Dagstuhl Seminar 99371:

Declarative Data Access on the Web

N. Spyratos, Paris, France
spyratos@lri.fr

K. Vidyasankar, Newfoundland, Canada
vidya@cs.mun.ca

G. Vossen, Münster, Germany
vossen@helios.uni-muenster.de

September 12–17, 1999

Contents

1 Introduction

Today, information is spreading to all sectors of society in ever increasing volumes. This information comes in multimedia digital form and is transmitted over world-wide networks. In particular, the World-Wide Web (WWW) renders it possible to obtain information that is distributed over the entire Internet. Since the Web (and the number of its users) continues to grow at a high speed, adequate tools are needed for finding, storing, and structuring the vast amount of information offered; for *locating, retrieving*, and *presenting* the information to the final user; for aiding the end-user in customizing the information obtained for personal usage. Although technology is advancing fast (e.g., Web browsers built into cellular phones), a lot remains to be done concerning the efficient retrieval of information from large digital collections (often called digital libraries), and its intelligible presentation to the end user.

From a database perspective, the information provided by the Web can be perceived as a huge, heterogeneous database which is distributed world-wide, and which is accessed by multiple users. From this point of view, it appears reasonable to try to adopt concepts and techniques from database technology and in particular from the area of *information retrieval* (IR) to the context of the Web. It makes sense to investigate to what extent they are applicable to a large-scale database such as the Web, or what kind of generalizations, extensions, or completely novel developments become necessary. Motivation to do so is obtained from a look at the present situation. Indeed, when accessing data sources on the Web with current browser and search-engine technology, a number of issues arise which deserve further study; these include:

1. *Locating the source*: Today's search engines are "primitive" devices for performing searches simply because they rarely do content-based retrieval; this in particular applies to search engines such as `AltaVista`, `Fireball`, `Lycos`, `Yahoo!`, and others. Instead, they mostly rely on searching indexes or directories, where distinct strategies in handling index information are exploited.

   Generally, users of the Web feel the need to develop advanced tools for locating information, for example based on graph navigation techniques, on content-based retrieval as known in IR, or on classification techniques used in present-day data mining. Moreover, a specification of desired results in a declarative way, preferably in an SQL-related language as done in various research prototypes, could aid in locating sources that are appropriate for a given query.

   When retrieving information from a "digital library" such as the Web, the *precision problem* and the *recall problem* arise. Due to the uncertainty associated to information retrieval, the answer to a query usually contains non-relevant data (this is the precision problem), while relevant data may be ranked low or neglected altogether (this is the recall problem).

   In other words, an automated device may retrieve imprecise data, or may not be certain whether or not to include some data found in the answer set. Once again, techniques based on content-oriented retrieval could help.

2. *Retrieving relevant information*: Once a desired data source has been located, the issue of retrieving relevant data arises. Two factors are important in this respect:

the *structure* of the data, i.e., data can be anywhere between highly structured and totally unstructured, and the multimedia *nature* of data.

Both factors are vastly orthogonal, and require different forms of exploration, treatment, and presentation going way beyond what today's browsers like `Netscape`, `Internet Explorer`, or `HotJava` have to offer. Again, information retrieval (and possibly data mining) techniques could help, for example for searching a video database found at some Web server for all scenes in which Humphrey Bogart kisses Lauren Becall. For specification of such search goals, we still envision an extension of SQL that allows to give possibly incomplete path and class descriptions.

Another important problem is providing *multi-modal retrieval*, i.e., allowing a single request to contain retrieval conditions expressed in different modes (visual, text, etc), and permit the user to select the most appropriate mode.

3. *Organizing retrieved data for personal usage*: Once data has been retrieved, it most likely needs to be reorganized for easy use in the applications the end-user has in mind. Often, data obtained from the Web is kept in a customized database for personal usage.

   The organization and maintenance of personal databases follows naturally as a topic from the other two. However, to the end we can imagine PC or workstation tools that help taking care of this so that these issues do not need fundamental research.

Even in the restricted and simple case that a user accesses a *single* data source through the Web, all these issues described above may arise. When multiple sources are accessed, the additional problem of *combining* and *integrating* information from these different sites comes up. As the information available through the Web becomes more and more complex and voluminous, the importance of providing adequate, application-oriented interfaces becomes a decisive factor. Indeed, users of cellular phones, office computers, or GPS-based navigation systems in cars, to name just a few, have vastly different requirements to their Web interfaces, ranging from simple textual to multimedia output. Considering the various environments from which information can be accessed (including mobile ones), the need to adapt interfaces to these environments arises. For example, an SQL query can hardly be expressed on the interface of a pager that provides one line of text only; on the other hand, a multimedia office computer can easily accommodate more sophisticated query facilities than SQL.

The goal of this seminar has been to study the problems of locating, retrieving, and presenting information on the Web. To this end, the seminar brought together researchers from the areas of database systems and data warehouses, information retrieval, multimedia presentations and interface design, as well as information integration, in order to discuss demanding questions and open problems in detail; the issues discussed specifically included:

- query languages for locating and retrieving Web data;
- appropriate user interfaces;

- techniques for query processing;
- information retrieval approaches;
- resource management and information organization;
- web-farming and data mining, for example as applied in areas such as electronic commerce;
- techniques for analyzing the information found in Web servers;
- integration of information found on the Web for the purpose of creating personalized databases;
- emerging Web standards, in particular XML and its proposals for query languages.

