Lecture Notes in Computer Science 1997
Edited by G. Goos, J. Hartmanis and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Dan Suciu Gottfried Vossen (Eds.)

The World Wide Web and Databases

Third International Workshop WebDB 2000
Dallas, TX, USA, May 18-19, 2000
Selected Papers

Springer

Volume Editors

Dan Suciu
University of Washington, Computer Science and Engineering
Seattle, WA 98195-2350, USA
E-mail: suciu@cs.washington.edu

Gottfried Vossen
Universität Münster, Wirtschaftsinformatik
Steinfurter Str. 109, 48149 Münster, Germany
E-mail: vossen@helios.uni-muenster.de

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

The World Wide Web and databases : selected papers / Third
International Workshop WebDB 2000, Dallas, TX, May 18 - 19, 2000. Dan
Suciu ; Gottfried Vossen (ed.). - Berlin ; Heidelberg ; New York ;
Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo :
Springer, 2001
(Lecture notes in computer science ; Vol. 1997)
ISBN 3-540-41826-1

CR Subject Classification (1998): H.5, C.2.4, H.4.3, H.2, H.3

ISSN 0302-9743
ISBN 3-540-41826-1 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by PTP Berlin, Stefan Sossna
Printed on acid-free paper SPIN 10782175 06/3142 5 4 3 2 1 0

Preface

With the development of the World-Wide Web, data management problems have branched out from the traditional framework in which tabular data is processed under the strict control of an application, and address today the rich variety of information that is found on the Web, considering a variety of flexible environments under which such data can be searched, classified, and processed. Database systems are coming forward today in a new role as the primary backend for the information provided on the Web. Most of today's Web accesses trigger some form of content generation from a database, while electronic commerce often triggers intensive DBMS-based applications. The research community has begun to revise data models, query languages, data integration techniques, indexes, query processing algorithms, and transaction concepts in order to cope with the characteristics and scale of the data on the Web. New problems have been identified, among them goal-oriented information gathering, management of semi-structured data, or database-style query languages for Web data, to name just a few. The *International Workshop on the Web and Databases* (WebDB) is a series of workshops intended to bring together researchers interested in the interaction between databases and the Web. This year's WebDB 2000 was the third in the series, and was held in Dallas, Texas, in conjunction with the ACM SIGMOD International Conference on Management of Data. After receiving a record number of paper submissions (69) the program committee accepted twenty papers to be presented during the workshop, in addition to an invited paper. The workshop attracted over 160 participants.

This volume contains a selection of the papers presented during the workshop, including the invited contribution. All papers contained herein have been further expanded by their authors and have undergone a final round of reviewing.

As could be seen from the workshop's papers, as well as from the selection of papers included in this volume, many researchers today relate their work to XML. Indeed, XML, an industry standard, offers a unifying format for the rich variety of data being shared today on the Web. There is tremendous potential here for the data management community. For the past years, research in avant-guard fields like semi-structured data, data integration, text databases, and combining database with relevance searches, applied only to very narrow settings, often hampering researchers' efforts to branch out of the traditional tabular-data processing framework. Today, XML makes these research areas not only relevant, but even imperative, opening the door for a dramatic impact. For researchers in data management, XML is seen as a linguistic framework which can express both data and meta data, and that can be stored as well as queried in a way that is familiar to a DBMS. This current interest was manifested in the workshop program through sessions on caching, querying, structuring and versioning, schema issues, and query processing, all centering around XML in one way or another.

The invited contribution is by Don Chamberlin (IBM ARC), one of the "fathers" of SQL, Jonathan Robie (Software AG USA), and Daniela Florescu (INRIA) on *Quilt: An XML Query Language for Heterogeneous Data Sources*. Quilt is a recent proposal for a query language that operates on collections of XML documents, and that searches them in a style that is familiar to the database user. It grew out of earlier proposals such XML-QL, XPath, and XQL, and combines features found there with properties of languages such as SQL and OQL. In particular, Quilt is a *functional* language whose main syntactical construct is the FLWR ("flower") expression which can bind variables in a *For* as well as a *Let* clause, then apply a predicate in a *Where* clause, and finally construct a result in a *Return* clause.

The section on *Information Gathering* has three contributions: In *Theme-Based Retrieval of Web News*, Nuno Maria and Mario J. Silva (Univ. Lisboa, Portugal) study the problem of populating a complex database of Web news with articles retrieved from heterogeneous Web sources. In *Using Metadata to Enhance a Web Information Gathering System*, Neel Sundaresan (IBM ARC), Jeonghee Yi (UCLA), and Anita Huang (IBM ARC) present the Grand Central Station (GCS) Web gathering system that enables users to find information regardless of location and format. GCS is composed of crawlers and summarizers, the former of which collect data, while the latter do content summarization in RDF, XML, or some custom format. In *Architecting a Network Query Engine for Producing Partial Results*, Jayavel Shanmugasundaram (Univ. Wisconsin and IBM ARC), Kristin Tufte (OGI), David DeWitt (Univ. Wisconsin), Jeffrey Naughton (Univ. Wisconsin), and Dave Maier (OGI) look at a new way of computing query results, namely by basing the processing on an initial part of the input instead of a complete input, which may be expensive to wait for on the Web.

