Web Data Management

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

Sourav S. Bhowmick Sanjay K. Madria
Wee Keong Ng

Web Data Management

A Warehouse Approach

With 106 Illustrations

Springer

Sourav S. Bhowmick
and Wee Keong Ng
School of Computer Engineering
Nanyang Technological University
50 Nanyang Avenue
Blk N4 2A-32
Nanyang, 639798
Singapore
assourav@ntu.edu.sg
awkng@ntu.edu.sg

Sanjay K. Madria
University of Missouri
Department of Computer Science
1870 Miner Circle Drive
310 Computer Science Building
Rolla, MO 65409
USA
madrias@umr.edu

Library of Congress Cataloging-in-Publication Data
Bhowmick, Sourav S.
Web data management : a warehouse approach / Sourav S. Bhowmick, Sanjay K. Madria, Wee Keong Ng.
p. cm. — (Springer professional computing)
Includes bibliographical references and index.
ISBN 0-387-00175-1 (alk. paper)
1. Web databases. 2. Database management. 3. Data warehousing. I. Madria, Sanjay Kumar. II. Ng, Wee Keong. III. Title. IV. Series.
QA76.9.W43B46 2003
005.75′8—dc21 2003050523

ISBN 0-387-00175-1 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10901038

Typesetting: Pages created by the author using a Springer TEX macro package.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Dedication

Sourav:
Dedicated to my parents,
Himanshu Kumar Saha Bhowmick, and
Gouri Saha Bhowmick, and
to my wife Rachelle
for her infinite love, patience, and support.

Sanjay:
To my parents Dr. M. L. Madria and Geeta Madria
for their encouragement, and
to my wife Ninu and
sons Priyank and Pranal
for their love and support.

Wee Keong:
To my parents and family.

Preface

Overview

The existence of different autonomous Web sites containing related information has given rise to the problem of integrating these sources effectively to provide a comprehensive integrated source of relevant information. The advent of e-commerce and the increasing trend of availability of commercial data on the Web has generated the need to analyze and manipulate these data to support corporate decision making. Decision support systems now must be able to harness and analyze Web data to provide organizations with a competitive edge. In a recent report on the future of database research known as the Asilomar Report, it has been predicted that in a few years from now, the majority of human information will be available on the Web. The Web is evolving at an alarming rate and is becoming increasingly chaotic without any consistent organization. To address these problems, traditional information retrieval techniques have been applied to document collection on the Internet, and a panoply of search engines and tools have been proposed and implemented. Such techniques are sometimes time consuming, and laborious, and the results obtained may be unsatisfactory. Thus there is a need to develop efficient tools for analyzing and managing Web data. In this book, we address the problem of efficient management of Web information from the database perspective. We build a data warehouse called WHOWEDA (**W**are**h**ouse **O**f **We**b **Da**ta) for managing and manipulating Web data. This problem is more challenging compared to its relational counterpart due to the irregular unstructured nature of Web data. This has led to rethinking and reusing existing techniques in a new way to address the current challenges in Web data management.

A web warehouse acts as an information server that supports information gathering and can provide value-added services such as personalization, summarization, transcoding, and knowledge discovery. A web warehouse can also be a shared information repository. By building a shared web warehouse (in a company), we aim to maximize the sharing of information, knowledge, and experience among users who share common interests. Users may access the warehouse data from appliances such as PDAs and cell phones. Because these devices do not have the same rendering capabilities as desktop computers, it is necessary for Web contents to be adapted, or

transcoded, for proper presentation on a variety of client devices. Second, for very large documents, such as high-quality pictures, or video files, it is reasonable and efficient to deliver a small segment to clients before sending the complete version. A web warehouse supports automated resource discovery by integrating technologies for the search engine, filtering, and clustering.