The meeting brought together 22 scientists from 6 different countries. During the week, 19 presentations were given with plenty of discussion time during and after each which even gave rise to another workshop proposal (to be submitted shortly); in addition, ample time were reserved for discussions of special topics in three small working groups. The remainder of this report contains the final program (Section 2), abstracts of the presentations (Section 3), working group summaries (Section 4), and the complete list of participants (Section 5).

We felt that all participants enjoyed the workshop and the pleasant and constructive atmosphere of the Dagstuhl Castle, which this time was coupled with splendid weather throughout the whole week. We particularly wish to thank the Dagstuhl staff for ensuring that everything ran so smoothly.

2 Final Program

The program consisted of talks and several working groups which reported their findings on the last day. In addition, there was the usual Wednesday afternoon excursion (this time the trip to the city of Trier). In detail, the program was as follows:

- Monday, September 13, 1999, morning
  Gottfried Vossen:
  Facets of Data Access on the Web

  Chair: G. Vossen

  1. Wolfgang May:
     Information Extraction from the Web with Florid
  2. Hua Shu:
     The Use of XML in Implementing Conditional Tables

- Monday, September 13, 1999, afternoon
  Chair: N. Spyratos

  1. Marc Rittberger:
     Criteria for Resource Selection
  2. Yuzuru Tanaka:
     Meme Market and Meme Base Architectures — Architectures for the Re-editing, Redistribution and Management of Intellectual Resources

- Tuesday, September 14, 1999, morning
  Chair: K. Vidyasankar

  1. Xiaowei Xu:
     Web Mining for E-Commerce
  2. Panos Constantopoulos:
     Context and Terminology Management: A Distributed Information Access Perspective

- Tuesday, September 14, 1999, afternoon
  Chair: K. Vidyasankar

  1. Christa Womser-Hacker:
     Multi-Lingual Information Retrieval
  2. George Samaras:
     Mobile Agents for Web Database Access
  3. Francois Goasdoué:
     A Knowledge-Based Approach for Information Integration: The PICSEL *System*

- Tuesday, September 14, 1999, evening
  Working Groups

- Wednesday, September 15, 1999, morning
  Chair: D. Laurent

  1. K. Vidyasankar:
     Qualifying Recentness of Data
  2. Ute Masermann:
     Schema-Independent Database Querying on the Web
  3. Henrik Loeser:
     Using ORDBSs for Web Data Access

- Wednesday, September 15, 1999, afternoon
  Trip to Trier: guided tour of the city

- Thursday, September 16, 1999, morning
  Chair: G. Moerkotte

  1. Ulrich Thiel:
     Personalized Data Access on the Web: User Profiles for Filtering and Enrichment
  2. Qiang Yang:
     Web Mining for Personalized Web Access

- Thursday, September 16, 1999, afternoon
  Chair: G. Samaras

  1. Jens Lechtenbörger:
     Web Farming
  2. Mikolaj Morzy:
     Web Data Acquisition and Usage
  3. Sanjay Kumar Madria:
     Web Warehousing
     (talk presented by K. Vidyasankar)

- Thursday, September 16, 1999, evening
  Working Groups cont'd

- Friday, September 17, 1999, morning
  Chair: G. Vossen

  1. Guido Moerkotte:
     Query Languages for XML
  2. Findings Reports by the Working Groups

3 Abstracts of Presentations

The following abstracts of presentations appear in alphabetical order of speakers' last names. The titles showing up in this section may vary slightly from the titles of the corresponding presentations as reported in the previous section.

Context and terminology management: A distributed information access perspective

Panos **Constantopoulos**, University of Crete & FORTH

In the Web an uncontrolled number of autonomous data source provides huge amounts of heterogeneous information. Semantic heterogeneity in particular is the most challenging and, probably, the most important to address. Within a strategy of preserving diversity and promoting agreement, rather than integration, we discuss how terminologies and contextual information can support overcoming the difficulties imposed by semantic heterogeneity on accessing information.

We first review the role of ontologies and thesauri in their scope, and the various levels of detail at which agreement can be achieved. Then we discuss issues related to incorporating terminology services in federations of digital collections and we present a 3-tier architecture which allows preserving some autonomy of the sources, while maintaining the consistency to terminology among all involved parties.

Terminology and its usage, as well as the aspect and detail of information stored in the various data sources, and its evolution, all depend on a context set by an individual user, a group of users, or an organizational entity. In order to explicitly account for such manifestations of context-dependence we introduce a notion of context as a conceptual modeling mechanism orthogonal to the others (classification, generalization, attribution). Context supports relativity and modularity in modeling, the use of synonyms, homonyms and anonyms, and the establishment of conceptual paths leading from frames of reference (contexts) as starting points to items relevant to particular information needs.

Context and terminology services can be used, separately or jointly, to provide guidance in exploring the information space, lexical support in formulating the information needs, and description of information resources.

