XML Full-Text Search: Challenges and Opportunities

Sihem Amer-Yahia
AT&T Labs Research
sihem@research.att.com

Jayavel Shanmugasundaram
Cornell University
jai@cs.cornell.edu

1 Motivation

An ever growing number of XML repositories are being made available for search. A lot of activity has been deployed in the past few years to query such repositories. In particular, full-text querying of text-rich XML documents has generated a wealth of issues that are being addressed by both the database (DB) and information retrieval (IR) communities. The DB community has traditionally focused on developing query languages and efficient evaluation algorithms for highly structured data. In contrast, the IR community has focused on searching unstructured data, and has developed various techniques for ranking query results and evaluating their effectiveness. Fortunately, recent trends in DB and IR research demonstrate a growing interest in adopting IR techniques in DBs and vice versa [1, 2, 3, 4, 5, 6, 7, 9].

In the past 5 years, the W3C has been putting a lot of effort in designing the XQuery 1.0 and XPath 2.0 languages that provide powerful primitives to navigate in XML documents. Many database researchers and practitioners have influenced the design of these languages and have been developing XQuery prototypes. On the other hand, in IR, INEX, the INitiative for the Evaluation of XML [8] has been created 3 years ago to put together XML documents to assess scoring and ranking methods for XML that accounts for document structure, in the same manner as TREC was designed for keyword retrieval. Several prototypes participate to INEX each year and the basic query language used within this effort is very similar to XPath.

The goal of this proposal is to provide a survey on existing research in XML full-text search in DB and IR including languages, appropriate scoring and ranking methods, implementation architectures and query evaluation algorithms and, summarize open research issues such as the joint optimization of queries on both structure and content. We believe that this tutorial is necessary to drive the attention of DB and IR researchers and practitioners to participate in solving these issues.

2 Tutorial Organization

The tutorial targets researchers in DB and IR, software and application developers, and the XML community.

It is organized in 3 parts. Each part is intended for 1 hour. The first part motivates full-text search in XML through a series of real XML documents and applications and describes challenging issues addressed by current research on XML full-text search. The second part reviews existing efforts in DB and IR including research projects and prototype architectures. The last part contains open research issues raised by the integration of structured queries and text search.

References

[1] S. Amer-Yahia, C. Botev, J. Shanmugasundaram. TeXQuery: A Full-Text Search Extension to XQuery. WWW 2004.

[2] E. W. Brown. Fast Evaluation of Structured Queries for Information Retrieval. SIGIR 1995.

[3] D. Carmel, Y. S. Maarek, M. Mandelbrod, Y. Mass, A. Soffer. Searching XML Documents via XML Fragments. SIGIR 2003.

[4] T. T. Chinenyanga, N. Kushmerick. Expressive and Efficient Ranked Querying of XML Data. WebDB 2001.

[5] S. Cohen et al. XSEarch: A Semantic Search Engine for XML. VLDB 2003.

[6] N. Fuhr, K. Grossjohann. XIRQL: An Extension of XQL for Information Retrieval. SIGIR 2000.

[7] L. Guo, F. Shao, C. Botev, J. Shanmugasundaram. XRANK: Ranked Keyword Search over XML Documents. SIGMOD 2003.

[8] Initiative for the Evaluation of XML Retrieval. *http://inex.is.informatik.uni-duisburg.de:2004/*

[9] A. Theobald, G. Weikum. The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking. EDBT 2002.

[10] H. Turtle, B. Croft. Inference Networks for Document Retrieval. SIGIR 1990.

[11] F. Weigel, H. Meuss, K. U. Schulz, F. Bry. Content and Structure in Indexing and Ranking XML. WebDB 2004.

[12] The World Wide Web Consortium. XQuery 1.0 and XPath 2.0 Full-Text. W3C Working Draft. *http://www.w3.org/TR/xquery-full-text/*

[13] The World Wide Web Consortium. XQuery 1.0 and XPath 2.0 Functions and Operators. W3C Working Draft. *http://www.w3.org/TR/xquery-operators/*

[14] C. Zhang, J. Naughton, D. DeWitt, Q. Luo, G. Lohman. On Supporting Containment Queries in Relational Database Management Systems. SIGMOD 2001.

Proceedings of the 31st VLDB Conference,
Trondheim, Norway, 2005