Report on the first Twente Data Management Workshop on XML Databases and Information Retrieval

Djoerd Hiemstra and Vojkan Mihajlović
University of Twente, The Netherlands
{d.hiemstra, v.mihajlovic}@utwente.nl

1 Introduction

The Database Group of the University of Twente initiated a new series of workshops called Twente Data Management workshops (TDM), starting with one on *XML Databases and Information Retrieval* which took place on 21 June 2004 at the University of Twente. We have set ourselves two goals for the workshop series: *i)* To provide a forum to share original ideas as well as research results on data management problems; *ii)* To bring together researchers from the database community and researchers from related research fields.

XML is becoming more and more accepted as the standard for both document-centric and data-centric information. This calls for new database systems as for instance studied in the DataX workshops (Mesiti et al. 2004); and it calls for new information retrieval systems as for instance studied in the INEX workshops (Fuhr and Lalmas 2004; see below for a short description of INEX).

We believe that for the development of such new systems, researchers have to take up an old but still largely unsolved challenge: The integration of *Databases* and *Information Retrieval*. Researchers from both fields should cooperate more than they have done in the past 30 years.

The programme committee of TDM'04 consisted of Arjen de Vries (CWI Amsterdam), Djoerd Hiemstra (University of Twente), Jaap Kamps (University of Amsterdam), Vojkan Mihajlović (University of Twente), Georgina Ramírez (CWI Amsterdam) and Börkur Sigurbjörnsson (University of Amsterdam). Each submission has been reviewed by at least one external reviewer and one of the committee members. We are proud to present a diverse and international programme with 10 contributions from 6 different countries covering 3 continents. The workshop attracted 45 participants.

Special thanks go to the Dutch research school for Information and Knowledge Systems (SIKS) for sponsoring the participation of Dutch PhD students, and to the Centre for Telematics and Information Technology (CTIT) for sponsoring the proceedings.

2 XML Information Retrieval: Achievements & Challenges

Norbert Fuhr of the University of Duisburg-Essen has been one of the few researchers who has been truly active both in database research and information retrieval research. He opened the workshop with a keynote presentation on achievements and challenges in XML information retrieval, focusing on the INEX workshops (Fuhr and Lalmas 2004). INEX (The Initiative for the Evaluation of XML Retrieval) provides an opportunity to evaluate XML information retrieval systems by means of a large-scale benchmark test collection. The test collection consists of a set of XML documents (IEEE journals), queries and so-called relevance assessments, i.e., assessments of the relevance of XML elements to the query. Many of the TDM participants have used the INEX collection to evaluate their research results.

Fuhr gave an excellent overview of approaches of various INEX participants. Most approaches are extensions of classical models and language models. New approaches and retrieval models should support vague interpretations of all types of data and conditions. As an example, consider the query for "an artist named Ulbrich living in Frankfurt, Germany about 100 years ago", and suppose the database contains the artist: "Joseph Maria Olbrich, Darmstadt, 1901". In order to match this record with some degree or probability of relevance we need phonetic similarity (Ulbrich vs. Olbrich), geographic similarity (Frankfurt is close to Darmstadt) and data similarity (1901 ≈ 1904). Another major issue in INEX is the evaluation itself: Assumptions of classical evaluation measures (e.g. documents as a separate retrieval units) are certainly invalid for XML elements. Hopefully, the new interactive track in INEX 2004 will answer some of the questions on evaluation metrics.

Jan Kuper of the University of Twente presented an invited paper on merging of incomplete, erroneous and mutually inconsistent data from different sources. The merger, that was presented at IJCAI (Kuper et al. 2003), aligns the elements of pairs of input XML documents, then groups together all alignments to combine more than two input documents, and finally applies a set of domain-specific reasoning rules. The approach was tested on information extraction results from descriptions of soccer matches, but can easily be applied to other (temporally) ordered XML data.

Jens Teubner of the University of Konstanz gave an invited talk on the implementation of XQuery. His message: standard relational algebra is (almost) all you need to execute XQuery queries (i.e. FLWOR expressions, element construction, arbitrary nesting, the whole lot!). All what is really needed apart from the standard relational algebra operators is a so-called 'row numbering' operator. This operator happens to mirror SQL:1999 OLAP ranking functionality, so nothing stands in the way of a XQuery implementation on a modern SQL-based relational database system. In fact, Teubner will present an XQuery implementation on SQL hosts at the VLDB conference this year (Grust and Teubner 2004).

The remainder of the workshop was split in two sessions, one on XML information retrieval issues and one on XML database issues.