A Knowledge Based Approach for Information Integration: The PIESEL System

Francois **Goasdoué**, University of Paris-Sud, France

Nowadays, a large amount of data are searchable on the web. Data are stored in information sources that can be heterogeneous and distributed. Information integration provides many interesting approaches like mediation, to allow users to access these data. Meditation aims at building a mediator which acts as an interface between users and information sources, giving users the illusion of querying a homogeneous and centralized system. To do this, a mediator provides a unique query language to users and a vocabulary from a semantic description (ontology) of a particular application domain. These ones are used to formulate queries.

In the context of mediation, we present a knowledge-based mediator: the PIESEL system. Its main characteristic is an integration of information sources fully driven by the semantic description of an application domain, and by a semantic description of integrated sources consisting in (i) one-to-one mappings between sources and domain relations (semantic views), (ii) semantic constraints over those views.

However, since XML emerges as a new standard for web documents, we show how easy it is for us to perform the integration of such documents in PIESEL. First, we show how we can capture the semantics of an XML document schema (DTD) thanks to the vocabulary of the application domain semantic description. Then we present a generic approach to connecting a mediator to an XML repository. In PIESEL, it consists of building a generic wrapper which translates a PIESEL query into an X-OQL query (X-OQL is an XML query language). This generic wrapper is not a traditional one, i.e., a fixed set of predefined queries. Our generic wrapper looks like a black box which dynamically generates for any PIESEL query the right X-OQL query.

Web Farming

Jens **Lechtenbörger**, University of Münster, Germany

In this talk, an overview on web farming is presented. The term web farming was coined by Richard Hackathorn to denote the process of integrating external web sources into enterprise data warehouses. Basically, web farming is a straightforward extension of data warehousing, which integrates operational data sources into a separate database for decision making purposes, to include external sources found on the web.

Particular attention is paid to technical aspects of the web farming process, which consists of web content discovery, acquisition, structuring and dissemination. It turns out that the acquisition of web content, i.e., business critical factors, can be achieved by means of a simple database schema, whereas structuring of these factors within the warehouse schema can be realized in terms of a natural extension of the warehouse schema to include a new fact table for each factor.

Using ORDBSs for Web Data Access

Henrik **Loeser**, University of Kaiserslautern, Germany

Object-relational Database Systems (ORDBSs) provide so-called extensibility, i.e., users can add their own types (user-defined types, UDTs), functions (UDFs), and access methods (UDAMs). The extensibility, partly standardized in SQL:1999, can be used to improve Web Information Systems (WISs) by building a Web site management system completely integrated into the ORDBS. By adding a specialized full-text index to the DBS and providing a seamless file system integration into SQL using UDAMs, all documents on the Web server can be indexed. Thus, by letting the DBS keep track of changes to the document base, an up-to-date local search engine can be provided. Moreover, on an intranet with user identification, personalized search can be offered, i.e., users can query all documents available to them. Therefore, the query result not only reflects the actual state of the document base, but also includes all available documents to a specific user, including sections of the Web site with authorization required.

Web Warehousing

Sanjay **Kumar Madria**, Purdue University (presented by K. Vidyasankar)

In this talk, we discuss the design of web warehouses to store materialized views of semistructured data. The objective is to populate the web warehouse based on the user defined query graphs and further manipulate the stored views to gather useful information. We have designed the web algebraic operations such as web select, web join and web coupling operators to manipulate the web views. We also discuss the query language, knowledge discovery and change management with respect to our web warehouse called WHOWEDA. We briefly discuss ongoing work and implementation of our model.

Schema-Independent Database Querying on the Web

Ute **Masermann**, University of Lübeck, Germany

The majority of the tools available for browsing and searching information on the Web is based on extracting information from *structured* documents. However, as information on the Web increasingly comes out of a database it is crucial to be able to search *databases* when working with the Web. Mainly due to the highly dynamic nature of the Web it is unlikely to know the underlying schemata of those databases.

We introduce an extension of SQL called *Reflective SQL* (RSQL) which treats data and queries in a uniform way. Queries are stored in particular *program relations* and can be evaluated by a new, LISP-style operator called *eval*. Program relations cannot only be constructed for given queries, but their contents can also be generated dynamically based on the current contents of the database. This kind of meta-programming allows to build schema-independent queries. RSQL serves as a basis for two novel query languages (a keyword based search and *Schema-Independent SQL* (SISQL)) which render it possible even for inexperienced users to formulate queries to databases in the absence of schema-knowledge. It is shown how these languages can be exploited as a search engine that works on databases instead of documents.

Information Extraction from the Web with FLORID

Wolfgang **May**, University of Freiburg, Germany

The talk presents an integrated architecture where Web exploration, wrapping, mediation, and querying is done in a monolithic system. The system is based on a unified framework – i.e., data model and language – in which all tasks are done. We regard the Web and its contents as a unit, represented in an object-oriented data model: the Web structure, given by its hyperlinks, the parse-trees of Web pages, and its contents all becomes part of the internal world model of the system. The advantage of this unified view is that the same data manipulation and querying language can be used for the Web structure and the application-semantic model: The model is complemented by a rule-based object-oriented language which is extended by Web access capabilities and structured document analysis and allows for accessing the Web, wrapping, mediating, and querying information. Due to this integration, a system in this architecture can be equipped with Web navigation and exploration functionality. We present generic rule patterns for typical extraction, integration, and restructuring tasks using this framework. We show the practicability of our approach by using the FLORID system. The approach is illustrated by two case-studies.

