IR Group Focus –

Dortmund University

University of Dortmund

Department of Computer Science

Information Retrieval Group

In this edition of the Informer we look at one of the many Information Retrieval groups on the continent, the Information Retrieval group at the University of Dortmund. Headed by Professor Norbert Fuhr, they are interested in a variety of applications of Information Retrieval, but lately have been focusing their attention to digital libraries, XML retrieval, and multimedia retrieval. In this article we present the history, the people and some of their current projects.

History and research topics of the Dortmund IR group

The Dortmund IR group started in 1991, when Norbert Fuhr was appointed Professor at the Computer Science Department of the University of Dortmund. In the same year, the specialist IR group of the German informatics society (GI) was founded. The charter of this specialist group defines Information Retrieval as a discipline which deals with uncertainty and vagueness in all kinds of information systems. Following this broad concept, the Dortmund IR group is mainly interested in extending IR models and methods for dealing with problems beyond the classical text retrieval task. In particular, the combination of concepts from IR and database systems is an ongoing theme of the work in Dortmund, with applications such as relational databases, multimedia information systems, distributed digital libraries, and XML documents.

As theoretic background for the new types of applications, the group combines Norbert Fuhr's earlier work on probabilistic IR models with logic-based approaches. A major outcome of this work was the development of probabilistic Datalog during the ESPRIT project FERMI (1994-1997), which focused on retrieval methods for multimedia documents. Based on this model, the retrieval engine HySpirit was implemented, which offers flexible and efficient retrieval mechanisms even for large data sets. Subsequently Dr. Thomas Rölleke a former member of the group, commercialized the HySpirit retrieval engine, founding a start up company bearing the same name in 1999.

During the mid-90’s, there was a move towards multimedia information systems and digital libraries (DLs) becoming an important area of application. Since 1995, the group has focused on developing techniques in this field. Some of the other areas that the group has recently worked and is currently working on include:

Networked IR

In the projects MeDoc (1995-1997), Interdoc (1998) and MIND (2001-2002), the group works on the development of new probabilistic models for resource selection and result fusion, addresses the issue of heterogeneity with respect to the database schemas and retrieval methods, and extends these approaches for retrieving multimedia data.

XML retrieval

The goal of the projects CARMEN (1999-2001) and CLASSIX (2002-2004) is the development of IR methods for XML documents. A major result is the development of the query language XIRQL and its implementation within the new retrieval engine HyREX, which is also used in CYCLADES project (2001-2003) for the retrieval of records from Open Archives.

User-oriented retrieval methods

Based on the ideas of Bates et al., the DAFFODIL project (2000-2004) develops a new front end for federated digital libraries that supports high-level search activities in an adaptive and proactive way.

Evaluation of Digital Libraries

Within the DELOS Network of Excellence (2000-2002), Norbert Fuhr leads the working group “Digital Library Test Suite” aiming at the development of evaluation methods and test beds for digital libraries. In cooperation with the FOCUS project (2000-2002), a test bed for full text retrieval of XML documents will be developed this year.

The members of the Dortmund Information Retrieval

Group Leader

Professor Norbert Fuhr

In 1991, Norbert became the first chairman of the then newly founded German Informatics Society – Specialist IR Group

IR Group Members

Mohammad Abolhassani

Gudrun Fischer

Gudrun is hopping between the thousands Open Archive data islands of CYCLADES!

Norbert Gövert

Norbert works on HyREX, but also enjoys Charles Schulz's Peanuts!

Kai Großjohann

Claus-Peter Klas

Henrik Nottelmann

Henrik will be attending the ECIR 2002 – Are you?

Projects

Continuing from their successes in the past, the Dortmund IR group are now working on a plethora of projects. Some of these include: a classification and intelligent search engine for information in XML format (CLASSIX), the development of effective methods for dealing with the retrieval of structured documents (FOCUS) and the development of distributed agents which facilitate user friendly access of digital libraries (DAFFOFIL). Following we present a few details about four of their major projects; HyREX, MIND, DAFFODIL, and CYCLADES.