The next section concentrates on techniques for *Caching* Web pages or views to speed up the handling of future requests. Luping Quan, Li Chen, and Elke A. Rundensteiner (Worcester Polytech) present *Argos: Efficient Refresh in an XQL-Based Web Caching System*. Qiong Luo, Jeffrey Naughton, Rajesekar Krishnamurthy, Pei Cao, and Yunrui Li (Univ. Wisconsin) study *Active Query Caching for Database Web Servers* as well as techniques for answering at a proxy server.

The first section devoted to XML is on *Querying XML*. Anja Theobald and Gerhard Weikum (Univ. Saarland, Germany) argue that XML query languages proposed so far are inadequate for Web searching since they do *Boolean* retrieval only and vastly ignore semantic relationships of data. They suggest *Adding Relevance to XML* by combining XML querying with an information retrieval search engine that has ontological knowledge. In *Evaluating Queries on Structure with eXtended Access Support Relations*, Thorsten Fiebig and Guido Moerkotte (Univ. Mannheim, Germany) present a scalable index structure that supports queries over the structure of XML documents. Next, Albrecht Schmidt, Martin Kersten, Menzo Windhouwer, and Florian Waas (CWI) present a data and an execution model for *Efficient Relational Storage and Retrieval of XML Documents*.

The second XML section is on *XML Structuring and Versioning*. It is started by Meike Klettke and Holger Meyer (Univ. Rostock, Germany) with *XML and Object-Relational Database Systems — Enhancing Structural Mappings Based on Statistics*. Arnaud Sahuguet (Univ. Pennsylvania), following a well-known movie title, discusses *Everything You Ever Wanted to Know About DTDs, but Were Afraid to Ask*. He explores how XML DTDs are being used today for specifying document structure and how and why they are abused. One of his findings is that most DTDs are incorrect, as they seem to be used more for documentation than for validation; moreover, many of the syntactic features of XML are not used in current DTDs. Finally several replacement candidates are discussed, such as XML Schemas, Schematron, and XDuce. The section concludes with *Version Management of XML Documents* by Shu-Yao Chien (UCLA), Vassilis Tsotras (UC Riverside), and Carlo Zaniolo (UCLA).

In the section entitled *Web Modeling*, Aldo Bongio, Stefano Ceri, Piero Fraternali, and Andrea Maurino (Politecnico di Milano, Italy) report on *Modeling Data Entry and Operations in WebML*. WebML, the Web Modeling Language, is an XML-based language for the conceptual and visual specification of Web sites that comes with a variety of design tools.

The next topic area to be studied is *Query Processing*. Gösta Grahne and Alex Thomo (Concordia University) present *An Optimization Technique for Answering Regular Path Queries* that does query rewriting in the context of semistructured data. Haruo Hosoya and Benjamin C. Pierce (University of Pennsylvania) present a preliminary report on *XDuce: A Typed XML Processing Language*. XDuce is a statically typed functional programming language for tree transformations and hence XML processing, which guarantees that programs never crash at run-time, and that resulting values always conform to specified types.

The final area is *Classification and Retrieval*. Panagiotis G. Ipeirotis, Luis Gravano (Columbia University), and Mehran Sahami (E.piphany, Inc.) discuss *Automatic Classification of Text Databases Through Query Probing*. David W. Embley and L. Xu (Brigham Young University) present *Record Location and Reconfiguration in Unstructured Multiple-Record Web Documents*, where the objective is to convert unstructured Web documents into structured database tables. The major technique employed for record location is a record recognition measure that is based on vector space modeling.

As can be seen from the above, WebDB 2000 covered a variety of topics and gave good insight into current research projects that are carried out at the intersection of databases and the Web. It clearly showed the rapidly increasing interest in issues related to Internet databases, and to applying database techniques to the Web; it also put the current XML hype somewhat into perspective.

We are particularly grateful to the members of our program committee, who had to do a lot of reading within very short time, and to Maggie Dunham, Leonidas Fegaras, Alex Delis, and the SIGMOD 2000 staff for doing almost all of the financial transactions as well as the local organization for us.

November 2000 Dan Suciu and Gottfried Vossen

Organization

WebDB 2000 took place for the third time, on May 18 and 19, 2000, at the Adam's Mark Hotel in Dallas, Texas, as in the previous year right after the ACM PODS/SIGMOD conferences.

Program Committee

Program Co-chairs:	Dan Suciu (University of Washington, USA)
	Gottfried Vossen (University of Münster, Germany)
Members:	Peter Buneman (University of Pennsylvania, USA)
	Stefano Ceri (Politecnico di Milano, Italy)
	Daniela Florescu (INRIA, France)
	Juliana Freire (Bell Labs, USA)
	Zoe Lacroix (Genelogic, USA)
	Laks Lakshmanan (Concordia University, Canada)
	Alon Levy (University of Washington, USA)
	Bertram Ludäscher (San Diego Supercomputer Center, USA)
	Gianni Mecca (Universita di Roma Tre, Italy)
	Renee Miller (University of Toronto, Canada)
	Guido Moerkotte (University of Mannheim, Germany)
	Frank Neven (Limburgs Universitaire Centrum, Belgium)
	Werner Nutt (DFKI Germany)
	Yannis Papakonstantionou (University of California, San Diego, USA)
	Louiqa Raschid (University of Maryland, USA)
	Shiva Shivakumar (Stanford University, USA)

Sponsoring Institutions

AT&T Labs Research, Florham Park, New Jersey, USA
Westfälische Wilhelms-Universität Münster, Germany

Table of Contents

Invited Contribution

Information Gathering

Caching

Querying XML

XML Structuring and Versioning

Web Modeling

Query Processing

Classification and Retrieval