Report on the First International Workshop on Efficient Web-based Information Systems

Zoé Lacroix
Arizona State University
PO Box 876106
Tempe AZ 85287-6106
USA
zoe.lacroix@asu.edu

Omar Boucelma
LSIS - Marseille
39, rue Joliot-Curie
13453 Marseille Cedex 13
France
omar.boucelma@cmi.univ-mrs.fr

1 Introduction

Many approaches have been developed in the past years to address efficient query processing for information systems, integrated systems, and Web-based systems. However, new techniques need to be developed to optimize query processing for Web-based information systems. These new techniques need to address several challenges including management and publication of semi-structured data, limited access to Web data sources, lack of published information about their content (statistics, metadata, etc.), complex coverage, etc.

The First International Workshop on Efficient Web-based Information Systems (EWIS) was held in Montpellier, France, in September 2002, in conjunction with the Eighth International Conference on Object-Oriented Information Systems. Its first occurrence attracted contributors and attendants from Europe, America, and Asia. The technical program included an invited talk, the presentation of seven papers and a final discussion involving all workshop participants. Presented papers were reviewed and published in the OOIS workshop proceedings published by Springer-Verlag in the Lecture Notes in Computer Science series (number 2426).

This first occurrence of EWIS mainly focused on semantics and navigation. The need for semantics was raised for data and tool integration. Indeed, the access to a Web-based information system is often limited for various reasons. First, Web data are semi-structured and need to be extracted from numerous sources, documents, or display features such as auxiliary text and banners. Also, the access to data is often complex and limited (APIs to various capabilities rather than a query language). Finally, each resource is autonomous and in constant and uncontrollable evolution (changes of the data themselves, their representation, the source capabilities, etc.)

The contributions to the workshop presented techniques to better access, manipulate, integrate and navigate Web data.

2 Semantic integration based on a mediated schema: report on two different approaches implemented in PICSEL and Xyleme

As a prelude to the workshop report, we include a summary provided by Marie-Christine Rousset.

In recent years, considerable research has been done about semantic integration both for structured and semi-structured data sources, and several information integration systems have been implemented. Most of them are based on the specification of a single *mediated schema* providing the semantics of a domain of interest, and on a set of *source descriptions* expressing the semantic mapping between the mediated schema and the schemas of the information sources that are related to the same domain of interest.

PICSEL [1] and Xyleme [2] are two recent information integration systems which are illustrative of two radically different choices concerning the expressivity of the mediated schema.

In PICSEL, it has been chosen to offer a formalism combining the expressive power of rules and classes for designing a rich mediated schema, thus enabling a fine-grained description of the contents of the sources. This approach is appropriate to build information integration systems related to application domains for which there exists several information sources containing closely related data. In such a setting, it is of primary importance to model the fine-grained differences between contents of sources to be able to answer precise queries in an efficient way. PICSEL has been used for building a mediator in the tourism domain in collaboration with France Telecom R&D and the Web travel agent Degriftour.[1] This travel agent makes available three databases on-line that contain different types of tourist products (flights, tours, stays in different places), each with its own specificities. For example, the *BonjourFrance* database offers a large variety of tourist products, all of them are located in France. The so-called *Degriftour* database offers flights, accommodations, and tours for many destinations all over the world. However, all it offers correspond to a departure date which is within the next two weeks. The *Reductour* database provides rather similar products but with a less strong constraint on the departure date (within eight months). Other differences exist between the contents of those three databases. For instance, *BonjourFrance* can provide rooms in Bed & Breakfast in addition to hotel rooms, while the only accommodations that are proposed in the two other sources are hotels, located exclusively out of France for *Degriftour*, and located exclusively in Europe for *Reductour*.

In Xyleme, the choice of a simple tree structure for mediated schemas has been guided by the goal of providing a very wide-area integration of XML sources that could scale up to the Web. The system architecture and the design choices have been motivated by "Web search engine"-like performance requirements, i.e. supporting many simultaneous queries over a Web-scale XML repository. Xyleme is based on a simple data model with data trees and tree types, and a simple query language based on tree queries with boolean conditions. The main components of the data model are a mediated schema, modeled by an abstract tree type, as a view of a set of tree types associated with actual data trees, called concrete tree types, and a mapping expressing the connection between the mediated

[1] See http://www.degriftour.fr/.

schema and the concrete tree types. The simplicity of the mapping relation (correspondences between tree paths) eases automatic mapping generation and distributed storage. The query language is intended to enable end-users to express simple Query-by-Example tree queries over the mediated schema. Xyleme optimizes such tree queries to ensure efficient response time.

3 Semantic data integration

Semantics can be used to facilitate the integration of Web resources and to improve query execution. The approaches presented in this section address both aspects.

The invited paper by Beneventano et al. presented query optimization techniques for semantic integration of heterogeneous structured and semi-structured sources. The described techniques are implemented in the ARTEMIS/MOMIS system that uses the formalization and reasoning capabilities of Description Logics. The system exploits Description Logics for both semantic integration and query rewriting for optimization. Semantic integration is performed by a semi-automatic extraction of inter-schema properties that express the relationships between the conceptual classes at the conceptual level, followed by the derivation of the schema mappings and inter-schema knowledge. The latter constitutes the semantic knowledge exploited by the system at query execution.