The HyREX Project

The HyREX project is an ongoing effort (funded as part of other project like e.g. CARMEN, CYCLADES and CLASSIX) for developing an IR engine for XML documents. The main collaborators from the Dortmund Information Retrieval group are Mohammad Abolhassani, Norbert Fuhr, Norbert Gövert, and Kai Großjohann.

The current W3C activities for the development of a standard query language for XML (XQuery) are targeting towards database-oriented applications and thus do not consider the needs of IR. In contrast, the Dortmund group focuses on document-oriented XML applications, where retrieval must take into account the intrinsic imprecision and vagueness of IR.

For this purpose, the query language XIRQL (XML IR Query Language) has been developed, which extends the XPath part of the (proposed standard) query language XQuery by the following features:

Weighting and ranking

Whereas XQuery supports Boolean retrieval only, XIRQL allows for weighting document terms as well as query terms. For the former, it is assumed that the weight of a term depends on its context (the definition of these contexts is given as part of an extended DTD). The underlying probabilistic model treats all term occurrences within the same index node as a single probabilistic event. Query processing produces a Boolean combination of these basic events, for which the correct probabilities can be computed (following the concept of event expressions from probabilistic Datalog).

Relevance-oriented search

Traditional IR queries specify only the requested content, but pose no restrictions on the structure of the result. In this case, the IR system should be able to retrieve the most relevant parts of XML documents by choosing the most specific element(s) that satisfy the query.

Data types and vague predicates

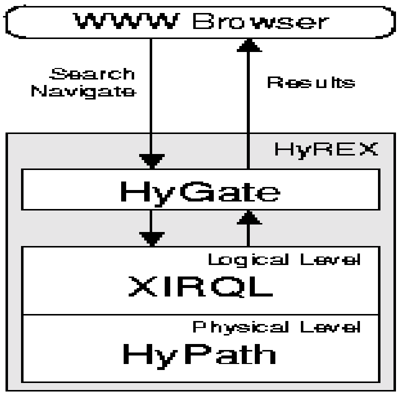
Since XML allows for a fine-grained mark up of elements, there should be the possibility to use special search predicates for different elements of various data types (e.g. person names, dates, technical measurement values, names of geographic regions). For each data type, the system must provide appropriate search predicates, most of which should be vague (e.g. phonetic similarity of names, approximate matching of dates, and closeness of geographic locations).

<XML> HyREX </XML>

It ain't cool if it ain't marked up!

Structural relativism

XML query languages allow for conditions with respect to the structure of the documents to be retrieved. In order to support uncertainty and vagueness for this type of conditions, appropriate methods ignore the difference between elements and attributes, searching for elements of a specific data type (e.g. search in all elements containing person names) or by exploiting hierarchies over element names defined in an ontology.



The architecture of HyREX

In contrast to XIRQL, XQuery offers additional operators for aggregation and restructuring of results. Further research will focus on appropriate extensions of XIRQL, i.e. probabilistic versions of the corresponding XQuery operators.

The XIRQL language is implemented within the HyREX (Hypermedia Retrieval Engine for XML) system. Its system architecture is similar to that of database management systems. Thus, there is a clear separation between the logical and the physical level.

At the logical level, XIRQL queries are transformed into a path algebra. A path describes the sequence of document nodes leading from the root of an XML document root to a specific element. The path algebra contains operators for manipulating sets of paths that describe intermediate results in query processing. After mapping a XIRQL query into a path algebra expression, the query optimization step transforms this expression into an equivalent one which (hopefully) can be processed more efficiently. Since users typically want to see the top ranking elements only, retrieval strategies focusing on these elements will be investigated.

The connection between the logical and the physical level is formed by the vague predicates, which take a value and/or structure conditions as arguments and return a list of paths as result. In order to perform efficient retrieval, appropriate index structures have to be available at the physical level. Whereas classical inverted lists support value conditions only (indicating occurrence/weights of terms), XIRQL queries may also contain conditions referring to element names and/or indexes as well as to sequence and aggregation of elements. Since inclusion of the necessary information in the inverted list entries will lead to large storage overheads, appropriate compression schemes are investigated.