The Web allows the information (both contents and structural modification) to change or disappear at any time and in any way. How many times have we noticed that bookmarked pages have suddenly disappeared or changed? Unless we store and archive these evolving pages, we will continue to lose some valuable knowledge over time. These rapid and often unpredictable changes or disappearances of information create the new problems of detecting, representing, and querying these changes. This is a challenging problem because the information sources in the Web are autonomous and typical database approaches to detect changes based on triggering mechanisms are not usable. Moreover, these information sources typically do not keep track of historical information in a format accessible to the outside user. When the versions of data are available, we can explore how a certain topic or community evolved over time. Web-related research and mining will benefit if the history of data can be warehoused. This will help in developing a change notification service that will notify users whenever there are changes of interest. The web warehouse can support many subscription services such as allowing changes to be detected, queried, and reported based on a user's query subscription.

Managing data in a web warehouse requires (1) design of a suitable data model for representing Web data in a repository, (2) development of suitable algebraic operators for retrieving data from the Web and manipulating the data stored in a warehouse, (3) tools for Web data visualization, and (4) design of change management and knowledge discovery tools. To address the first issue, we propose a data model called WHOM (WareHouse Object Model) to represent HTML and XML documents in the warehouse. To address the second issue, we define a set of web algebraic operators to manipulate Web data. These operators build new web tables by extracting relevant data from the Web, and generating new web tables from existing ones. To address the next issue, we introduce a set of data visualization operators to add flexibility in viewing query results coupled from the Web. Finally, we propose algorithms to perform change management and knowledge discovery in the web warehouse.

Organization and Features

We begin by introducing the characteristics of Web data in Chapter 1. We motivate the need for new warehousing techniques by describing the limitations of Web data and how conventional data warehousing techniques are ill-equipped to manage heterogeneous autonomous Web data. Within this context, we describe how a web warehouse differs from those studied in traditional data warehousing literature. We present an overview of our framework for modeling and manipulating Web data in the web warehouse. We present the conceptual architecture of the web warehouse, and identify its key modules and the subproblems they address. We define the scope

of this book by identifying the portion of the architecture and their subproblems that is addressed here. Then we briefly describe the key research issues raised by the need for storing and managing data in the web warehouse. Finally, we highlight the contributions of the book.

In Chapter 2, we discuss prior work in the Web data management area. We focus on high-level similarities and differences between prior work and our work, deferring detailed comparisons to later chapters that present our techniques in detail. We focus on three classes of systems based on the task they perform related to information management on the Web: modeling and querying the Web, information extraction and integration and, Web site construction and restructuring. Furthermore, we discuss recent research in XML data modeling, query languages, and data warehousing systems for Web data. The knowledgeable reader may omit this chapter, and perhaps refer back to comparisons while reading later chapters of the book.

In Chapter 3, we describe the issues that we have considered in modeling warehouse data. We provide a brief overview of WHOM, the data model for the web warehouse. We present a simple and general model for representing metadata, structure, and content of Web documents and hyperlinks as trees called node and link metadata trees, and node and link data trees. Within this context, we identify HTML elements and attributes considered useful in the context of the web warehouse for generating tree representations of content and structure of HTML documents.

Chapter 4 describes a flexible scheme to impose constraints on metadata, content and structure of HTML and XML data. An important feature of our scheme is that it allows us to impose constraints on a specific portion of Web documents or hyperlinks, on attributes associated with HTML or XML elements, and on the hierarchical structure of Web documents, instead of simple keyword-based constraints similar to the search engines. It also presents a mechanism to associate two sets of documents or hyperlinks using comparison predicates based on their metadata, content, or structural properties.

In Chapter 5, we present a mechanism to represent constraints imposed on the hyperlinked connection in a set of Web documents (called *connectivity*) in WHOM). An important feature of our approach is that it can represent interdocument relationships based on partial knowledge of the user about the hyperlinked structure. We discuss the syntax and semantics of a connectivity element. In this context, we motivate the syntax and semantics of connectivities by identifying various real examples.

