The Role of Structural Summaries for XML Retrieval

Mirella M. Moro[1], Zografoula Vagena[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre, RS, Brazil

[2] IBM Almaden Research Center
San Jose, CA, USA

mirella@inf.ufrgs.br, zovagena@us.ibm.com

1. Description and Objectives

A Structural Summary of an XML document is a dynamically generated and maintained graph structure that preserves the structural characteristics of the document in a compact form. The versatility of structural summaries has been established with their extensive usage for diverse retrieval tasks. Within traditional XML query processing those structures have been used as primary indexes on the structure, as well as for (a) structure discovery, (b) query formulation, rewrite and optimization, (c) storage of statistics, and other important metadata information. At the same time, structural summaries have appeared within other XML retrieval scenarios including (a) XML keyword search, (b) information discovery within P2P systems, and (c) message routing within publish/subscribe systems. This tutorial introduces the concept of XML Structural Summaries and describes their role within XML retrieval. It covers the usage of those summaries for Database-style query processing, as well as Information Retrieval-style search tasks in the context of both centralized and distributed environments. Finally, it concludes with a presentation of new retrieval scenarios that can potentially be favorably supported by those summaries.

2. Content and Outline

- Review of XML Basic Concepts: definition of XML documents and XML query processing; research efforts on efficient XML query processing.
- Structural Summaries: main techniques for processing Structural Summaries (DataGuides [1], A(k)-indexes [2] and Suffix Trees [3]); how to employ them as access methods.
- Structural Summaries within XML Query Processing: different techniques for using structural summaries while processing XML queries; limits of structural summaries (cases where the structural summaries provide false positives; extents size [4]).
- Structural Summaries within Other Retrieval Contexts: use of structural summaries in a stream processing application (publish-subscribe systems) [5,6].
- Current Work and Future Directions: new XML retrieval scenarios (such as web services, digital libraries and mashups) and hints on the potential usage of Structural Summaries to address the unique requirements of those applications.

3. Prerequisites

None.

4. Authors' Bibliography

Mirella M. Moro holds a Ph.D. in Computer Science (University of California Riverside - UCR) and has Masters and Bachelors in Computer Science as well (Universidade Federal do Rio Grande do Sul - UFRGS, Brazil). She is currently working as a researcher collaborator at Instituto de Informática – UFRGS, and her research areas of interest include XML query optimization, stream processing, content-based dissemination systems and digital libraries.

Zografoula Vagena holds a Ph.D. in Computer Science (University of California, Riverside) and has a Bachelors in Electrical Engineering and Computer Science as well (School of Electrical and Computer Engineering, NTU Athens, Greece). Her research interests include Databases and IR; Text Indexing and Retrieval, Query Processing and Optimization, XML Data Management, Document Management, Preference-Based Query Processing, Index Structures, Version Management, Transaction Time Databases, Time series. Currently, she is a Post-Doctoral Researcher at IBM Almaden Research Center (San Jose, Ca, USA) and is working on the cooperation of IR and DB-style indexing and query processing methods.

5. Resources

Additional material is available at the presenter's webpage: http://www.inf.ufrgs.br/~mirella

References

[1] Roy Goldman and Jennifer Widom. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In Proc. of VLDB, 1997.

[2] Raghav Kaushik, Pradeep Shenoy, Philip Bohannon, and Ehud Gudes. Exploiting Local Similarity for Indexing Paths in Graph-Structured Data. In Proc. of ICDE, 2002.

[3] Edward M. McCreight. A Space-Economical Suffix Tree Construction Algorithm. Journal of ACM, 23(2):262–272, 1976.

[4] Mirella M. Moro, Zografoula Vagena, and Vassilis J. Tsotras. Evaluating Structural Summaries as Access Methods for XML. In Proc. of WWW, 2006.

[5] Zografoula Vagena, Mirella M. Moro, and Vassilis J. Tsotras. RoxSum: Leveraging Data Aggregation and Batch Processing for XML Routing. In Proc. of ICDE, 2007.

[6] Zografoula Vagena, Mirella M. Moro, and Vassilis J. Tsotras. Value-Aware RoXSum: Effective Message Aggregation for XML-Aware Information Dissemination. In Proc. of WebDB, 2007.