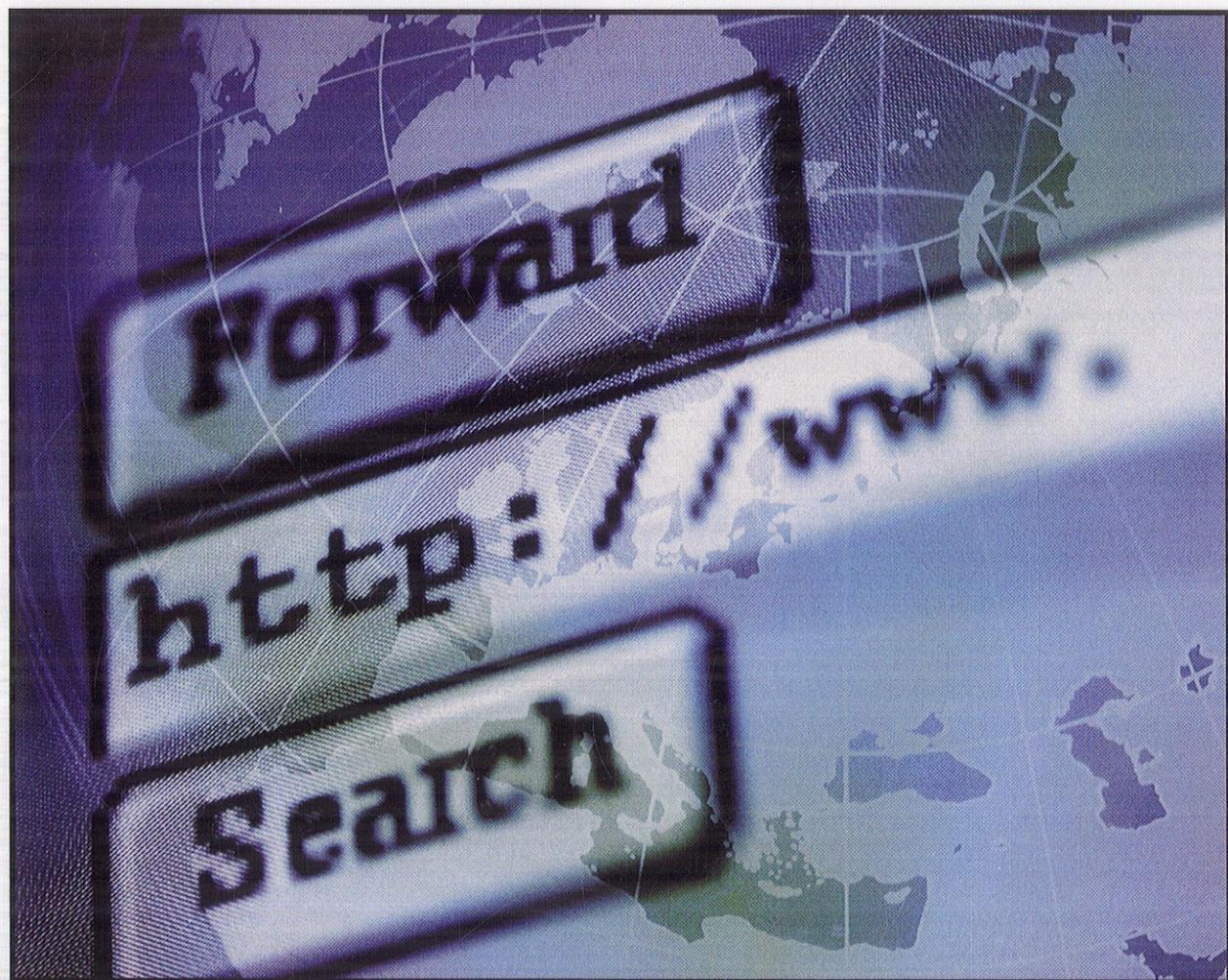# Open and Novel Issues in XML Database Applications

Future Directions and Advanced Technologies

Eric Pardede

Detailed Table of Contents

Section I
XML Repositories and Modelling

Chapter I

        Mary Ann Malloy, The MITRE Corporation, USA
        Irena Mlynkova, Charles University, Czech Republic

As XML technologies have become a standard for data representation, it is inevitable to propose and implement efficient techniques for managing XML data. A natural alternative is to exploit tools and functions offered by relational database systems. Unfortunately, this approach has many detractors, especially due to inefficiency caused by structural differences between XML data and relations. But, on the other hand, relational databases represent a mature, verified and reliable technology for managing any kind of data including XML documents. In this chapter, the authors provide an overview and classification of existing approaches to XML data management in relational databases. They view the problem from both state-of-the-practice and state-of-the-art perspectives, then describe the current best known solutions, their advantages and disadvantages. Finally, they discuss some open issues and their possible solutions.

