR) system evaluation activity began in 1997 at the US Text REtrieval Conferences (TREC), very little IR system testing work had been done for languages other than English and almost all existing cross-language systems were designed to handle no more than two languages. Since its beginnings, CLEF[1] has worked hard to change this situation and to promote the development of systems capable of searching over multiple languages. For this reason, each year, the campaign has proposed a set of core evaluation tracks designed to test monolingual, bilingual and multilingual text retrieval systems. The aim has been to encourage groups to work their way up gradually from mono- to multilingual text retrieval, providing them with facilities to test and compare search and access techniques over languages and pushing them to investigate the issues involved in processing a number of languages with different characteristics. Over the years, the language combinations provided have increased and the tasks offered have grown in complexity. In 2000, the main CLEF multilingual corpus consisted of approximately 360,000 newspaper and news agency documents for four languages; by 2005 it has grown to well over 2 million documents and twelve languages. The CLEF 2003 and 2005 multilingual tracks have included tasks which entail searching a collection in eight languages and returning the results in a single ranked list.

The evaluation environment for these tracks adopted an automatic scoring method, based on the well-known Cranfield methodology (Cleverdon, 1997), and has adapted it to the multilingual context. The test collections consist of sets of "topics" describing information needs and sets of documents to be searched to find relevant information. All language dependent tasks such as topic creation and relevance assessment are performed in a distributed setting by native speakers. Evaluation is done for each ranking of documents with respect to a topic by the usual computation of recall and precision. In the first years, CLEF focussed mainly on testing overall performance of off-line text retrieval systems, where good system performance is equated with good retrieval effectiveness. The results in terms of participation and of the different approaches and techniques tested have been impressive. Much work has been done on fine-tuning for individual languages, while other efforts have concentrated on developing language-independent strategies. The issues involved in cross-language text retrieval have thus been investigated in depth and improvement in system performance over time has been demonstrated. A discussion of these results and the lessons learned can be found in Braschler & Peters (2004).

However, this is only one part of the CLIR problem. A multilingual system evaluation activity that meets the needs of the potential application communities must also provide facilities to investigate many other issues:

- not just document retrieval, but also targeted information location and extraction;

[1] CLEF was launched when it was decided to move the coordination of the existing CLIR track at TREC to Europe. CLEF 2000 and 2001 were sponsored by the 5FP DELOS Network of Excellence; CLEF 2002 and 2003 were funded independently by the European Commission under the IST programme. From 2004 CLEF is again organised as an activity of DELOS under the 6FP; the DELOS groups involved in the coordination of CLEF are ISTI-CNR, Italy; Language and Interaction Lab., SICS, Swedish Institute for Computer Science; Dept. of Information Engineering, University of Padua, Italy.

- not just text but also multimedia data, e.g. collections containing images or spoken documents;
- not just system performance but also wider usability issues that affect the users' ability to recognize relevant information and refine search results even if documents are written in an unfamiliar language.

For this reason, over the years we have gradually increased the evaluation tasks offered to the participants in the annual campaigns, in order to stimulate the development of CLIR systems that include such functionality. In 2001 we introduced a track to investigate document selection questions and to support mechanisms for interactive query formulation and refinement (iCLEF). Experiments in this track over the years have included user-assisted term translation and user-assisted query reformulation. In 2002 CLEF began to pay attention to the issues involved in cross-language spoken document retrieval (CL-SDR – now CL-SR); in 2003, and again in 2004, the CL-SDR track aimed at evaluating CLIR systems on noisy automatic transcripts of spoken documents with known story boundaries. The results of the experiments showed that, as expected, bilingual performance was lower for all participants than the comparative English monolingual run. The degree of degraded performance was shown to depend on the translation resources used. In 2003, we offered two pilot experiments for multiple language question answering (QA@CLEF) and cross-language retrieval on image collections (ImageCLEF). Both tracks attracted a lot of attention and have now been confirmed as regular CLEF tracks. The QA track is seen as very important not only because it has generated the first cross-language QA test suite but also because it has stimulated monolingual QA system development for languages other than English – almost all previous studies in this area were restricted to English. The ImageCLEF track is very popular and has encouraged much testing of systems that combine both visual and text data in their search. In 2005, two new tracks have been introduced: a multilingual web track (WebCLEF) and a track for cross-language geographical information retrieval (GeoCLEF).

