Report on WebDB'2000:

3rd International Workshop on the Web and Databases

D. Suciu
AT&T Labs Research, USA
suciu@research.att.com

G. Vossen
University of Münster, Germany
vossen@helios.uni-muenster.de

June 2000

1 Introduction

With the advent of the World-Wide Web, data management has branched out from a traditional relational-server framework to deal with the variety of information available on the WWW. At the same time, database systems have come forward in a new role as the primary source 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 (e.g., searching a product catalog, generating entries in an order database, enabling customer relationship management or user profiling). 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 WebDB (*International Workshop on the Web and Databases*) series of workshops is intended to bring together researchers interested in the interaction between databases and the Web.

This year, WebDB 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. The workshop drew tremendous attention: The number of submissions ended up to be almost 80 papers, 20 of which were accepted for presentation. Attendance almost doubled from last year to a record-high of around 160 participants, many of which stayed until the very last session on Friday afternoon. The program was put together by an international program committee consisting of Peter Buneman (University of Pennsylvania), Stefano Ceri (Politechnico di Milano), Daniela Florescu (INRIA), Juliana Freire (Bell Labs), Zoe Lacroix (Genelogic), Laks Lakshmanan (Concordia University), Alon Levy (University of Washington), Bertram Ludäscher (San Diego Supercomputer Center), Gianni Mecca (Universita di Roma Tre),

Renee Miller (University of Toronto), Guido Moerkotte (University of Mannheim), Frank Neven (Limburgs Universitaire Centrum), Werner Nutt (DFKI Germany), Yannis Papakonstantionou (University of California, San Diego), Louiqa Raschid (University of Maryland), and Shiva Shivakumar (Stanford University).

2 Topics Covered

As could be seen from the submissions as well as from the papers finally accepted, a major topic in the database area right now, both a theoretical as well as from a practical point of view, is XML. Indeed, XML is seen as answer to the question of how to handle semi-structured data coming from the Web, how to organize large collections of data collected from different sources, and how to exchange data between various data sites. As a consequence, XML is more and more seen as a linguistic framework which can express data type and schema information, yet that can also 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 talk was given by Don Chamberlin (IBM Almaden Research Center), one of the "fathers" of SQL, on *Quilt: An XML Query Language for Heterogeneous Data Sources*. Quilt is a very 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 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.

Besides the invited talk, the workshop program consisted of 7 more sessions. It started out with a session on *Information Gathering* that saw three presentations: In *Theme-based Retrieval of Web News*, Nuno Maria and Mario J. Silva (University of Lisboa, Portugal) studied 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 Almaden Research Center), Jeonghee Yi (University of California, Los Angeles), and Anita Huang (IBM Almaden Research Center) presented 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 (University of Wisconsin and IBM Almaden Research Center), Kristin Tufte (OGI), David DeWitt (University of Wisconsin), Jeffrey Naughton (University of Wisconsin), and Dave Maier (OGI) looked 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 session on *Caching* concentrated on techniques to cache Web pages or views for speeding up the handling of future requests. Luping Quan, Li Chen, and Elke A.

Rudensteiner (Worcester Polytech) presented *Argos: Efficient Refresh in an XQL-Based Web Caching System.* Qiong Luo, Jeffrey F. Naughton, Rajesekar Krishnamurthy, Pei Cao, and Yunrui Li (University of Wisconsin) studied *Active Query Caching for Database Web Servers* as well as techniques for answering at a proxy server.

Session 3 was the first devoted to XML, here to *Querying XML.* Anja Theobald and Gerhard Weikum (University of the Saarland, Germany) argued that XML query languages proposed so far are inadequate for Web searching since they vastly ignore semantic relationships of data and suggested *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 (University of Mannheim, Germany) presented a scalable index structure that supports queries over the structure of XML documents. Next, Albrecht Schmidt, Martin Kersten, Menzo Windhouwer, and Florian Waas (CWI, The Netherlands) presented a data and an execution model for *Efficient Relational Storage and Retrieval of XML Documents.*

On Friday morning, the invited talk was followed by a session on *XML Structuring and Versioning.* It was opened by Meike Klettke and Holger Meyer (University of Rostock, Germany) with a talk on *XML and Object-Relational Database Systems — Enhancing Structural Mappings Based on Statistics.* That was followed by a highly vivid presentation by Arnaud Sahuguet (University of Pennsylvania) on *Everything You Ever Wanted to Know About DTDs, but Were Afraid to Ask.* He explored how XML DTDs are being used today for specifying document structure and how and why they are abused. One of his findings was 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 were discussed, such as XML Schemas, Schematron, and XDuce. The session concluded by a talk on *Version Management of XML Documents* by Shu-Yao Chien (University of California, Los Angeles), Vassilis Tsotras (University of California, Riverside), and Carlo Zaniolo (University of California, Los Angeles).

Session 6 on *Schema Issues* was opened by AnHai Doan, Pedro Domingos, and Alon Y. Levy (University of Washington) with a talk on *Learning Source Descriptions for Data Integration.* They considered the problem of automating the task of mapping between source schemas and mediated schema in data integration. Next, Aldo Bongio, Stefano Ceri, Piero Fraternali, and Andrea Maurino (Politecnico di Milano, Italy) reported 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. George A. Mihaila (University of Toronto), Louiqa Raschid (University of Maryland), and Maria-Esther Vidal (University of Maryland and University Simon Bolivar, Venezuela) discussed *Using Quality of Data Metadata for Source Selection and Ranking* in the WebSemantics System they have developed. This system can describe data repositories in WS-XML documents, publish them on the Web, locate repositories using WSQL queries, and provide transparent access to types, domains, data, and metadata.

In Session 7 on *Query Processing*, Gösta Grahne and Alex Thomo (Concordia University) presented *An Optimization Technique for Answering Regular Path Queries* that

does query rewriting in the context of semistructured data. *XPERANTO: A Middleware for Publishing Object-Relational Data as XML Documents*, by Michael Carey, Daniela Florescu, Zachary Ives, Ying Lu, Jayavel Shanmugasundaram, Eugene Shekita, and Subbu Subramanian (IBM Almaden Research Center), is an architecture that provides queryable XML views over OR databases. To this end, query processing comprises parsing and rewriting based on a query graph model that extends known techniques, and translation into SQL; a relational result undergoes XML tagging to turn into an end result. Haruo Hosoya and Benjamin C. Pierce (University of Pennsylvania) presented 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 session on *Classification and Retrieval* saw three more talks: Panagiotis G. Ipeirotis, Luis Gravano (Columbia University), and Mehran Sahami (E.piphany, Inc.) discussed *Automatic Classification of Text Databases Through Query Probing*. David W. Embley and L. Xu (Brigham Young University) talked on *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. Finally, Taher H. Haveliwala, Aristides Gionis, and Piotr Indyk (Stanford University) presented *Scalable Techniques for Clustering the Web*. The goals of this project are to generate a fine-grained Web clustering based on related topics and similarity search, where the underlying notion of similarity scales to high volumes of documents.

3 Conclusions and Outlook

In conclusion, WebDB 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.

There will be a post-workshop proceedings published in Springer-Verlag's LNCS series which will contain most of the contribution in more polished form; hopefully, that volume will be out by the end of the year. In the meantime, check out
`http://dbms.uni-muenster.de/events/webdb2000`
or
`http://www.research.att.com/conf/webdb2000`
for further information, in particular for online versions of the papers presented at the workshop. Next year, a follow-up workshop will be organized by Gianni Mecca (Universita di Roma Tre) and Jerome Simeon (Bell Labs).