In Chapter 6, we present a mechanism for querying the Web. The complete syntax is unveiled and some examples are given to demonstrate the expressive power of the query mechanism. Some of the important features of our query mechanism are the ability to query metadata, content, internal and external (hyperlink) structure of Web documents based on partial knowledge, ability to express constraints on tag attributes and tagless segments of data, the ability to express conjunctive as well as disjunctive query conditions compactly, the ability to control execution of a web query and preservation of the topological structure of hyperlinked documents

in the query results. We also discuss various properties, validity conditions, and limitations of the query mechanism.

In Chapter 7, we present a novel method for describing schema of a set of relevant Web data. An important feature of our schema is that it represents a collection of Web documents relevant to a user, instead of representing any set of Web documents. We also describe the syntax, semantics, and properties of a web schema and introduce the notion of web tables. Again, an important advantage of our web schema is that it provides the flexibility to represent irregular, heterogeneous structured data. We present the mechanism for generating a web schema in the context of a web warehouse.

In Chapter 8, we focus on how web tables are generated and are further manipulated in the web warehouse by a set of web algebraic operators. The web algebra provides a formal foundation for data representation and manipulation for the web warehouse. Each web operator accepts one or two web tables as input and produces a web table as output. A set of simple web schemas and web tuples are produced each time a web operator is applied. The global web coupling operator extracts web tuples from the Web. In particular, portions of the World Wide Web (WWW) are extracted when it is applied to the WWW. The web union, web cartesian product, and the web join are binary operators on web tables. Web select extracts a subset of web tuples from a web table. Web project removes some of the nodes from the web tuples in a web table. The web distinct operator removes duplicate web tuples from a web bag.

A user may wish to view web tuples in a different framework. In Chapter 9, we introduce a set of data visualization operators such as web nest, web unnest, web coalesce, web pack, web unpack, and web sort to add flexibility in viewing query results coupled from the Web. The web nest and web coalesce operators are similar in nature. Both of these operators concatenate a set of web tuples over identical nodes and produce a set of directed graphs as output. The web pack and web sort operations produce a web table as output. The web pack operator enables us to group web tuples based on the domain name or host name of the instances of a specified node type identifier or the keyword set in these nodes. A web sort, on the other hand, sorts web tuples based on the total number of nodes or total number of local, global or interior links in each tuple. Web unnest, web expand and web unpack perform the inverse functions of web nest, web coalesce and web pack respectively.

In Chapter 10, our focus is on detecting and representing changes given old and new versions of a set of interlinked Web documents, retrieved in response to a user's query. We present a mechanism to detect relevant changes using web algebraic operators such as web join and outer web join. Web join is used to detect identical documents residing in two web tables, whereas outer web join, a derivative of web join, is used to identify dangling web tuples. We discuss how to represent these changes using delta web tables. We have designed and discussed formal algorithms for the generation of delta web tables.

In Chapter 11, we introduce the concept of the web bag in the context of the web warehouse. Informally, a web bag is a web table that allows multiple occurrences of identical web tuples. We have used the web bag to discover useful knowledge from

a web table such as visible documents (or Web sites), luminous documents, and luminous paths. In this chapter, we formally discuss the semantics and properties of web bags. We provide formal algorithms for various types of knowledge discovery in a web warehouse using the web bag and illustrate them with examples.

In Chapter 12, we conclude by summarizing the contributions of this book and discussion on promising directions for future work in this area. Readers can benefit by exploring the research directions given in this chapter.

Audiences

This book furnishes a detailed presentation of relevant concepts, models, and methods in a clear, simple style, providing an authoritative and comprehensive survey and resource for web database management systems developers and enterprise Web site developers. The book also outlines many research directions and possible extensions to the ideas presented, which makes this book very helpful for students doing research in this area.

The book has very strong emphasis and theoretical perspective on designing the web data model, schema development, web algebraic operators, web data visualization, change management, and knowledge discovery. The solid theoretical foundation will provide a good platform for building the web query language and other tools for the manipulation of warehoused data. The implementation of the discussed algorithms will be good exercises for undergraduate and graduate students to learn more about these operators from a system perspective. Similarly, the development of tools for application developers will serve as the foundation for building the web warehouse.

