Web Data Management

The Internet and World Wide Web have revolutionized access to information. Users now store information across multiple platforms from personal computers, to smartphones, to Web sites such as YouTube and Picasa. As a consequence, data management concepts, methods, and techniques are increasingly focused on distribution concerns. That information largely resides in the network, as do the tools that process this information.

This book explains the foundations of XML, the Web standard for data management, with a focus on data distribution. It covers the many facets of distributed data management on the Web, such as description logics, that are already emerging in today's data integration applications and herald tomorrow's semantic Web. It also introduces the machinery used to manipulate the unprecedented amount of data collected on the Web. Several "Putting into Practice" chapters describe detailed practical applications of the technologies and techniques.

Striking a balance between the conceptual and the practical, the book will serve as an introduction to the new global information systems for Web professionals as well as for master's level courses.

Serge Abiteboul is a researcher at INRIA Saclay and ENS Cachan and cofounder of the start-up Xyleme. His previous books include the textbook *Foundations of Databases*.

Ioana Manolescu is a researcher at INRIA Saclay, and the scientific leader of the LEO team, joint between INRIA and University Paris XI.

Philippe Rigaux is a Professor of Computer Science at the Conservatoire National des Arts et Métiers. He has coauthored six books, including *Spatial Databases (2001)*.

Marie-Christine Rousset is a Professor of Computer Science at the University of Grenoble.

Pierre Senellart is an Associate Professor in the DBWeb team at Télécom ParisTech, the leading French engineering school specializing in information technology.

WEB DATA MANAGEMENT

Serge Abiteboul
INRIA Saclay – Île-de-France

Ioana Manolescu
INRIA Saclay – Île-de-France

Philippe Rigaux
CNAM, France

Marie-Christine Rousset
University of Grenoble, France

Pierre Senellart
Télécom ParisTech, France

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107012431

First published 2012

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Web data management / Serge Abiteboul... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 978-1-107-01243-1 (hardback)
1. Web databases. 2. Database management. 3. Electronic commerce.
I. Abiteboul, S. (Serge)
QA76.9.W43.W4154 2011
005.7402854678–dc23 2011037996

ISBN 978-1-107-01243-1 Hardback

Additional resources for this publication at http://webdam.inria.fr/Jorge/

Cambridge University Press & Assessment has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

Introduction *page* ix

Part 1 Modeling Web Data

1 Data Model 3

1.1 Semistructured Data 3
1.2 XML 5
1.3 Web Data Management with XML 16
1.4 The XML World 18
1.5 Further Reading 28
1.6 Exercises 29

2 XPath and XQuery 32

2.1 Introduction 32
2.2 Basics 32
2.3 XPath 42
2.4 FLWOR Expressions in XQuery 54
2.5 XPath Foundations 62
2.6 Further Reading 67
2.7 Exercises 69

3 Typing 72

3.1 Motivating Typing 72
3.2 Automata 75
3.3 Schema Languages for XML 80
3.4 Typing Graph Data 89
3.5 Further Reading 91
3.6 Exercises 92

4 XML Query Evaluation 95
4.1 Fragmenting XML Documents on Disk 97
4.2 XML Node Identifiers 99
4.3 XML Query Evaluation Techniques 103
4.4 Further Reading 112
4.5 Exercises 113

5 Putting into Practice: Managing an XML Database with EXIST 116
5.1 Prerequisites 116
5.2 Installing EXIST 117
5.3 Getting Started with EXIST 118
5.4 Running XPath and XQuery Queries with the Sandbox 120
5.5 Programming with EXIST 123
5.6 Projects 127

6 Putting into Practice: Tree Pattern Evaluation Using SAX 131
6.1 Tree-Pattern Dialects 131
6.2 CTP Evaluation 134
6.3 Extensions to Richer Tree Patterns 138

Part 2 Web Data Semantics and Integration

7 Ontologies, RDF, and OWL 143
7.1 Introduction 143
7.2 Ontologies by Example 145
7.3 RDF, RDFS, and OWL 148
7.4 Ontologies and (Description) Logics 159
7.5 Further Reading 169
7.6 Exercises 170

8 Querying Data Through Ontologies 171
8.1 Introduction 171
8.2 Querying RDF Data: Notation and Semantics 172
8.3 Querying Through RDFS Ontologies 176
8.4 Answering Queries Through DL-LITE Ontologies 179
8.5 Further Reading 194
8.6 Exercises 195

9 Data Integration 196
9.1 Introduction 196
9.2 Containment of Conjunctive Queries 199
9.3 Global-as-View Mediation 200
9.4 Local-as-View Mediation 204
9.5 Ontology-Based Mediators 215
9.6 Peer-to-Peer Data Management Systems 222
9.7 Further Reading 229
9.8 Exercises 229