Papers and slides of the group, and the FLORID system can be found at
http://www.informatik.uni-freiburg.de/ dbis/

Query Languages for XML

Guido **Moerkotte**, University of Mannheim, Germany

We discussed a little bit existing query languages for XML. Where do they come from, how do they look like, and where do they go. More specifically examples for 4 XML query languages where presented. Then, a list of requirements for query languages for XML was discussed and these requirement matched against the query languages. The result was a disaster. So I introduced a new XML-query language, called YAXQL, that fulfills all the requirements. An example illustrated YAXQL. There was further some discussion on the features of YAXQL.

Web Data Acquisition and Usage

Mikolaj **Morzy**, University of Münster, Germany

The World Wide Web is becoming more and more popular medium for commerce. The number of e-customers increases dynamically every month and so does the financial value of electronic markets. E-commerce differs from traditional merchandise in many aspects and poses new challenges, among them the question of how to efficiently gather the data about customers. These data can be extracted from web server logs, user navigation paths or histories of user's purchases and they can be used to better understand customer behavior patterns, to predict customer's needs and expectations or to formulate accurate customer characteristics. The quality of these data are crucial for successful web marketing as well. This talk presents various methods for customer data acquisition from the World Wide Web. It addresses the questions concerning user privacy and anonymity. Obtaining information from the users who are highly unwilling to reveal any details about their behavior or identity is a nontrivial technical problem and some attempts, which have tried to solve this problem, are described. The data collected from the Internet contain large amounts of valuable knowledge, which could be used to improve the quality of service of a company, a web-site, etc. Prior to that the knowledge has to be extracted from the massive repository, which could be a warehouse, a database, a web log. Data mining techniques, such as association rule discovery, classification or clustering, could be applied to the data collected from the Internet. The discussion of the application of these techniques and the impact they could have on the Web concludes the talk.

Criteria for Resource Selection

Marc **Rittberger**, University of Konstanz, Germany

There are a lot of different types of resources on the web as web pages, web sites, databases, electronic markets or information about experts. In respect to electronic markets one can differ six levels of quality: Content, presentation, interaction, system, provider and personal level. Based on this levels we present an index number system to evaluate electronic market places. We used the index number system to evaluate 12 electronic market places at the Lake of Constance area. For evaluating the electronic market places in the Lake of Constance area we worked with three different user models, which represent typical users at the Lake of Constance area. As another example of identifying resources on the web we show a useful way to retrieve a group of experts which are relevant to a special topic. We introduce a model to represent a group of experts in an open hypertext system like the KHS and discuss the necessity of an elaborated representation of experts.

Mobile Agents for Web Database Access
George **Samaras**, University of Cyprus

Wireless mobile computing breaks the stationary barrier and allows users to compute and access information anywhere and anytime. However, the severe restrictions induced by wireless connectivity and mobility have a great impact on the design and structure of mobile computing applications and motivate the development of new computing models. To this end, a number of extensions to the traditional distributed system architectures have been proposed. These new software models, however, are static and require a priori set up and configuration. This in effect limits their potential in dynamically serving the mobile client; the client cannot access a site where an appropriate model is not configured in advance. The contribution of this work is twofold. First, it shows how an implementation of the proposed models using mobile agents eliminates this limitation and enhances the utilization of the models. Second, new frameworks for web-based distributed access to databases are proposed and implemented via mobile agents.

The Use of XML in Handling Dynamic Views
Hua **Shu**, Karlstad University, Sweden

This talk discusses the specification of the extended relational data model using an XML application called ERDBML (Extended Relational Database Markup Language). The design goal of ERDBML is to allow authors to describe the content of relational and object-relational databases in order to facilitate storage and exchange of data between the databases and applications, in particular the applications that dynamically maintain materialized views. Our previous work showed that conditional tables, i.e. tables containing variables and logical conditions associated with the tuples, are useful as a data model in dynamically maintaining materialized views. In implementing this approach to view maintenance using conditional tables, we need an efficient way to specify conditional tables. The mark-up language is designed for this purpose. The language is compatible with MathML (Mathematical Markup language) and can be easily extended to handle other abstract data types.

Meme Market and Meme Base Architectures — Architectures for the Re-editing, Redistribution and Management of Intellectual Resources

Yuzuru **Tanaka**, Hokkaido University, Japan

Computers are expanding their target of augmentation from individuals and groups to societies. While people in a group share a definite goal, people in a society share their achievements in the form of their knowledge. The augmentation of societies requires a new type of media that can carry varieties of knowledge resources, replicate themselves, recombine themselves, and be naturally selected by their environment. They may be called Meme media since they carry what R. Dawkins called "memes". An accumulation of memes in a society forms a meme pool that functions like a gene pool. When economic activities are introduced, a meme pool becomes a meme market where providers and distributors of memes carry their business without prohibiting users' re-editing and redistributing memes. Meme pools and meme markets will bring about rapid accumulations of memes, and require new technologies for the management and retrieval of memes. This talk has proposed system architectures for meme media, meme pools, meme markets, and meme bases.