3 XML Information Retrieval

Paul Ogilvie of Carnegie Mellon University presented new challenges for the XML retrieval community based on a question answering scenario. In this scenario, textual documents are processed using information extraction techniques (as in Kuper's example above) to build an XML knowledge base. Questions are converted into knowledge base queries. Ogilvie finished his presentation by stating that it is crucial that researchers start investigating similar scenarios which are fundamentally different from the INEX scenario.

Jovan Pehcevski of RMIT University Melbourne presented a study in which he compared the effectiveness of a standard information retrieval system (Zettair), a native XML database system (eXist), and a combination of the two on the INEX benchmark. The experiments show that the hybrid system is 1.8 times more effective than the information retrieval system alone, and 3 times more effective than the XML database system alone.

Börkur Sigurbjörnsson of the University of Amsterdam presented an approach to XML information retrieval that decomposes a query in multiple content queries, whose results are combined in a final step based on the structural constraints of the original query. Content queries are processed following a language modeling approach, which will be presented at SIGIR later this year (Kamps et al. 2004). In a thorough evaluation on the INEX collection, Sigurbjörnsson showed that it is hard to beat a baseline run that ignores all structure.

Interestingly, Pehcevski's experiments support Sigurbjörnsson's experiments: a standard information retrieval system – ignoring all structure – already gives reasonable results. Maybe it is time for another challenge? For instance retrieval from automatically constructed fact databases as suggested by Ogilvie...

Emilie Septáková of the VSB-Technical University of Ostrave, Czech Republic, closed the session by presenting an algebraic approach to the approximate tree embedding problem.

4 XML Databases and XML Query Languages

The second session focused on XML database issues. Georgina Ramírez of CWI opened this session by making a good case for the 'database approach' to XML information retrieval. She showed that join indices on path expressions together with some simple cost estimation can significantly decrease query processing time on the INEX test collection.

Whereas Ramírez focused on physical query optimisation, Maurice van Keulen of the University of Twente presented an approach to logical query optimisation of XPath queries (i.e., without a cost model) which utilises standard relational algebra. Van Keulen's message is similar to Teubner's message: relational optimisation is all you need. In fact, these results will come in handy for Teubner (see above), as his query plans are huge and provide many opportunities for optimisation.

Benjamin Piwowarski of the University Paris 6 presented a new ("yet another", as said by himself) algebra for XML information retrieval. Interestingly, the algebra is relatively simple, relying on simple operators (set operators and XPath axis step operators) and a special *about* operator for information retrieval queries. The algebra will be used in the INEX 2004 system.

Piwowarski's presentation was an excellent introduction to that of Maarten Marx of the University of Amsterdam, who started his presentation with some questions: Is one algebra really different from another algebra? Is your algebra better than his algebra? (e.g., is Piwowarski's algebra better than Fuhr's algebra?) Marx studies such questions by giving characterisations of the expressive power of XPath in terms of first order logic. In his PODS 2004 paper (Marx 2004), for which he received the best paper award, Marx presented an extended XPath version in which every first order definable set of nodes can be expressed. At TDM'04 Marx showed that XPath 1.0 *does* define an elegant and meaningful subset of this extended version: one that can define the same sets as first order logic in two variables.

5 Conclusion and Outlook

The TDM workshop on XML databases and information retrieval showed the success of 'cross-over research' where ideas from the database community and the information retrieval community are combined to develop XML retrieval systems of the future. We hope the workshop initiated new research in this direction. TDM'04 should set the stage for workshops to come. A second workshop in the TDM workshop series is planned in June 2005.

References

Fuhr, N. and M. Lalmas (2004). Report on the INEX 2003 workshop. *SIGIR Forum 38*(1).

Grust, T. and J. Teubner (2004). XQuery on SQL hosts. In *30th International Conference on Very Large Databases (VLDB)*.

Kamps, J., M. de Rijke, and B. Sigurbjörnsson (2004). Length normalisation in XML information retrieval. In *27th ACM Conference on Information Retrieval (SIGIR)*.

Kuper, J. et al. (2003). Multimedia indexing and retrieval through multi-source merging. In *18th International Joint Conference on Artificial Intelligence (IJCAI)*.

Marx, M. (2004). Conditional XPath, the first order complete XPath dialect. In *Proceedings of 23rd ACM Symposium on Principles of Database Systems (PODS)*.

Mesiti, M., et al. (2004). Report on the EDBT'04 workshop on database technologies for handling XML information on the web. *SIGMOD Record 33*(2).

Mihajlović, V. and D. Hiemstra (Eds.) (2004). Proceedings of the first Twente Data Management Workshop (TDM'04) on XML Databases and Information Retrieval. *Centre for Telematics and Information Technology*, ISSN 0929-0627. http://www.cs.utwente.nl/∼tdm