CLEF 2005 has thus offered eight tracks designed to evaluate the performance of systems for:

- mono-, bi- and multilingual document retrieval on news collections (Ad hoc)
- mono- and cross-language domain-specific retrieval (Domain-specific)
- interactive cross-language retrieval (iCLEF)
- multiple language question answering (QA@CLEF)
- cross-language retrieval on image collections (ImageCLEF)
- cross-language spoken document retrieval (CL-SR)
- multilingual web track (WebCLEF)
- cross-language geographic information retrieval (GeoCLEF)

About eighty groups from 27 different countries, distributed over 5 continents, have registered to participate in one or more of these tracks. Eight different document collections have been used in CLEF 2005 to build the test collections for the different tasks:

- CLEF multilingual comparable corpus of more than 2 million news documents in 12 languages[2]
- The GIRT-4 social science database in English and German and the Russian Social Science Corpus
- St Andrews historical photographic archive
- CasImage radiological medical database with case notes in French and English
- IRMA collection in English and German for automatic medical image annotation
- Malach collection of spontaneous conversational speech derived from the Shoah archives
- EuroGOV, a multilingual collection of about 2 million webpages crawled from European governmental sites

It is generally agreed that evaluation campaigns play an active role in advancing system development. The goal of CLEF at the moment is to attempt to narrow the gap between the R&D community and application world. We have begun to do this by including activities that investigate different kinds of user-system interaction and that test system performance on collections that are not just text-oriented but consider the needs of other media. We intend to continue in this direction. To find out more about CLEF past and future activities and for an extensive bibliography, see: http://www.clef-campaign.org.

[2] The CLEF comparable corpus currently contains news documents for the same time period (1994-95) in ten languages: Dutch, English, Finnish, French, German, Italian, Portuguese, Russian, Spanish, and Swedish, and for 2002 in Bulgarian and Hungarian.

Project participants

University of Padua: Maristella Agosti, Giorgio Di Nunzio, Nicola Ferro
Swedish Institute of Computer Science – SICS: Preben Hansen, Jussi Karlgren

References

Braschler, M. & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements, Information Retrieval, 7(1-2) pp 7-31.

Cleverdon, C. (1977). The Cranfield Tests on Index Language Devices. In: K. Spärck-Jones and P. Willett, eds. Readings in Information Retrieval, Morgan Kaufmann, 1997. pp 47-59.

A Digital Library Testbed Framework for the Evaluation of Architectures, Services and Execution Dynamics

Norbert Fuhr, Claus-Peter Klas – University of Duisburg-Essen
Hanne Albrechtsen – RISOE
Sarantos Kapidakis – Ionian University
Andras Micsik – MTA SZTAKI, Hungary

Keywords
H.3.7 [Digital Libraries]: Dissemination – System Issues – User Issues - Evaluation.

Objectives
Today most digital library (DL) evaluations use specific systems, which are difficult to compare. The aim of this effort is to provide a standard testbed framework for comparative evaluation of DL systems. Based on a theoretical framework for DL evaluation, we will develop a framework system to *guide* the stakeholders of an DL evaluation and to *provide* a communication and sharing platform. Thus, stakeholders can work more efficiently by using the DELOS evaluation framework for their scientific research.

Standard Evaluation Framework
Theoretically, our work is based on new DL evaluation frameworks like the Evaluation Computer [Kovacs & Micsik 04] and the DL interaction model [Tsakonas, Kapidakis, Papatheodorou 04], in combination with well-tested frameworks like the Cognitive Systems Engineering framework [Albrechtsen, Andersen, Cleal & Pejtersen, 2004; Rasmussen, Pejtersen & Goodstein, 1994]. On the practical side, we use the Daffodil-system [Klas et al. 04], which is a frontend system consisting of a rich collection of services for accessing federations of DLs. The current prototype application (http://www.daffodil.de) integrates about 15 DLs. Within this task, we want to generalize Daffodil to a flexible toolbox for DL evaluation, to provide easy ways for

- applying Daffodil to other digital object collections,
- integrating new DL services into Daffodil (both local and remote services),
- replacing/disabling services or user interface components of Daffodil.

The flexible architecture of Daffodil is a good base for implementing such a testbed. Based on the theoretical models mentioned above, we will investigate the area of evaluations for which Daffodil could be used, and then we will enhance the system appropriately, such that it is a generic DL testbed e.g. for testing new algorithms, service types, user interface paradigms, storage mechanisms, etc..

Standard Logging Framework
For analyzing and evaluating the application of this testbed framework, as well as for comparison with and between other systems, we will develop a logging framework for events in DL usage. In connection with the basic reference model of digital libraries, events of the DL can be defined on different levels of the architecture and on different levels of abstractions. Currently, we have system events and multiple levels of GUI interaction events. We assume that all other levels needed for evaluations possible can be derived through transformations.

