Assessing XML Data Management with XMark

Albrecht Schmidt[1] Florian Waas[2] Martin Kersten[1] Michael J. Carey[3]
Ioana Manolescu[4] Ralph Busse[5]

[1] CWI, Kruislaan 413, 1090 GB Amsterdam, The Netherlands, *firstname.lastname*@cwi.nl
[2] Microsoft Corporation, Redmond (WA), USA, florianw@microsoft.com
[3] BEA Systems, Inc., USA, mcarey@bea.com
[4] INRIA-Rocquencourt, 78153 Le Chesnay Cedex, France, Ioana.Manolescu@inria.fr
[5] FHG-IPSI, Dolivostr. 15, 64293 Darmstadt, Germany, busse@ipsi.fraunhofer.de

1 Motivation

We discuss some of the experiences we gathered during the development and deployment of XMark, a tool to assess the infrastructure and performance of XML Data Management Systems. Since the appearance of the first XML database prototypes in research institutions and development labs, topics like validation, performance evaluation and optimization of XML query processors have received significant interest. The XMark benchmark follows a tradition in database research and provides a framework to assess the abilities and performance of XML processing system: it helps users to see how a query component integrates into an application and how it copes with a variety of query types that are typically encountered in real-world scenarios. To this end, XMark offers an application scenario and a set of queries; each query is intended to challenge a particular aspect of the query processor like the performance of full-text search combined with structural information or joins. Furthermore, we have designed and made available a benchmark document generator that allows for efficient generation of databases of different sizes ranging from small to very large. In short, XMark attempts to cover the major aspects of XML query processing ranging from small to large document and from textual queries to data analysis and *ad hoc* queries.

2 Some Lessons Learned

During the experiments we conducted the following points of interest came up: (1) The physical XML mapping has a far-reaching influence on the complexity of query plans. Each mapping favors certain types of queries and enables efficient execution plans for them. No mapping was able to outperform the others across the board in our experiments. (2) The complexity of query plans is often aggravated by information-loss during translation from the declarative high-level query language to the low-level execution algebra. There often appears to be a semantic gap between the two – at least in the implementations we inspected. Thus, cost-based query optimizers tend to consider search spaces that are larger than necessary. Further research might lead to query algebras that reduce the gap. (3) Meta-data access can be a dominant factor in query execution especially in simple lookup queries with small result sizes. It is possible that rather complex relationships have to be extracted from the database's meta data store. (4) Schema information often enables better database schema design and is also useful in query optimization since it introduces syntactic and semantic constraints that can guide the search for a good execution plan, reduce storage requirements or enable clustering. (5) An XML query engine is usually only one part of a large system and has to integrate well with the other components.

More and complementary information especially about results obtained from running the benchmark on several platforms can be found at the places listed in the References below.

References

[1] A. Schmidt, M. Kersten, D. Florescu, M. Carey, I. Manolescu, and F. Waas. The XML Store Benchmark Project, 2000. http://www.xml-benchmark.org.

[2] A. Schmidt, F. Waas, M. Kersten, M. Carey, I. Manolescu, and R. Busse. XMark: A Benchmark for XML Data Management. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, Hong Kong, China, August 2002. to appear.

[3] A. Schmidt, F. Waas, M. Kersten, D. Florescu, I. Manolescu, M. Carey, and R. Busse. The XML Benchmark Project. Technical Report INS-R0103, CWI, Amsterdam, The Netherlands, April 2001.

Proceedings of the 28th VLDB Conference,
Hong Kong, China, 2002