Prerequisites

This book assumes that readers have some introductory knowledge of relational database systems and HTML or XML as well as some knowledge of HTML or XML syntax and semantics for understanding of the initial chapters. A database course at the undergraduate or graduate level or familiarity with the concepts of relational schema, data model, and algebraic operators is a sufficient prerequisite for digesting the concept described in the later chapters. For professionals, working knowledge of relational database systems and HTML programming is sufficient to grasp the ideas presented throughout this book. Some exposure to the internals of search engines will help in comparing some of the methodology described in the context of the web warehouse. A good knowledge of C++/Java programming language at a beginner's level is sufficient to code the algorithm described herein. For readers interested in learning the area of Web data management, the book provides many examples throughout the chapters, which highlight and explain the intrinsic details.

Acknowledgments

It is a great pleasure for us to acknowledge the assistance and contributions of a large number of individuals to this effort. First, we would like to thank our publisher Springer-Verlag for their support. In particular, we would like to acknowledge the efforts, help, and patience of Wayne Wheeler, Wayne Yuhasz, Frank Ganz, and Timothy Tailor, our primary contacts for this edition.

The work reported in this book grew out of the Web Warehousing Project (WHOWEDA) at the Nanyang Technological University, Singapore. In this project, we explored various aspects of the web warehousing problem. Building a warehouse that accommodates data from the WWW has required us to rethink nearly every aspect of conventional data warehouses. Consequently, quite a few doctoral and master's dissertations have resulted from this project. Specifically, the chapters in this book are larger extensions in terms of scope and details of some of the papers published in journals and conferences and some initial chapters of Sourav's thesis work. Consequently, Dr. Wee Keong Ng, who was also Sourav's advisor, deserves the first thank you. Not only did he introduce Sourav to interesting topics in the database field, he was also always willing to discuss ideas, no matter how strange they were. In addition to Dr. Wee Keong Ng, the WHOWEDA project would have not been successful without the contributions made by Dr. Lim Ee-Peng who advised on many technical issues related to the WHOWEDA project.

In addition, we would also like to express our gratitude to all the group members, past and present, in the WHOWEDA project team. In particular, Feng Qiong, Cao Yinyan, Luah Aik Kee, Pallavi Priyadarshini, Ang Kho Kiong, and Huang Chee Thong made substantial contributions to the implementation of some of the components of WHOWEDA.

Quite a few people have helped us with the initial vetting of the text for this book. It is our pleasure to acknowledge them all here. We would like to thank Samuel Mulder for carefully proofreading the complete book in a short span of time and suggesting the changes which have been incorporated. We would also like to acknowledge Erwin Leonardi and Zhao Qiankun (graduate students in NTU) for refining some of the contents of this book.

Sourav S. Bhowmick would like to acknowledge his parents who gave him incredible support throughout the years. Thanks to Diya, his cute and precocious two year old niece, who has already taught him a lot; nothing matters more than drinking lots of milk, smiling a lot, and sleeping whenever you want to. A special thanks goes to his wife Rachelle, for her constant love, support, and encouragement. A special thanks goes to Rod Learmonth who was Sourav's mentor and a great motivator during his days in Griffith University. He was the major force behind the development of Sourav's aspiration to pursue a doctoral degree.

Sanjay would also like to mention his source of encouragement: his parents and the constant love and affection of his wife Ninu and sons Priyank and Pranal for giving him time out to work on the book at various stages, specially making a visit to Singapore in December 2002. He would also like to thank his friends and students who also helped him in many ways in completing the book.

Finally, we would like to thank the School of Computer Engineering of Nanyang Technological University, Singapore for the generous resources and financial support provided for the WHOWEDA project. We would also like to thank the Computer Science Department at the University of Missouri-Rolla for allowing the use of their resources to help complete the book.

DR. SOURAV S. BHOWMICK, DR. SANJAY MADRIA, DR. WEE KEONG NG
Nanyang Technological University, Singapore
University of Missouri-Rolla, USA
April 5th, 2003

Contents