Querying and Exchanging XML and RDF on the Web

Axel Polleres
Siemens AG Austria
Siemensstraße 90, 1210 Vienna, Austria
axel.polleres@siemens.com

Sherif Sakr
NICTA and University of New South Wales
Sydney, Australia
sherif.sakr@nicta.com.au

ABSTRACT

XML and RDF are the W3C standard for representing and exchanging data and metadata about Web resources. Alongside, XQuery and SPARQL have been acknowledged as the standard query languages for querying the XML and RDF data models. In particular, they represent the counterpart of the SQL language in the world of relational databases. With the continued increase of XML and RDF usage on the Web and in corporate environments, efficient and scalable management of XML and RDF data represent fundamental challenges at the core of the vision of Web data management. Therefore, several techniques and systems have been proposed to tackle this problem. In this tutorial, we provide background on the fundamentals of querying and exchanging XML and RDF data and the integration between them. In addition, we provide a comprehensive overview of the main representatives of XML and RDF data management techniques and systems that covers an in-depth analysis of their different technical design decisions.

Keywords

XML, RDF, XQuery, SPARQL, XSPARQL, XML EXI, RDF HDT.

1. INTRODUCTION

The eXtensible Markup Language (XML) has been introduced by the end of the 1990's in order to create a standard data-format for the World Wide Web which can be easily handled by computers as well as by humans. XML has found practical application in numerous domains including data interchange, streaming data and data storage. In 2001 XQuery was recommended by the World Wide Web Consortium (W3C) as the standard XML query language. XQuery is based on a hierarchical and ordered document model which supports a wide variety of constructs and use cases. The language addresses a wide range of requirements, thus incorporating a rich set of features. On the mean time, the Resource Description Framework (RDF) has been introduced as another W3C recommendation that has rapidly gained popularity as a mean of expressing and exchanging semantic metadata, i.e., data that specifies semantic information about data. RDF was originally designed for the representation and processing of metadata about remote information sources and defines a model for describing relationships among resources in terms of uniquely identified attributes and values, but has emerged as a standard, graph-based data model, alternative to XML. The SPARQL query language (with its new release SPARQL 1.1) is the official W3C standard for querying and extracting information from RDF graphs. It is based on a powerful graph matching facility which allows binding variables to components in the input RDF graph and supports conjunctions and disjunctions of triple patterns. In essence, both XQuery and SPARQL represent the counterpart to select-project-join queries in the relational model.

With the continued increase of XML and RDF usage on the Web and in corporate environments, a gap becomes apparent between these two formats. Convenient languages and tools to transform between XML and RDF or merge sources of either format are missing, since both XQuery and SPARQL alone only insufficiently address this task. Thus, XSPARQL [2] has been presented as a hybrid language that provides an integration framework for XML, RDF and in its next release even JSON and relational data by partially combining several languages such as XQuery, SPARQL and SQL. The first session of the tutorial will provide a comprehensive background on the fundamentals of querying and exchanging XML and RDF data and the integration between them.

In practice, efficient and scalable management of XML and RDF data is a fundamental challenge at the core of the vision of Web data management. Thus, several techniques and systems have been proposed to tackle this problem [4, 10, 9]. These systems can be broadly classified into two main categories: native storage and query processing systems and relational-based storage and query processing systems. In the second part of the tutorial, we will provide a comprehensive overview of the main representatives of these two approaches. We will provide an in-depth discussion of the different design decisions. In addition, we will report about the results of recent benchmarking studies [7, 11, 12] in this domain and sketch potential directions for future work in the field of scalable Web data management.

2. TUTORIAL OUTLINE

The intended length of the proposed tutorial is 3 hours over two sessions. The first session will focus on providing a comprehensive background on the fundamentals of querying and exchanging XML and RDF data on the Web. In particular, the first session will cover the following list of items:

- The W3C standard query languages for XML and RDF: XQuery and SPARQL.
- SPARQL 1.0 vs. SPARQL 1.1[1]: new powerful features to query Web data.
- The bottlenecks of integrating XML and RDF data and the proposal of the XSPARQL query language [2] as well as possible implementations.

WWW2012, April 2012, Lyon, France.
.

[1] `http://www.w3.org/TR/sparql11-query/`

- Efficient formats for data exchange on the web: XML EXI[2] and RDF HDT[3].

In the second session, we will cover the state-of-the-art of systems for supporting large scale XML and RDF data management [10]. In addition, we will provide an in-depth analysis for some of these systems with a focus on systems where implementation details are published in scholarly articles (e.g. Pathfinder [5, 6], RDF-3x [8], OWLIM [3], Hexastore [13], SW-Store [1]) in addition to open source projects and commercially available systems (e.g. Apache Jena[4], AllegroGraph[5], Virtuoso[6], Oracle Semantic Web Technologies[7]) . In particular, we will explain design choices of these systems, analyze demands and access patterns of different applications and enumerate desiderata for semantic Web data management systems.