10 Putting into Practice: Wrappers and Data Extraction with XSLT 231
10.1 Extracting Data from Web Pages 232
10.2 Restructuring Data 234

11 Putting into Practice: Ontologies in Practice (by Fabian M. Suchanek) 236
11.1 Exploring and Installing Yago 236
11.2 Querying Yago 237
11.3 Web Access to Ontologies 238

12 Putting into Practice: Mashups with YAHOO! PIPES and XProc 240
12.1 YAHOO! PIPES: A Graphical Mashup Editor 240
12.2 XProc: An XML Pipeline Language 241

Part 3 Building Web Scale Applications

13 Web Search 247
13.1 The World Wide Web 248
13.2 Parsing the Web 250
13.3 Web Information Retrieval 257
13.4 Web Graph Mining 272
13.5 Hot Topics in Web Search 280
13.6 Further Reading 281
13.7 Exercises 283

14 An Introduction to Distributed Systems 287
14.1 Basics of Distributed Systems 288
14.2 Failure Management 295
14.3 Required Properties of a Distributed System 299
14.4 Particularities of P2P Networks 303
14.5 Case Study: A Distributed File System for Very Large Files 305
14.6 Further Reading 308

15 Distributed Access Structures 310
15.1 Hash-Based Structures 311
15.2 Distributed Indexing: Search Trees 325
15.3 Further Reading 336
15.4 Exercises 337

16 Distributed Computing with MAPREDUCE and PIG 339
16.1 MAPREDUCE 341
16.2 PIG 348
16.3 Further Reading 359
16.4 Exercises 361

17 Putting into Practice: Full-Text Indexing with LUCENE (by Nicolas Travers) 364

17.1 Preliminary: A LUCENE Sandbox 364
17.2 Indexing Plain Text with LUCENE – A Full Example 364
17.3 Put It into Practice! 371
17.4 LUCENE – Tuning the Scoring (Project) 372

18 Putting into Practice: Recommendation Methodologies (by Alban Galland) 374

18.1 Introduction to Recommendation Systems 374
18.2 Prerequisites 375
18.3 Data Analysis 377
18.4 Generating Some Recommendations 380
18.5 Projects 385

19 Putting into Practice: Large-Scale Data Management with HADOOP 387

19.1 Installing and Running HADOOP 388
19.2 Running MAPREDUCE Jobs 391
19.3 PIGLATIN Scripts 395
19.4 Running in Cluster Mode (Optional) 395
19.5 Exercises 397

20 Putting into Practice: COUCHDB, a JSON Semistructured Database 400

20.1 Introduction to the COUCHDB Document Database 401
20.2 Putting COUCHDB into Practice! 417
20.3 Further Reading 419

Bibliography 421
Index 431

Introduction

The Internet and the Web have revolutionized access to information. Individuals are depending more and more on the Web to find or publish information, download music and movies, and interact with friends in social networking Web sites. Following a parallel trend, companies go more and more toward Web solutions in their daily activity by using Web services (e.g., agenda) as well as by moving some applications into the cloud (e.g., with Amazon Web services). The growth of this immense information source is witnessed by the number of newly connected people, by the interactions among them facilitated by the social networking platforms, and above all by the huge amount of data covering all aspects of human activity. With the Web, information has moved from data isolated in very protected islands (typically relational databases) to information freely available to any machine or any individual connected to the Internet.

Perhaps the best illustration comes from a typical modern Web user. She has information stored on PCs, a personal laptop, and a professional computer, but also possibly on some server at work, on her smartphone, in an e-book, and so on. Also, she maintains information in personal Web sites or social network Web sites. She may store pictures in Picasa, movies in YouTube, bookmarks in Firefox Sync, and the like. So, even an individual is now facing the management of a complex distributed collection of data. On a different scale, public or private organizations also have to deal with information produced and stored in different places, or collected on the Web, either as a side effect of their activity (e.g., worldwide e-commerce or auction sites) or because they directly attempt to understand, organize and analyze data collected on the Web (e.g., search engines, digital libraries, or Web intelligence companies).