Chapter II

        Mirella M. Moro, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
        Lipyeow Lim, IBM T.J. Watson Research Center, USA
        Yuan-Chi Chang, IBM T.J. Watson Research Center, USA

It is well known that XML has been widely adopted for its flexible and self-describing nature. However, relational data will continue to co-exist with XML for several different reasons one of which is the high cost of transferring everything to XML. In this context, data designers face the problem of modeling both relational and XML data within an integrated environment. This chapter highlights important questions on hybrid XML-relational database design and discusses use cases, requirements, and deficiencies in existing design methodologies especially in the light of data and schema evolution. The authors' analysis results in several design guidelines and a series of challenges to be addressed by future research.

Chapter III

Vassiliki Koutsonikola, Aristotle University of Thessaloniki, Greece
Athena Vakali, Aristotle University of Thessaloniki, Greece

Nowadays, XML has become the standard for representing and exchanging data over the Web and several approaches have been proposed for efficiently managing, storing, querying, and representing XML data originating from diverse and often heterogeneous sources. The Lightweight Directory Access Protocol is a promising technology for XML data storage and retrieval since it facilitates access to information organized under a variety of frameworks and applications. As an open, vendor-neutral standard, LDAP provides an extendable architecture for centralized storage and management of information that needs to be available for today's distributed systems and services. The similarities between XML and LDAP data representation have led to the idea of processing XML data within the LDAP framework. This chapter focuses on the topic of LDAP and XML integration with emphasis on the storage and retrieval approaches implemented so far. Moreover, the chapter includes an overview and survey of the theoretical background and the adopted practices as realized in the most popular and emerging frameworks which tune XML and LDAP.

Chapter IV

Giovanna Guerrini, Università degli Studi di Genova, Italy
Marco Mesiti, Università degli Studi di Milano, Italy

The large dynamicity of XML documents on the Web has created the need to adequately support structural changes and to account for the possibility of evolving and versioning the schemas describing XML document structures. This chapter discusses and compares the support for schema evolution and versioning provided by commercial systems as well as the most relevant approaches and prototypes proposed and developed by the research community.

Section II
Querying and Processing XML

Chapter V

Mingzhu Wei, Worcester Polytechnic Institute, USA
Ming Li, Worcester Polytechnic Institute, USA
Elke A. Rundensteiner, Worcester Polytechnic Institute, USA
Murali Mani, Worcester Polytechnic Institute, USA
Hong Su, Oracle Cooperation, USA

Stream applications bring the challenge of efficiently processing queries on sequentially accessible XML data streams. In this chapter, the authors study the current techniques and open challenges of XML stream processing. Firstly, they examine the input data semantics in XML streams and introduce the state-of-the-art of XML stream processing. Secondly, they compare and contrast the automaton-based and algebra-based techniques used in XML stream query execution. Thirdly, they study different optimization strategies that have been investigated for XML stream processing – in particular, we discuss cost-based optimization as well as schema-based optimization strategies. Lastly but not least, they list several key open challenges in XML stream processing.

Sven Groppe, University of Lübeck, Germany
Jinghua Groppe, University of Lübeck, Germany
Christoph Reinke, University of Lübeck, Germany
Nils Hoeller, University of Lübeck, Germany
Volker Linnemann, University of Lübeck, Germany

The widespread usage of XML in the last few years has resulted in the development of a number of XML query languages like XSLT or the later developed XQuery language. Today, there are many products like databases in the area of XML processing that support either XSLT or XQuery, but not both of them. This may prevent users from employing their favourite XML query language. In this chapter, the authors show that both languages have comparable expression power and present a scheme for translating XQuery to XSLT and vice versa. This translation scheme enables a user to employ either XSLT or XQuery for each product which supports one of these languages. They also summarize in this chapter both current and future trends and research issues and also consider those that might emerge in the common area of XSLT and XQuery and which are particular to XSLT.

Mirella M. Moro, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
Zografoula Vagena, Microsoft Research, UK
Vassilis J. Tsotras, University of California, Riverside, USA

Content-based routing is a form of data delivery whereby the flow of messages is driven by their content rather than the IP address of their destination. With the recognition of XML as the standard for data exchange, specialized XML routing services become necessary. In this chapter, the authors first demonstrate the relevance of such systems by presenting different world application scenarios where XML routing systems are needed and/or employed. Then, they present a survey of the current state-of-the-art. Lastly, they attempt to identify issues and problems that have yet to be investigated. This discussion will help identify open problems and issues and suggest directions for further research in the context of such systems.

Philippe Poulard, The French National Institute for Research in Computer Science and
Control (INRIA), France

XML engines are usually designed to solve a single class of problems: transformations of XML structures, validations of XML instances, Web publishing, and so forth. As the relevant operations or declarations are described with XML vocabularies, their respective engines have to face similar issues such as unmarshalling, and at runtime data handling. In order to address such issues, the author proposes an innovative and reliable XPath-based framework, Active Tags, that unifies XML technologies in a coherent system where various XML languages can cooperate. In our approach, we focus on its type system that enhances the XML Data Model, specifically by allowing XPath expressions to be applied on non-XML objects, and on the ability of the engine to mix declarative languages with imperative constructs. This valuable feature is illustrated with the Active Schema Language, which allows the building of dynamic-content models.

