REPORT on the EDBT'04 Workshop on Database Technologies for Handling XML Information on the Web

Marco Mesiti
University of Milano

Barbara Catania
University of Genova

Giovanna Guerrini
University of Pisa

Akmal Chaudhri
IBM developerWorks

The EDBT'04 Workshop on *Database Technologies for Handling XML Information on the Web* (*DataX'04*) was held in Heraklion, Crete, on Sunday 14 March, 2004, and attracted approximately 30 participants from different countries.

The *DataX'04* workshop was organized to bring together academics, users, and vendors from the XML database community, to discuss new trends in research on XML data management. The workshop benefited from experiences gained with the *XML-Based Data Management Workshop* (XMLDM) [2] held in conjunction with EDBT'02, and other workshops and conferences related to XML data management in which the DataX'04 organization and program committee members were involved.

In response to the call for papers, 35 high quality submissions were received. Each paper was carefully reviewed by at least three members of the program committee and external reviewers. As a result of this process, 11 papers were presented at the workshop, covering a large variety of topics, ranging from document indexing and querying to schema evolution management and applications. In addition, an invited talk by Prof. Stefano Ceri was presented and a panel was organized to discuss new research trends in XML data management. The most significant papers of the workshop are included in the joint EDBT'04 post-workshop proceedings published by Springer-Verlag, in the LNCS series.

The workshop was organized into four sessions, preceded by the invited talk and followed by the panel. In what follows, the most relevant ideas and issues discussed at the workshop are reported.

Invited Talk. The workshop started with the presentation of the invited paper *XML Challenges for the Database Community: Past, Present, and Future* by Daniele Braga, Alessandro Campi, and Stefano Ceri. Four years ago, the "future impact" of XML upon the database community was predicted in an EDBT invited paper [1]. In this follow-up paper, Stefano Ceri and his coauthors discuss what was accurate and what was inaccurate in that first paper, and look at the new challenges that may have an impact in the future. In retrospect, they observe that four years ago XMLSchema and XQuery were in their infancy. It was easy to foresee a leading role for XML as a data interchange standard, but they could not anticipate the role of XML in the so-called "Web services revolution" and in successful standards and protocols such as SOAP and WSDL. They predict that the "next steps" should lead to the progressive XML-ization of databases, and discuss the evolution required by XQuery (not only extensions but also simplifications) and the technological challenges posed for "native" XML repositories.

Session 1: XML metadata and evolution of XML Schemas. In the first session, two papers addressed issues related to the use of XML for representing metadata and to the evolution of the schema of XML documents.

Knowledge Management Framework for the Collaborative Distribution of Information, by Jérôme Godard, Frédéric Andrès, William Grosky, and Kinji Ono, presents a collaborative framework for managing and accessing multimedia resources. Resources are associated with XML annotations containing

metadata describing their content, which are exploited to organize and compare resources through an ontology. Environmental entities (i.e., user, community, device) are associated with XML profiles, which are used for evaluating the relevance of a resource to a user. Then services are proposed for automatically dispatching resources on devices and tailoring the view of a resource in the requesting environment.

Keeping pace with Evolving XML-based Specifications, by Marvin Tan and Angela Goh, presents an extension of XML Schema that allows one to incorporate information about changes from earlier versions. By using this extension, documents conforming to different versions of XML-based standards and specifications, most of which are still evolving and are characterized by frequent changes, can be handled.

Session 2: Indexing XML documents. In the second session, four papers were presented, focusing on indexing techniques for XML documents.

L-Tree: a Dynamic Labeling Structure for Ordered XML Data, by Yi Chen, George Mihaila, Sriram Padmanabhan, and Rajesh Bordawekar, presents the L-tree, a dynamic structure used to maintain an order-preserving labeling scheme for XML data in the presence of updates. The L-Tree helps in assigning and updating labels of data items guaranteeing a good trade-off between the number of bits used for label representation and the re-labeling costs associated with XML updates.

Implementation of XPath Axes in the Multi-dimensional Approach to Indexing XML Data, by Michal Krátký, Jaroslav Pokorný, and Václav Snášel, revises multidimensional indexing approaches for XML documents and introduces a novel approach for the implementation of an XPath subset, supporting both text queries and path expressions.

In *Dynamic Range Labeling for XML Trees*, by Takeharu Eda, Yasushi Sakurai, Toshiyuki Amagasa, Masatoshi Yoshikawa, Shunsuke Uemura, and Takashi Honishi, the problem of efficiently evaluating path expressions is addressed by labeling nodes in such a way that frequent updates to the XML tree do not require bulk node re-labeling. The most sophisticated technique proposed uses histograms to keep information concerning update operations and can manage growing data sets.

FliX: A Flexible Framework for Indexing Complex XML Document Collections, by Ralf Schenkel, addresses the relevant problem of indexing heterogeneous collections of interlinked XML documents. It describes a framework (FliX) which exploits path indexing strategies (such as Pre/Post Order scheme, HOPI, APEX) to build compact indexes on meta-documents. Using these indexes, path expressions are executed firstly on meta-documents and then are evaluated on the linked documents.