As a consequence, a major trend in the evolution of data management concepts, methods, and techniques is their increasing focus on distribution concerns: Since information now mostly resides in the network, so do the tools that process this information to make sense of it. Consider for instance the management of internal reports in a company. Typically, many collections of reports may be maintained in different local branches. To offer a unique company-wide query access to the *global* collection, one has to integrate these different collections. This leads to data management within a wide area network. Because of slow communications, the company may prefer to maintain such a large collection in a unique central repository. (This is not always possible for organizational reasons.) If the collection is a massive data set, it may rapidly outrange the capacity of a single computer. One may then choose to distribute the collection *locally* on a cluster of machines. Indeed, one may even prefer this solution simply because buying a cluster of cheap computers may be much cheaper than buying a single high-end machine with the same throughput as the cluster. This leads to data management within a local area network, with very fast communication. An extreme example that combines both aspects is Web search: The global collection is distributed on a wide area network (all documents on the Web) and the index is maintained on a local area network (e.g., a Google farm).

The use of global-area-network distribution is typical for Web data: Data relevant for a particular application may come from a large number of Web servers. Local-area-network distribution is also typical because of scalability challenges raised by the quantity of relevant data as well as the number of users and query load. Mastering the challenges opened by data distribution is the key to handling Web-scale data management.

MOTIVATION FOR THE BOOK

Distributed data management is not a new idea. Research labs and database companies have tackled the problem for decades. Since System R* or SDD-1, a number of distributed database systems have been developed with major technical achievements. There exist for instance very sophisticated tools for distributed transaction processing or parallel query processing. The main achievements in this context have been complex algorithms, notably for concurrency control (e.g., commit protocols), and global query processing through localization.

Popular software tools in this area are ETLs (for extract, transform, and load). To support performance needs, data are imported using ETLs from operational databases into warehouses and replicated there for local processing (e.g., OLAP or online analytical processing). Although a lot of techniques have been developed for propagating updates to the warehouse, this is much less frequently used. Data in warehouses are refreshed periodically, possibly using synchronization techniques in the style of that used for version control systems.

With the Web, the need for distributed data management has widely increased. Also, with Web standards and notably standards for Web services, the management of distributed information has been greatly simplified. For example, the simple task of making a database available on the network that was typically requiring hours with platforms such as Corba, can now be achieved in minutes. The software that is needed is widely available and often with free licenses. This is bringing back to light distributed data management.

The ambition of this book is to cover the many facets of distributed data management on the Web. We will explain the foundations of the Web standard for data management, XML. We will travel in logical countries (e.g., description logic) that provide foundations for the Semantic Web, which is emerging in modern data integration applications. We will show the beauty of software tools that everyone is already using today, for example Web search engines. And finally, we will explain the impressive machinery used nowadays to manipulate an unprecedented amount of data.

We are witnessing an emergence of a new global information system created, explored, and shared by the whole of humankind. This book aims to expose the recent achievements that help make this system usable.

SCOPE AND ORGANIZATION OF THE BOOK

Databases are a fantastic playground where theory and systems meet. The foundation of relational databases was first-order logic, and at the same time, relational systems are among the most popular software systems ever designed. In this book, theory and systems will also meet. We will encounter deep theory (e.g., logics for describing knowledge, automata for typing trees). We will also describe elegant algorithms and data structures such as PageRank or Distributed Hash Tables. We believe that all these aspects are needed to grasp the reality of Web data management.

We present this material in different core chapters that form, in our opinion, the principles of the topic. They include exercises and notes for further reading. We also see as essential putting this material into practice, so that it does not remain too abstract. This is realized in PiP (for Putting into Practice) chapters. For instance, after we present the abstract model for XML in core chapters, we propose a PiP for XML APIs (Application Programming Interfaces for XML), and one for eXist (an Open Source XML database). The approach is followed for the other topics addressed by the book. Our main concern is to deliver content that reaches a good balance between the conceptual aspects that help make sense of the often unstructured, heterogeneous, and distributed content of the Web and the practical tools that let practitioners acquire a concrete experience. Also, because software or environments typically evolve faster than core material, the PiP chapters are complemented by teaching material that can be found on a Web site.

The book is organized in three parts. The first part covers Web data modeling and representation, the second is devoted to semantic issues, and the last one delves into the low levels of Web scale data handling systems. We next detail these three parts.

Part 1: Modeling Web Data

The HTML Web is a fantastic means of sharing information. But HTML is fully oriented toward visual presentation and keyword search, which makes it appropriate for humans but much less so for access by software applications. This motivated the introduction of a *semistructured data model*, namely XML, which is well suited for both humans and machines. XML describes *content* and promotes machine-to-machine communication and data exchange. XML is a generic data exchange format that can be easily specialized to meet the needs of a wide range of data uses.

Because XML is a universal format for data exchange, systems can easily exchange information in a wide variety of fields, from bioinformatics to e-commerce. This universality is also essential to facilitate data integration. A main advantage (compared to previous exchange formats) is that the language comes equipped with an array of available software tools such as parsers, programming interfaces, and manipulation languages that facilitate the development of XML-based applications. Last but not least, the standard for distributed computation over the Internet is based on Web services and on the exchange of XML data.