3. LEARNING OUTCOME

This tutorial is intended to benefit researchers and system designers in the broad area of scalable query engines for XML and RDF. The tutorial would benefit both designers of the query engines as well as users of these engines since a survey of the current systems and an in-depth understanding will is essential for choosing the appropriate system as well as designing an effective system. This tutorial does not require any knowledge on XML or RDF query engines. After attending this tutorial, the audience will have:

- An overview of the W3C standard query languages for XML and RDF.
- A good understanding of the challenges of implementing efficient and scalable XQuery and SPARQL query processors over large XML and RDF repositories and the possibilities of integrating them.
- A good understanding how to efficiently exchange big amounts of RDF and XML data on the Web.
- A comprehensive review of the state-of-the-art in Web data storage management and query processing techniques.
- Highlights for potential research directions to improve the state-of-the-art and support the efforts towards achieving the broad vision of the Web data management.

4. PRESENTERS

Axel Polleres has obtained a doctorate in Computer Science from TU xsVienna in 2003 and a Habilitation (venia docendi) in the subject of information systems at the same university in 2011. He worked at Univ. of Innsbruck from 2003-2006, at Univ. Rey Juan Carlos, Madrid, from 2006-2007, and at the Digital Enterprise Research Institute at the National Univ. of Ireland, Galway, as project leader and lecturer, from 2007-2011. In June 2011, Dr. Polleres joined Siemens AG's Corporate Technology Research division as Senior Research Scientist. His research is focuses on querying and reasoning about ontologies and linked data, rules and query languages, Semantic Web technologies and standards, Web Services and Knowledge Management. He has worked in several European and national research projects in these areas, is actively contributing to W3C standards such as RIF and SPARQL1.1, and has held tutorials and courses at different universities. Dr. Polleres has published over 100 scientific articles.

Sherif Sakr is a Research Scientist in the Software Systems Research Group at National ICT Australia (NICTA), Sydney, Australia. He is also a Conjoint Lecturer in The School of Computer Science and Engineering (CSE) at University of New South Wales (UNSW), Australia. He received his PhD degree in Computer Science from Konstanz University, Germany in 2007. He received his BSc and MSc degree in Computer Science from the Information Systems department at the Faculty of Computers and Information in Cairo University, Egypt, in 2000 and 2003 respectively. In 2011, Sherif held a Visiting Researcher position at Microsoft Research Laboratories, Redmond, USA. He has published more than 40 refereed research publications in international journals and conferences such as: VLDB, SIGMOD, WWW, ER, BPM, ICWS, JCSS, JCST, JDM and IEEE COMST. Dr. Sakr's research interest is data and information management in general, particularly in areas of indexing techniques, query processing and optimization techniques, semi-structured and graph data management, semantic Web and social networks.

5. REFERENCES

[1] D. Abadi, A. Marcus, S. Madden, and K. Hollenbach. SW-Store: a vertically partitioned DBMS for Semantic Web data management. *VLDB J.*, 18(2), 2009.

[2] W. Akhtar, J. Kopecký, T. Krennwallner, and A. Polleres. XSPARQL: Traveling between the XML and RDF Worlds - and Avoiding the XSLT Pilgrimage. In *ESWC*, 2008.

[3] B. Bishop, A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, and R. Velkov. OWLIM: A family of scalable semantic repositories. *Semantic Web*, 2(1), 2011.

[4] G. Gou and R. Chirkova. Efficiently Querying Large XML Data Repositories: A Survey. *IEEE TKDE*, 19(10), 2007.

[5] T. Grust, M. Mayr, J. Rittinger, S. Sakr, and J. Teubner. A SQL:1999 Code Generator for the Pathfinder XQuery Compiler. In *SIGMOD*, 2007.

[6] T. Grust, S. Sakr, and J. Teubner. XQuery on SQL Hosts. In *VLDB*, 2004.

[7] M. Morsey, J. Lehmann, S. Auer, and A. Ngonga Ngomo. DBpedia SPARQL Benchmark - Performance Assessment with Real Queries on Real Data. In *International Semantic Web Conference (1)*, 2011.

[8] T. Neumann and G. Weikum. RDF-3X: a RISC-style engine for RDF. *PVLDB*, 1(1), 2008.

[9] S. Sakr. XML compression techniques: A survey and comparison. *J. Comput. Syst. Sci.*, 75(5), 2009.

[10] S. Sakr and G. Al-Naymat. Relational Processing of RDF Queries: A Survey. *SIGMOD Record*, 38(4), 2009.

[11] M. Schmidt, T. Hornung, N. Küchlin, G. Lausen, and C. Pinkel. An Experimental Comparison of RDF Data Management Approaches in a SPARQL Benchmark Scenario. In *ISWC*, 2008.

[12] M. Schmidt, T. Hornung, G. Lausen, and C. Pinkel. SP^{2}Bench: A SPARQL Performance Benchmark. In *ICDE*, 2009.

[13] C. Weiss, P. Karras, and A. Bernstein. Hexastore: sextuple indexing for semantic web data management. *PVLDB*, 1(1), 2008.

[2] http://www.w3.org/XML/EXI/

[3] http://www.w3.org/Submission/2011/SUBM-HDT-20110330/

[4] http://incubator.apache.org/jena/

[5] http://www.franz.com/agraph/allegrograph/

[6] http://www.openlinksw.com/dataspace/dav/wiki/Main/VOSRDF

[7] http://www.oracle.com/technetwork/database/options/semantic-tech/