The paper by Hemnani and Bressan focused on techniques to combine syntactic and semantic knowledge for the automatic extraction of Web data. Their experimental results show that both recall and precision are improved when both approaches are combined, whereas efficiency is not affected.

In addition to the published papers, Fèlix Saltor (University of Catalonia, Spain) spontaneously accepted to present the BarceLona Object Oriented Model and Methodology (BLOOM).[2] BLOOM was motivated by the growing need of integration and cooperation between independent, autonomous, heterogeneous and distributed databases. The system is currently evolving from a tightly coupled database federation into a system that enables more loose integration of resources.

The final discussion of the workshop addressed issues related to semantic data integration. It is an area that received growing attention in the past few years. The Semantic Web activity of the World Wide Web Consortium (W3C) is now leading this effort.

4 Metadata for tool integration

For many applications it is critical not only to integrate data but also to exploit the various source capabilities available on the Web. Three papers addressed issues related to integrate efficiently source capabilities.

The paper by Katchaounov, Risch and Zürcher presented ORWISE, a system that uses an object-relational wrapper (ORW) mediator system to integrate multiple Internet search engines (ISE). ORWISE addresses the issues specific to internet search engines and Web resources in general, such as their flexibility in data representation (semi-structured), their various capabilities, their autonomy, and constant evolution. Both the source capabilities and data are modeled and exploited by the mediator, that allows transparent queries to integrated sources

[2] More information on the BLOOM project can be found at http://www-lsi.upc.es/bloom/.

with different capabilities and structure.

The paper by Synodinos and Avgeriou presented an XML-based multi-tier model named WOnDA (Write Once, Deliver Anywhere) that supports the creation and deployment of hypermedia applications.

An alternative use of semantics was proposed by Zaïane and Strilets with an approach to support users when submitting queries to search engines. They use similarity measures to compare queries and automatically generate similar queries that users can use to refine their queries.

The concluding discussion acknowledged the numerous efforts to provide homogeneous representation of applications and query capabilities. It was noted that the agent community and the database community could both contribute. Standards are being designed to facilitate the integration of resources. For the specific integration of search engines, a participant introduced the Z39.50 Information Retrieval Standard.[3] Another participant discussed the Web Feature Server (WFS) specifications in the context of integration of geographical information systems.

5 Efficient navigation

XPath is the language for Web navigation designed by the World Wide Web consortium. The languages XSLT and XQuery respectively designed to manipulate XML stylesheets, and XML data and documents both extend XPath. The ability to handle efficiently XPath queries is therefore critical. The existing benchmarks for XML management systems such as X007,[4] XMach,[5] and X-Mark[6] do not address, in general, issues specific to navigation. However, a Web query is likely to contain an XPath subquery. Two papers presented approaches to optimizing XPath queries.

The paper by He and Dyreson presents an optimization technique for XPath called warp-edge optimization. Warp edges are shortcuts into the XML tree. They represent edges such as from an element to a sibling, or to a grandchild that are computed and stored during query evaluation. The exploitation of these pre-computed structures significantly enhances the evaluation of XPath queries.

The paper by Turau presents a caching system for database driven Web sites using XML and XSLT that uses Java objects to increase scalability and to reduce the load on the backend systems storing the data.

The final discussion addressed many aspects related to efficient XPath processing. The idea of a benchmark for XPath was discussed. Such benchmark could be used in the context of both XSLT and XQuery.

6 Conclusion

The first occurrence of the workshop was a success. It received a lot of attention from researchers who contacted the organizers to share their interest and make sure that a second workshop will be organized in 2003. The 2002 occurrence of EWIS focused mainly on the need to represent and exploit semantics in various

[3] More information about NISO Z39.50-1995 - Information Retrieval (Z39.50) can be found at http://www.ifla.org/VI/5/op/udtop3/udtop3.htm.

[4] See http://www.comp.nus.edu.sg/ ebh/X007.

[5] See http://dbs.uni-leipzig.de/en/projekte/XML/XmlBenchmarking.html.

[6] See http://monetdb.cwi.nl/xml/.

ways. Another focus resided in the need to execute efficiently Web query languages, in particular XPath, XSLT, and XQuery. The workshop concluded in a French bistro in Montpellier over a copious dinner in a friendly atmosphere.

Acknowledgments We wish to thank the organizers of OOIS 2002, in particular, Jean-Pierre Bruel, as well as all the members of our Program committee listed below.

References

[1] F. Goasdoué, V. Lattès, and M.-C. Rousset. The use of CARIN language and algorithms for information integration: The PICSEL system. *International Journal of Cooperative Information Systems (IJCIS)*, 9(4):383–401, 2000.

[2] L. Xyleme. A dynamic warehouse for xml data of the web. *IEEE Data Engineering Bulletin*, 24(2):40–47, June 2001.