Chapter IX

   Stéphane Bressan, National University of Singapore, Singapore
   Wee Hyong Tok, National University of Singapore, Singapore
   Xue Zhao, National University of Singapore, Singapore

Since XML technologies have become a standard for data representation, a great amount of discussion has been generated by the persisting open issues and their possible solutions. In this chapter, the authors consider the design space for XML query processing techniques that can handle ad hoc and continuous XPath or XQuery queries over XML data streams. They present the state-of-art techniques in continuous and progressive XML query processing. They also discuss several open issues and future trends.

Section III
XML Personalization and Security

Chapter X

   Fabio Grandi, Università di Bologna, Italy
   Federica Mandreoli, Università di Modena e Reggio Emilia, Italy
   Riccardo Martoglia, Università di Modena e Reggio Emilia, Italy

In several application fields including legal and medical domains, XML documents are "versioned" along different dimensions of interest, whose nature depends on the application needs such as time, space and security. Specifically, temporal and semantic versioning is particularly demanding in a broad range of application domains where temporal versioning can be used to maintain histories of the underlying resources along various time dimensions, and semantic versioning can then be used to model limited applicability of resources to individual cases or contexts. The selection and reconstruction of the version(s) of interest for a user means the retrieval of those fragments of documents that match both the implicit and explicit user needs, which can be formalized as what are called personalization queries. In this chapter, the authors focus on the design and implementation issues of a personalization query processor. They consider different design options and, among them, they introduce an in-depth study of a native solution by showing, also through experimental evaluation, how some of the best performing technological solutions available today for XML data management can be successfully extended and optimally combined in order to support personalization queries.

Chapter XI

Tran Khanh Dang, National University of Ho Chi Minh City, Vietnam

In an outsourced XML database service model, organizations rely upon the premises of external service providers for the storage and retrieval management of their XML data. Since, typically, service providers are not fully trusted, this model introduces numerous interesting research challenges. Among them, the most crucial security research questions relate to data confidentiality, user and data privacy, query assurance, secure auditing, and secure and efficient storage model. Although there exists a large number of related research works on these topics, the authors are still at the initial stage and the research results are still far from practical maturity. In this chapter, they extensively discuss all potential security issues mentioned above and the existing solutions, and present open research issues relevant to security requirements in outsourced XML databases.

Section IV
XML for Advanced Applications

Chapter XII

Marco Mesiti, DICO, Università di Milano, Italy
Ernesto Jiménez Ruiz, DLSI, Universitat Jaume I, Spain
Ismael Sanz, DICC, Universitat Jaume I, Spain
Rafael Berlanga Llavori, DLSI, Universitat Jaume I, Spain
Giorgio Valentini, DSI, Università di Milano, Italy
Paolo Perlasca, DICO, Università di Milano, Italy
David Manset, Maat Gknowledge, Valencia, Spain

There is a proliferation of research and industrial organizations that produce sources of huge amounts of biological data issuing from experimentation with biological systems. In order to make these heterogeneous data sources easy to use, several efforts at data integration are currently being undertaken based mainly on XML. Starting from a discussion of the main biological data types and system interactions that need to be represented, the authors discuss the main approaches proposed for their modelling through XML. Then, they show the current efforts in biological data integration and how an increasing amount of semantic information is required in terms of vocabulary control and ontologies. Finally, future research directions in biological data integration are discussed.

Chapter XIII

Doulkifli Boukraa, University of Jijel, Algeria
Riadh Ben Messaoud, University of Nabeul, Tunisia
Omar Boussaid, University Lumière Lyon 2, France

Current data warehouses deal for the most part with numerical data. However, decision makers need to analyze data presented in all formats which we qualify as complex data. Warehousing Complex data is a new challenge for the scientific community. Indeed, it requires revisiting the whole warehousing process

in order to take into account the complex structure of data; therefore, many concepts of data warehousing will need to be redefined. In particular, modeling complex data in a unique format for analysis purposes is a challenge. In this chapter, the authors present a complex data warehouse model at both conceptual and logical levels. They show how XML is suitable for capturing the main concepts of our model, and present the main issues related to these data warehouses.

Section V
XML Benchmarking

Since XML technologies have become a standard for data representation, numerous methods for processing XML data emerge every day. Consequently, it is necessary to compare the newly proposed methods with the existing ones, as well as analyze the effect of a particular method when applied to various types of data. In this chapter, the authors provide an overview of existing approaches to XML benchmarking from the perspective of various applications and they show that to date the problem has been highly marginalized. Therefore, in the second part of the chapter they discuss persisting open issues and their possible solutions.