First, the talk has given a brief review of IntelligentPad and IntelligentBox as 2D and 3D meme media architectures. Then, it has proposed a meme pool architecture using a special pad, a component of IntelligentPad, called PiazzaPad. Each PiazzaPad is associated with a Piazza server, which allows users to register and retrieve pads to and from its storage file. When a PiazzaPad is loaded and opened on the desktop, all the pads registered in its associated storage file are also loaded from its server and arranged on itself. Each PiazzaPad works as a piazza where people come together to open a market. PiazzaPads allow us to publish any pads or any boxes in any PiazzaPads just by drag and drop operations. PiazzaPads can be also dropped in any other PiazzaPad. PiazzaPads enable end users to open his own gallery of pads in the Internet or in some other's or public pad gallery. These pad galleries work as flea markets, shops, shopping centers, community message boards, community halls, or plazas.

The super-distribution of pads using request modules and account modules introduces business activities in the meme pool and makes it work as an international meme market.

For the management of meme media objects, we may consider two alternative ways. If we have to manage a large number of pads of a few different forms, we can keep the form information outside the databases; we only need to store the state information of pads in the databases. Such a database is called a form base. If we have to manage composite pads of a large number of different forms, we need database facilities not only to manage the states of the pads but also to manage their different forms. Such a database is termed a pad base. Pad bases are instance bases by their nature, and require further research. IntelligentPad allows us to treat any composite pad as a database value. This means any object that can be represented as a composite pad can be stored in relational databases. These objects include interactive multimedia objects such as images, movies and sounds, interactive charts and tables, interactive maps, database access forms, varieties of application tools, and compound documents embedding any of

these objects.

Meme media and meme base architectures are applied to various applications. Researchers on nuclear reaction database have applied them to the exchange and publication of experimental data and analysis tools. They distribute these data and tools for others to re-edit and to reuse these. The Miyako system, an interactive digital archive system for Kyoto cultural heritage, uses IntelligentPad to store archived contents in a relational database, and to access them through an interactive hypermedia environment. IntelligentPad and IntelligentBox can be also applied to visualization or interactive animation of database records. Meme media and meme base architectures provide a generic framework for such systems.

Personalized Data Access on the Web: User Profiles for Filtering and Enrichment

Ulrich **Thiel**, GMD-IPSI

User-initiated access to Web-sites via browsing or querying was conceived for traditional web-sites which basically can be regarded as information stores. With the advent of highly dynamic web-sites, however, alternative ways of access are needed. For instance, users of a news agency site may submit a profile, and receive a stream of messages after non-relevant messages were filtered out.

To address the needs of Web-based information services with highly dynamic offers, advanced filter methods can be employed in combination with enrichment techniques and automatic page layout. This approach was adopted in the TREVI (Text REtrieval of Vital Information) project. Based on the results of syntactic and semantic parsing of an incoming newswire message, the TREVI system applies Bayesian networks and dedicated subject identification rules to index the new message. A set of filtering strategies is available to decide on the relevance of the message w.r.t. a given user profile. A user-specified background repository, e.g., an archive containing previous messages, is then accessed to enrich the incoming news.

The system incorporates elaborated presentation components, using different distribution modes. TREVI results can be included in electronic journals or published as web-pages, or sent via email.

Qualifying Recentness of Data

K. **Vidyasankar**, Memorial University of Newfoundland, Canada

Underlying any activity in the World-Wide Web is the finding of new or "recent" data. Recentness can be measured in different ways. In this talk, we identify five different notions of recentness: We define them formally in terms of system executions in shared read/write variables. We then give several examples illustrating the occurrences and usefulness of these recentness notions in Web applications.

Facets of Data Access on the Web

Gottfried **Vossen**, University of Münster, Germany

The last five or six years have seen a tremendous growth and increase in computer usage, triggered by the launch of the Web. While previously the Internet was largely used for email and other information exchange services, the Web has made it a platform for education, entertainment, communication, commerce, and many other things. Web tools have emerged from simple directories and search engines into malls and portals, and the Web has already brought along novel business models, such as banner advertising and auctions. The use of database systems in the context of the Web is ubiquitous, and many attempts have been made of transferring database techniques to the Web. Nevertheless, important problems remain, such as how provide, find, structure, and organize Web information. There are also open issues w.r.t. security, privacy, language support, mark-up formats, etc.

This seminar will study some of these problems, try to identify views people can have on the Web and how to utilize approaches from database systems, information retrieval, digital libraries and related areas.

Multi-Lingual Information RETRIEVAL

Christa **Womser-Hacker**, University of Hildesheim, Germany

"Die Grenzen meiner Sprache bedeuten die Grenzen meiner Welt." (Ludwig Wittgenstein, Tractatus logico-philosophicus, p. 67)

The handling of multilinguality is an important challenge for user-friendly information systems. In my talk, I focussed on information process starting with HCI, different search paradigms, and the two main issues in Information Retrieval which are the representation of items and the search for items. I talked about the indexing process providing some examples. Concerning access to multi-lingual information resources, I differentiated between multilinguality and crosslinguality. The connected linguistic problems range from language identification to sophisticated NLP and machine translation techniques applied to multi-lingual environments. Within an IR process several ways to integrate multi-lingual concepts were considered:

- multilinguality at the object level,
- multilinguality at the index level,
- multilinguality at the query level.