Session 3: Querying XML documents. In this session three papers were presented, dealing with different perspectives of XML query processing.

A Statistical Approach for XML Query Size Estimation, by Mong Li Lee, Hanyu Li, Wynne Hsu, and Beng Chin Ooi, presents a statistical method for estimating the result size of XML queries. The system extracts two pieces of summarized information – node ratio and node frequency – from every distinct parent-child path in XML files. Experimental results are quite impressive compared to other techniques.

Answering Queries on XML Data by means of Association Rules, by Elena Baralis, Paolo Garza, Elisa Quintarelli, and Letizia Tanca, proposes patterns corresponding to summarized representations of XML data, based on the extraction of association rules from XML datasets. Such patterns can be directly used to accelerate the query process. A graph-based formalism is introduced to represent these patterns.

Prune the XML before you Search it: XML Transformations for Query Optimization, by Stéphane Bressan, Zoé Lacroix, Ying Guang Li, and Anna Maddalena, describes an optimization technique for XQuery evaluation over XML documents. The basic idea is to project out at run-time, before query execution, parts of the documents that are not relevant to the query. To this aim, a symbolic representation of the relevant portion of the input query is created and used to reduce the document size. Experiments show that the proposed pruning approach, coupled with appropriate indexing structures, significantly reduces the query execution time.

Session 4: XML applications. In the fourth session, two papers were presented describing some applications to the protection of XML information and to multimedia data.

XML-based Revocation and Delegation in a Distributed Environment, by Konstantina E. Stoupa, Athena I. Vakali, Fang Li, and Ioannis A. Tsoukalas, proposes an XML-based distributed delegation module which can be integrated into a distributed role-based access control mechanism for protecting networks. The whole environment is XML-based and authorizations as well as delegations and revocations are expressed through DTDs.

A Transcode-and-Prefetch Method of XML Contents Containing Multiple Multimedia Data for Mobile Terminals, by Maria Hong, Dong-Yeop Ryu, and Young-Hwan Lim, presents methods to efficiently transfer multimedia information represented in XML to mobile clients under bandwidth constraints. It proposes prefetching and transcoding of information segments.

Panel: What's next in XML and databases. The concluding panel discussion was on new cutting edge research topics on XML data management.

Minos Garofalakis (Bell Labs, USA) sees the issue of XML data reduction and approximation as a promising direction for future research. Through the use of concise synopses, XML query engines can provide users with very fast, approximate results to their interactive data analysis/exploration queries. Although some initial steps have been taken, several important problems in this area remain open, including the effective handling of XML document values in the summary and designing a generic framework for characterizing XML data synopses.

Ioana Manolescu (INRIA, France) believes that research should target the XQuery language; XQuery is soon to be issued as a W3C standard. This opens up new issues on XQuery implementations: defining storage and execution models, optimization techniques, and cost models. Among the promising applications of XML data management, warehousing of XML data and Web services deserve significant attention, due to the increase of XML data on the Web.

George Mihaila (IBM Watson Research Center, USA) claims that we still lack approaches for versioning (both at schema and data levels), efficient stream processing, binary encoding of XML documents, and XML data federation. Moreover, implementations of indexing structures for XQuery and structural joins are also deemed very relevant.

Ralf Schenkel (Max Planck Institute, Germany) sees the issue of querying XML data with a vague knowledge of schema information as a relevant research topic. The query engine should thus be able to return approximate results with a relevance feedback that can take into account the structural similarity between the document and the query and the use of ontologies to identify similar concepts. This could be the way to answer queries relying on what the user means instead of what the user specifies.

Bhavani Thuraisingham (NSF, USA) sees the security of XML and RDF information in the semantic web as a leading research direction. We also need to examine the Web Rules Language to specify security policies in XML and RDF. Closely related to security is privacy. We also need to examine the specification and enforcement of privacy policies. Finally, we need to develop trust mechanisms for the semantic web. This also includes developing a trust negotiation language as well as developing mechanisms for enforcing trust. In summary, security, privacy and trust for the semantic web are important research areas.

Vasilis Vassalos (Athens University of Economics and Business, Greece) believes that XML is best suited for data integration and Web service integration. The whole family of XML standards, such as XQuery, XML Schema, RDF, and WSDL, come into play in addressing these issues. XML should also become the format to describe and natively store and process data with strong semistructured characteristics, such as (most importantly) bioinformatic data.

References.

[1] S. Ceri, *et al.* XML: Current Developments and Future Challenges for the Database Community. In Proc. of EDBT 2000, LNCS 1777, pages 3-17, 2000.

[2] A. Chaudhri *et al.* XML-Based Data Management and Multimedia Engineering - EDBT 2002 Workshops. LNCS 2490, 2002.