http://webdam.inria.fr/

Web Data Management

Introduction

Serge Abiteboul
INRIA Saclay & ENS Cachan

Ioana Manolescu
INRIA Saclay & Paris-Sud University

Philippe Rigaux
CNAM Paris & INRIA Saclay

Marie-Christine Rousset
Grenoble University

Pierre Senellart
Télécom ParisTech

http://webdam.inria.fr/Jorge/

Contents

Internet and the Web have revolutionized access to information. Individuals are more and more depending on the Web where they find or publish information, download music and movies, or interact with friends in social networking Websites. Following a parallel trend, companies go more and more towards 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 at some server at work, on her smartphone, an e-book, etc. 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, etc. So, even an individual is now facing the management of a complex distributed collection of data. At 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., world-wide e-commerce or auction sites) or because they directly attempt at understanding, organizing and analyzing 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 than 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 open by data distribution is the key to handle 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 is imported using ETLs from operational databases into warehouses and replicated there for local processing, (e.g., OLAP or on-line analytical processing). Although a lot of techniques have been developed for propagating updates to the warehouse, this is much less 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 that 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 amount of data of unprecedented size.

We are witnessing an emergence of a new, global information system created, explored, and shared by the whole humankind. The book aims at exposing 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 foundations 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 to put 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 a 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 in 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 I: 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 for accesses by software applications. This motivated the introduction of a *semistructured data model*, namely XML, that is well suited both for 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 usages.

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 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 what are 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 II: 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 key tool for integrating different data sources, providing more precise answers, or (semi automatically) 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 III: 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 rate. 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 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, material of the book has already been tested, both at the undergraduate and graduate levels. The PiPs 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 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, or 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), David Gross-Amblard (INRIA Saclay), Fran cois Goasdoué (Univ. Paris-Sud), Fabrice Jouanot (Univ. Grenoble), Pekka Kilpeläinen (Univ. of Eastern Finland), Witold Litwin (Univ. Paris-Dauphine), Laurent d'Orazio (Univ. Clermont-Ferrand), Fabian Suchanek (INRIA Saclay), Nicolas Travers (CNAM).

We are also grateful to the students at CNAM, ENS Cachan, Grenoble, Paris-Sud, or Télćom ParisTech who followed portions of this course and helped, by their questions and comments, improving 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.