We present a first preliminary XML schema for these log events, which is similar to the proposed format in [Fox et al 02], but more comprehensive, due to the richer functionality of Daffodil. For each event, a minimum set of attributes is described in order to cover a large portion of events occurring in the Digital Library. The set of attributes are under discussion and are defined according to the evaluation purposes. In addition, this set must be defined such that it can be extended easily, in order to allow for richer descriptions of events.

Current activities
The first activities are to develop an evaluation framework based on the current state of the art in digitial library evaluation. Second, a first draft of the digital library event logging schema is developed, implemented and preliminary evaluated within the Daffodil system. The existing analysing tools are reprogrammed to meet the new standard. Third, the Daffodil application, currently restricted to the area of computer science, will also be applied to the area of archaeology. Finally, new services and visualizations will be integrated and evaluated.

Expected Results

- *Community Building in Evaluation Research.* The lack of globally accepted abstract models and methodologies in this evaluation discipline can be counterbalanced by collecting, publishing and analyzing current research activities. Overview of evaluation activities and their interrelations may help to define good practice in the field and can help to build the community of researchers as well.

- *Establishment of Primary Data Repositories for Data Mining.* The open access to primary data of evaluation (transaction logs, surveys, monitored events etc.) as a tradition could be borrowed from other research fields. In this aspect, proper anonymization of evaluation primary data is a problem to solve, as privacy can be a basic concern here. Common repositories and infrastructure for storing primary and secondary data are suggested along with the collaborative formation of evaluation best practices, and modular building blocks for evaluations.
- *Standardized Logging Format.* Further use and dissemination of common logging standards is also considered useful. Logging could be extended to include user behavior and system internal activity as well in order to support the personalization and intelligent user interface design processes.
- *Evaluation Specifics of Digital Libraries.* How does a digital library relate to other complex networked information systems (e.g. archives, portals, knowledge bases, etc.) with respect to evaluation? Is it possible to connect or integrate DL evaluation to the evaluation of various web-based information services?
- *Verification of DL execution versus a specification of DL dynamics.* In this scenario a specification of expected event sequences is given and one can check if the actual execution corresponds to the specification. This verification can be based on a description technique developed for specification of expected DL event sequences.
- *Quantitative analysis of DL logs.* For the DL log format specified, we will develop a set of tools for statistical analysis of DL usage logs. These tools will allow computing e.g. frequencies of certain operations or combinations thereof, as well as analyzing corresponding execution times. For more detailed analysis, we will provide a tool for converting selected attribute combinations into the standard input format for data mining tools, the ARFF format.
-

Project participants

UNI DUISBURG - Norbert Fuhr fuhr@uni-duisburg.de, Claus-Peter Klas klas@uni-duisburg.de
RISOE - Hanne Albrechtsen hanne.albrechtsen@risoe.dk
MTA SZTAKI DSD: Laszlo Kovacs laszlo.kovacs@sztaki.hu, Andras Micsik micsik@dsd.sztaki.hu
Ionian University: Sarantos Kapidakis sarantos@ionio.gr

References

1. Klas, C.P., Fuhr, N., Schaefer, A.: Evaluating strategic support for information access in the DAFFODIL system. In Heery, R., Lyon, L., eds.: Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2004). Lecture Notes in Computer Science, Heidelberg et al., Springer (2004)
2. Schaefer, A., Jordan, M., Klas, C.P., Fuhr, N.: Active support for query formulation in virtual digital libraries: A case study with DAFFODIL. To appear in: Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2005).
3. Kovács, L., Micsik, A.: The evaluation computer: a model for structuring evaluation activities. DELOS Workshop on the Evaluation of Digital Libraries, Padua, Italy, October 4-5, 2004
4. Rasmussen, J.; Pejtersen, A.M. and L.P. Goodstein (1994). *Cognitive Systems Engineering*. New York: Whiley.
5. Albrechtsen, H.; Andersen, H.H.K.; Cleal, B.R.; Pejtersen, A.M (2004). Categorical complexity in knowledge integration: Empirical evaluation of a cross-cultural film research collaboratory. In: *Knowledge organization and the global information society. Proceedings. 8. International ISKO conference, London (GB), 13-16 Jul*
6. Tsakonas, G., Kapidakis, S., Papatheodorou, C. (2004). Evaluation of user interaction in digital libraries. DELOS Workshop WP7, Padua

Printed by ISTI-CNR
at Gruppo ALI, Pisa, Italy
July 2005