At the system level, I described some efforts on multilinguality and cross- language retrieval (e.g. Eurospider, Mulinex, Twenty-One). Finally, I concluded by talking about the TREC initiative in the US which includes cross-language IR and intends to reach a deeper understanding of the performance of such systems (http://www.nist.nlpir.gov).

References:

- Berry, M.W., Young, P.G. (1995), Using Latent Semantic Indexing for Multilingual Information Retrieval. In: Computers and Humanities, Vol. 29, Dordrecht et al., Kluwer Academic Publishers, December, 1995. pp. 413- 429.
- Erbach, G., Neumann, G., Uszkoreit, H., MULINEX: Multilingual Indexing, Navigation and Editing Extensions for the World-Wide-Web. In: THIRD DELOS WORKSHOP. Cross-Language Information Retrieval. Zurich, 5-7 March 1997. http://www.ercim.org/publications/ws-proceedings/DELOS3/index.html
- Harman, D.K. (1995c), The TREC Conferences. In: Kuhlen, R., Rittberger, M. (Eds.), Hypertext - Information Retrieval - Multimedia. Synergieeffekte elektronischer Informationssysteme. Proceedings HIM '95, S. 9-28.
- Klavans, J., et al. A Natural Language Approach to Multi-Word Term Conflation. In: THIRD DELOS WORKSHOP. Cross-Language Information Retrieval. Zurich, 5-7 March 1997. http://www.ercim.org/publications/ws-proceedings/DELOS3/index.html
- Kraaij, W., The EU Project "Twenty-One" and Cross-Language IR. In: THIRD DELOS WORKSHOP. Cross-Language Information Retrieval. Zurich, 5-7 March 1997. http://www.ercim.org/publications/ws-proceedings/DELOS3/index.html

- Oard, D.W. (1997), Cross-Language Text Retrieval Research in the USA. In: THIRD DELOS WORKSHOP. Cross-Language Information Retrieval. Zurich, 5-7 March 1997.
  http://www.ercim.org/publications/ws-proceedings/DELOS3/index.html

- Peters, C., Picchi, E., Using Linguistic Tools and Resources in Cross- Language Retrieval. In: THIRD DELOS WORKSHOP. Cross-Language Information Retrieval. Zurich, 5-7 March 1997.
  http://www.ercim.org/publications/ws-proceedings/DELOS3/index.html

- Sheridan, P., Schuble, P., Cross-Language Multi-Media Information Retrieval. In: THIRD DELOS WORKSHOP. Cross-Language Information Retrieval. Zurich, 5-7 March 1997.
  http://www.ercim.org/publications/ws-proceedings/DELOS3/index.html

- Thurmair, G., Womser-Hacker, C. (1996), Multilingualität im wissensbasierten Faktenretrieval. In: Krause, J., Herfurth, M., Marx, J. (Hrsg.), Herausforderungen an die Informationswirtschaft. Informationsverdichtung, Informationsbewertung und Datenvisualisierung, (ISI'96). Konstanz. pp. 121-132

- Wittgenstein, L. (1984), Tractatus logico-philosophicus. Werkausgabe Band I. Suhrkamp Wissenschaft.

Web Mining for E-Commerce

Xiaowei **Xu**, Coporate Technology, Siemens AG

In this talk we presented an architecture for data mining. The goal is to extract knowledge which is helpful for e-business to make right decision at right time. To achieve that goal, the Web log data, transaction data, corporate product data, and external information about competitors are loaded and integrated into a warehouse. We give an example of Web mining for e-commerce. The requirements on data mining algorithms are analysed. We have studied how to design algorithms to satisfy these requirements. As a case study, a density-based clustering algorithm is presented.

Web Mining for Personalized Web Access

Diang **Yang**, Simon Fraser University, Canada
and Microsoft Research, China

Multimedia access on the world wide web can be a costly process. One way to alleviate the problem is to learn user's oder user group's preference models, and to use the models to predict the next multimedia file the user might access. In this presentation, we discuss how to construct accurate prediction models for data access from usage logs and product profiles. We also discuss how to select data sources based on a query optimization model. Practical issues regarding the lack of information are also discussed.

4 Working Groups

During the week three different working groups were formed, each of which met several times to discuss particular issues in depth. Following are the summaries of two of these groups, which were presented on the last morning of the seminar. For the third, entitled *Context-Sensitivity in Web Searching*, we were unable to obtain a transcript.

4.1 Impact of XML on the Web

The group has discussed the possible impact of XML on the following Web-related tasks:

1. Web-side management,
2. data exchange on the Web,
3. improvement of search engines,
4. data integration, and
5. data reuse.

It seems clear that XML can help accomplish many of these tasks. For example, XML can help improving search engines since DTDs provide useful information for indexing. The reason why database people care about XML is that they want to handle semi-structural data. Therefore, XML will have an impact on database research and on certain applications on the Web that involve semi-structured data. The impact on database research is obvious if one considers the various XML query languages under development.