This part proposes a wide but concise picture of the state-of-the-art languages and tools that constitute the XML world. We do not provide a comprehensive view of the specifications, but rather explain the main mechanisms and the rationales behind the specifications. After reading this part, the reader should be familiar enough with the semistructured data approach to understand its foundations and be able to pick up the appropriate tools when needed.

Part 2: Web Data Semantics and Integration

On the Web, given a particular need, it may be difficult to find a resource that is relevant to it. Also, given a relevant resource, it is not easy to understand what it provides and how to use it. To solve such limitations and facilitate interoperability, the Semantic Web vision has been proposed. The key idea is to also publish *semantic descriptions* of Web resources. These descriptions rely on *semantic annotations*, typically on logical assertions that relate resources to some terms in predefined *ontologies*.

An ontology is a formal description providing human users or machines a shared understanding of a given domain. Because of the logic inside, one can reason with ontologies, which is a key tool for integrating different data sources, providing more precise answers, or (semiautomatically) discovering and using new relevant resources.

In this part, we describe the main concepts of the Semantic Web. The goal is to familiarize the reader with ontologies: what they are, how to use them for query answering, how to use them for data integration.

Part 3: Building Web Scale Applications

At this stage of the book, we know how to exchange data and how to publish and understand semantics for this data. We are now facing the possibly huge scale of Web data. We will present main techniques and algorithms that have been developed for scaling to huge volumes of information and huge query rates. The few numbers that one may want to keep in mind are billions of Web documents, millions of Web servers, billions of queries per month for a top Web search engine, and a constant scale-up of these figures. Even a much smaller operation, such as a company-wide center, may have to store millions of documents and serve millions of queries.

How do you design software for that scale?

We will describe the basics of full-text search in general, and Web search in particular. Indexing is at the core of Web search and distributed data access. We will consider how to index huge collections in a distributed manner. We will also present specific techniques developed for large-scale distributed computing.

This part puts an emphasis on existing systems, taken as illustrative examples of more generic techniques. Our approach to explain distributed indexing techniques, for instance, starts from the standard centralized case, explains the issues raised by distribution, and shows how these issues have been tackled in some of the most prominent systems. Because many of these technologies have been implemented in Open Source platforms, they also form the basis of the PiP chapters proposed in this part.

INTENDED AUDIENCE

The book is meant as an introduction to the fascinating area of data management on the Web. It can serve as the material for a master course. Some of it may also be used in undergraduate courses. Indeed, the material in this book has already been tested, at both the undergraduate and graduate levels. The PiP chapters are meant to be the basis of labs or projects. Most of the material deals with well-established concepts, languages, algorithms, and tools. Occasionally, we included more speculative material issued from ongoing research dedicated to the emergence of this vision. This is to better illustrate important concepts we wanted to highlight. The book's content can thus also serve as an academic introduction to research issues regarding Web data management.

Among other viewpoints, one can view the Web as a very large library. In our travel within Web territories, we will be accompanied by a librarian, Jorge. This is in homage to Jorge Luis Borges whose short story *The Library of Babel* introduces a library preserving the whole human of knowledge.

COMPANION WEB SITE

A companion Web site for this book, available at `http://webdam.inria.fr/Jorge/`, contains electronic versions of this book, as well as additional materials (extra chapters, exercise solutions, lecture slides, etc.) pertaining to Web data management. In particular, all examples, data sets, and software used in the PiP chapters are available there.

ACKNOWLEDGMENTS

We would like to thank the following people who helped us to collect, organize, and improve the content of this book: Stanislav Barton (Internet Memory Foundation), Michael Benedikt (Oxford Univ.), Véronique Benzaken (Univ. Paris–Sud), Balder ten Cate (UCSC), Irini Fundulaki (FORTH Institute), Alban Galland (INRIA Saclay), François Goasdoué (Univ. Paris–Sud), David Gross-Amblard (INRIA Saclay), Fabrice Jouanot (Univ. Grenoble), Witold Litwin (Univ. Paris–Dauphine), Laurent d'Orazio (Univ. Clermont-Ferrand), Fabian Suchanek (INRIA Saclay), and Nicolas Travers (CNAM).

We are also grateful to the students at CNAM, ENS Cachan, Grenoble, Paris–Sud, and Télécom ParisTech who followed portions of this course and helped, through their questions and comments, to improve it.

This book has been developed as part of the Webdam project. The Webdam project is funded by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013), ERC grant Webdam, agreement 226513.