The development of a user interface to an XML IR engine poses a number of new challenges. HyREX currently supports only a simple Web browser interface where users may enter XIRQL queries and receive ranked lists of answers. For query formulation, several variants based on the concept of query by example are under consideration; as an example, either the DTD, the logical structure or the final layout of a specific document can be used. Visualization of results has to cope with the fact that different matches may occur within the same document, where even a match may contain others; here variants of tile bars and tree maps are studied.

Open Source software of the HyREX is available from http://ls6-www.cs.uni-dortmund.de/ir/projects/hyrex/. The current version allows for efficient retrieval of XML document collections up to the gigabyte range.

MIND

The MIND is a Resource Selection and Data Fusion system for Multimedia International Digital Libraries. The project began in early 2001 and is due to be completed by mid 2003. It is sponsored by EU FP5 and involves several other institutes including the University of Strathclyde, the Universita di Firenze, the University of Sheffield and Carnegie Mellon University.

MIND
Resource Selection and Data Fusion for Multimedia International Digital Libraries

The all seeing mind's eye!

Norbert Fuhr and Henrik Nottelmann lead development of the project at the University of Dortmund.

The MIND project addresses problems associated with the emergence of thousands of heterogeneous multimedia Digital libraries distributed internationally on multiple platforms. Users typically have problems with resource selection, as they are unaware of the contents of each individual library in terms of quantity, quality, information type, provenance and likely relevance. When a set of relevant libraries has been selected, the user must organize and interpret the information in a common format and environment. This is performed through visual evaluation and ad hoc integration, which forces users to restrict their attention to a small subset of the information retrieved.

MIND attempts to assist users to know where to search, how to query different media, and how to combine information from diverse sources.

The University of Dortmund's Information Retrieval Group addresses the issues of resource selection and heterogeneity:

Resource selection

The basis is a decision-theoretic framework (developed by Dortmund) which will be refined within this project. Each database has assigned costs (covering retrieval quality, communication time, and monetary costs). Given a query (containing the number of documents to retrieve), the task is to compute (for efficiency, this number should be zero mostly) for every database the number of documents to retrieve from that database. Of course, the sum should equal the user-specified number of documents to retrieve, and the overall costs should be minimized.

Heterogeneity

The existing databases differ in terms of content and structure (schema) of its documents (e.g., they can distinct "editor" and "author"). Thus, the user query (specified against a global schema) must be translated for every database into a query fitting the database schema. This query transformation is based on uncertain predicate logic rules which will be learned from an example set.

DAFFODIL picks the best from a bouquet of digital libraries.

DAFFODIL

The DAFFODIL project (Distributed Agents for User-friendly Access of Digital Libraries) develops an agent-based front end for federated digital libraries. Based on the ideas of Bates et al., strategic support for information searches is provided by offering multiple ways for accessing literature; standard metadata search can be enhanced by invoking a thesaurus, browsing through a classification leads to documents of a selected category, the author network tool displays coauthor relationships in a graph, citing/cited publications of a given document are retrieved via the references tool, and the journal and conference tools support browsing through the respective tables of contents. In specific search situations, proactive agents suggest the invocation of these methods to the user, and the system is able to adapt its behavior to the user's preferences. In the near future, components for personalization and collaboration will be integrated in the DAFFODIL system. You can try it out at http://www.daffodil.de/!

CYCLADES

The Cyclades project aims to provide an open collaborative virtual archives environments. To be completed in mid 2003, Gudrun Fischer and Norbert Fuhr are working in partnership with Consiglio Nazionale delle Ricerche-Istituto di Elaborazione dell'Informazione, (IEI-CNR), European Research Consortium for Informatics and Mathematics ERCIM, Foundation for Research and Technology (FORTH) and Fraunhofer-Gesellschaft.

The CYCLADES logo!

The standards defined by the Open Archives initiative (OAi) provide uniform access (by defining a gatherer interface) to open, heterogeneous and distributed digital archives. The CYCLADES project aims at developing services on top of the OAi standard that support single users and user groups in their work with OAi conform archives. These services will include information retrieval in distributed archives, searching and browsing in multi level hypertext, collecting relevance feedback and on-line annotations, and user profiling.