However, we believe that XML will have little impact on the Web if a number of prerequisites are not satisfied. The following is a list of the prerequisites discussed in the working group:

1. There are sufficiently many tools that support XML (browsers, editors, XML servers, etc),
2. XML (and DTDs) is (are) directly accessible as the source and not hidden behind XSL,
3. there are standardized DTDs,
4. there is data in XML format.

Prerequisites 1. and 3. are the less problematic ones. Indeed, there are plenty of tools for XML (e.g., XML vs. DBMS, etc) and the available tools will be improved.

It is uncertain whether prerequisite 2. will be satisfied or not. Two notions of visibility need to be considered: Global visibility means that a document is visible to everyone. Local visibility means that it is visible to a restricted community. We believe that XML will have a major impact on the Web only when XML is globally visible.

Another open question is whether there will be enough data in XML on the Web; at the moment there is very little, if at all.

A conclusion of this working group is that the future of XML depends on whether the data management community believes in XML and provides data in XML format or not. If many people believe in XML, then they will provide data in XML, which in turn enables XML to have a major impact on the Web.

Participants: Loeser, Masermann, May, Moerkotte, Shu, Vossen

4.2 Web Mining for Web Access and Web Access for Web Mining

To start with, we asked the following questions:

- What data to obtain for Web mining?
- How to get the data for Web mining?
- What is the Business purpose of Web mining?
- What to mine in the data?
- How to use the knowledge obtained?
- How to mine competitors' information online?
- How to evaluate the interestingness of mined knowledge?

The first part of the discussion centered on the objectives of Web mining. We divided the objectives into two categories: Ecommerce and System Performance related. In Ecommerce related Web mining, the objective is to mine users' preferences and needs, and to use this knowledge to enhance business. In this category, the sub-category of business to business Ecommerce is easier to do, because the activities are more regular and more cooperative. On the other hand, Business to Customer type of Ecommerce is harder, because of anonymous users and groups. Further, proxies and ISP's tend to provide incomplete log information. Moreover, there was discussion on issues such as sparse purchasing that produces more non-comparable objects, which increases the difficulty of Web mining..

It was also pointed out that in Ecommerce related activities, it is possible to study the structure of other business' web structure to formulate one's own and to provide targeted advertisement.

The second category of Web mining relates to system performance related activities. The goal is to enhance the performance of Web-based systems through a variety of technologies, such as the push technology, caching, pre- sending and pre-fetching of videos and multimedia files.

The second part of the discussion focused on the difficulties of Web mining. The first question raised was the difference between Web mining from regular data mining. It was observed that in the latter, lots of incomplete information and inaccurate information about both users and products exist. The dynamic nature of the Web data determines

that data change too often; the lack of a satisfactory ontology entails that no classification hierarchies are universally available. It was also pointed out that the existence of multiple servers and server logs motivates the need for distributed Web mining.

The participants of the working group also wondered aloud new applications of Web mining. For example, there is a particular need for a watchdog of E-Commerce, who uses Web mining to study the trust-worthiness of a E-Commerce company. This raises a more general question: How to ensure that data is current, correct, complete, consistent, and "good".

Other applications of Web access include Web Farming , an activity where data is obtained externally. It was noticed that perhaps a better name could be found when the "seed" of the farm was actually planted by other people. There may be a need to set up insurance services for measuring data "goodness"; likewise, there may be needs to find new forms of non-relational representation for the linkage information and unstructured data on the Web.

In conclusion, the working group felt that Web access can benefit greatly from Web mining, and that further research and development activities in Web Mining require more effective Web access as support.

The participants of the working group were: Xiao-Wei Xu, Nicolas SPYRATOS, George SAMARAS , Yuzuru TANAKA, Gottfried VOSSEN, Nicolas MORZY, Dominique LAURENT, Qiang Yang and Mirian Halfeld Ferrari. The note taker was Qiang Yang.

5 List of Participants

Panos Constantopoulos
FORTH & University of Crete
Institute of Computer Science
Science & Technology Park of Crete
Vassilika Vouton
P.O. Box 1385
GR-71110 Heraklion, Crete, Greece

phone: +30-81-391-634
fax: +30-81-391-601
e-mail: panos@ics.forth.gr

Francois Goasdoué
LRI, U.R.A. 410 du CNRS
Bât. 490 - Université de Paris-Sud
F-91405 Orsay Cedex, France

phone: +33-1-69 15 58 46
fax: +33-1-69 15 65 86
e-mail: goasdoue@lri.fr
url: www.lri.fr/ goasdoue/

Mirian Halfeld-Ferrari
Université de Tours
IUP Informatique
3, Place Jean Jaures
F-41000 Blois, France

phone: +33-0254 55 27 72
e-mail: mirian@univ-tours.fr

Dominique Laurent
Université de Tours
IUP Informatique
3, Place Jean Jaures
F-41000 Blois, France

phone: +33-2-54 55 21 11
fax: +33-2-54 55 21 32
e-mail: laurent@univ-tours.fr

Jens Lechtenbörger
Institut für Wirtschaftsinformatik
Universität Münster
Steinfurter Strasse 107
D-48149 Münster, Germany

phone: +49-251-8338-158
fax: +49-251-8338-159
e-mail: lechten@helios.uni-muenster.de