Architecture

The architecture consists of five main processes:

Access Service – which enables the harvesting and indexing of Metadata, the storage and retrieval of metadata records and the archival of such information.

Search & Browse Service – in the form of multilevel hypertext searching and browsing .

Collaborative Work Service – which facilitates the collaboration between individual scholars, members of project groups, and wider communities, via the use of Shared workspaces, Hierarchy of folders and Rating and annotation.

Collection Service - which allows structuring the information space and the provision for topic-based virtual archives.

A Filtering & Recommendation Service Recommend records to users, communities, and collections and filters query results based on their respective profiles.

Selected Publications

Fuhr, N.; Gövert, N. (2002). Index Compression vs. Retrieval Time of Inverted Files for XML Documents. (Submitted for publication.)

Fuhr, N.; Großjohann, K. (2002). XIRQL: An XML Query Language Based on Information Retrieval Concepts. (Submitted for publication.)

Fuhr, N.; Weikum, G. (2002). Classification and Intelligent Search on Information in XML. In: *IEEE Data Engineering Bulletin* Volume 25.

Fuhr, N. (2001). Information retrieval methods for multimedia objects. In: *State-of-the-art in Content-Based Image and Video Retrieval*, pages 191-212. Kluwer Academic Publishers, Boston, Dordrecht, London.

Fuhr, N. (2001). Language Models and Uncertain Inference in Information Retrieval. In: Callan, J.; Croft, B.; Lafferty, J. (eds.). *Proc. Workshop on Language Modelling and Information Retrieval*, pages 6-11. Carnegie Mellon University, Pittsburgh, PA.

Fuhr, N.; Großjohann, K. (2001). XIRQL: A Query Language for Information Retrieval in XML Documents. In: *Proceedings of the 24th Annual International Conference on Research and development in Information Retrieval*, pages 172-180. ACM, New York.

Fuhr, N.; Klas, C.-P. (2001). Combining RDF and Agent-Based Architectures for Semantic Interoperability in Digital Libraries. In: *Proceedings of the DELOS Workshop on Interoperability in Digital Libraries.*

Fuhr, N.; Hansen, P.; Mabe, M.; Micsik, A.; Solvberg, I. (2001). Digital Libraries: A Generic Classification and Evaluation Scheme. In: *Proceedings European Conference on Digital Libraries*, pages 187-199. Springer, Berlin et al.

Fischer, G.; Fuhr, N. (2001). Heterogeneity in Open Archives Metadata. In: *Proceedings of the Workshop on Experimental OAI based Digital Library Systems.*

Gövert, N. (2001). Bilingual Information Retrieval with HyREX and Internet Translation Services. In: *Proceedings of the CLEF 2000 Workshop, LNCS 2069*, pages 237-244. Springer.

Frommholz, I. (2001). Categorizing Web Documents in Hierarchical Catalogues. In: *Proceedings ECIR '01.*

Nottelmann, H.; Fuhr, N. (2001). Learning probabilistic Datalog rules for information classification and transformation. In: *Proceedings of the 10th International Conference on Information and Knowledge Management*, pages 387-394. ACM, New York.

Nottelmann, H.; Fuhr, N. (2001). MIND: An architecture for multimedia information retrieval in federated digital libraries. In: *Proceedings of the DELOS-Workshop on Interoperability in Digital Libraries.* DELOS-Network of Excellence on Digital Libraries.

Fuhr, N. (2000). Probabilistic Datalog: Implementing Logical Information Retrieval for advanced Applications. In: *Journal of the American Society for Information Science 51(2)*, pages 95-110.

Fuhr, N (1999). A Decision-Theoretic Approach to Database Selection in Networked IR. In: *ACM Transactions on Information Systems 17(3)*, pages 229-249.

Further Information

For further information about the Dortmund Information Retrieval group or the projects they are involved with contact:

Professor Norbert Fuhr
Email: fuhr@cs.uni-dortmund.de

Or visit the group’s website:
http://ls6-www.cs.uni-dortmund.de/ir/