Henrik Loeser
Fachbereich Informatik, AB DBIS
Universität Kaiserslautern
Postfach 3049
D-67653 Kaiserslautern, Germany

phone: +49-631-205-3283
fax: +49-631-205-3299
e-mail: loeser@gmx.de
url: www.uni-kl.de/AG-Haerder/Loeser/

Ute Masermann
Medizinische Universität Lübeck
Institut für Informationssysteme
Osterweide 8
D-23562 Lübeck, Germany

e-mail: maserman@informatik.mu-luebeck.de
url: www.informatik.mu-luebeck.de/
maserman/

Wolfgang May
Institut für Informatik
Universität Freiburg
01-026, Gebäude 51
Am Flughafen 17
D-79110 Freiburg, Germany

phone: +49-761-203-8131
fax: +49-761-203-8122
e-mail: may@informatik.uni-freiburg.de
url: www.informatik.uni-freiburg.de/ may/

Guido Moerkotte
Lehrstuhl Praktische Informatik III
Universität Mannheim
D7, 27
D-68131 Mannheim, Germany

phone: +49-621-292-8820
fax: +49-621-292-8818
e-mail:
moer@pi3.informatik.uni-mannheim.de
url: pi3.informatik.uni-mannheim.de/

Nicolas Morzy
Institut für Wirtschaftsinformatik
Universität Münster
Steinfurter Strasse 107
D-48149 Münster, Germany

phone: +49-251-8338-154
e-mail: morzy@helios.uni-muenster.de
url: wwwmath.uni-muenster.de/informatik/
u/dbis/Morzy/index.html

Marc Rittberger
Institut für Informationswissenschaft
Universität Konstanz
Postfach D 87
D-78457 Konstanz, Germany

e-mail: marc.rittberger@uni-konstanz.de
url: www.inf-wiss.uni-konstanz.de/people/
mr.html

George Samaras
Dept. of Computer Science
University of Cyprus
75 Kallipoleos Street
P.O. Box 20537
CY-1678 Nicosia, Cyprus

phone: +357-2-892233
e-mail: cssamara@turing.cs.ucy.ac.cy
url: www.cs.ucy.ac.cy/ cssamara

Hua Shu
Dept. of Computer Science
Karlstad University
S-65634 Karlstad, Sweden

phone: +46-54-700-2156
fax: +46-54-700-5060
e-mail: hua.shu@kau.se
url: www.cs.kau.se/ hua/

Nicolas Spyratos
Laboratoire de Recherche en Informatique
U.R.A. 410 du CNRS
Bât. 490 - Université de Paris-Sud
F-91405 Orsay Cedex, France

e-mail: spyratos@lri.fr
url: www.lri.fr/Francais/Recherche/bd.html

Yuzuru Tanaka
Meme Media Laboratory
Hokkaido University
N13 - W8
J-060 8628 Sapporo, Japan

phone: +81-11-706-7252
fax: +81-11-706-7808
e-mail: tanaka@meme.hokudai.ac.jp

Ulrich Thiel
GMD-IPSI
Dolivostrasse 15
D-64293 Darmstadt, Germany

phone: +49-6151-869-855
fax: +49-6151-869-818
e-mail: thiel@darmstadt.gmd.de

Yannis Tzitzikas
FORTH & University of Crete
Institute of Computer Science
Science & Technology Park of Crete
Vassilika Vouton
P.O. Box 1385
GR-71110 Heraklion, Crete, Greece

e-mail: tzitzik@csi.forth.gr
url: www.csi.forth.gr/ tzitzik

K. Vidyasankar
Dept. of Computer Science
Memorial University of Newfoundland
St. John's, Newfoundland
Canada A1B 3X5

phone: +1-709-737-4369
fax: +1-709-737-2009
e-mail: vidya@cs.mun.ca
url: www.cs.mun.ca/ vidya/

Gottfried Vossen
Institut für Wirtschaftsinformatik
Universität Münster
Steinfurter Strasse 107
D-48149 Münster, Germany

phone: +49-251-83-38151/0
fax: +49-251-83-38159
e-mail: vossen@helios.uni-muenster.de
url: wwwmath.uni-muenster.de/informatik/u/dbis/Vossen/index.html

Christa Womser-Hacker
FB Sprachen und Technik
Universität Hildesheim
Marienburger Platz 22
D-31141 Hildesheim, Germany

phone: +49-5121-883-833
fax: +49-5121-883-802
e-mail: womser@cl.uni-hildesheim.de
url: www.uni-hildesheim.de/aiw

Xiaowei Xu
Siemens AG
Corporate Technology
ZT IK4
Otto-Hahn-Ring 6
D-81730 München, Germany

fax: +49-89-636-49767
e-mail: Xiaowei.Xu@mchp.siemens.de

Qiang Yang
School of Computer Science
Simon Fraser University
Burnaby, British Columbia
Canada V5A 1S6

phone: +1-604-291-5415
fax: +1-604-291 3045
e-mail: qyang@cs.sfu.ca
url: fas.sfu.ca/cs/people/Faculty/Yang