The explosive growth of the Web has dramatically changed the way in which information is managed and accessed. In particular, nowadays, the Web has evolved rapidly from a simple information-sharing environment (offering only static text and images) to a rich framework of dynamic and interactive services (such as video and audio conferencing, e-commerce, and distance learning). This enormous growth and diversity in terms of access devices, bandwidth, information sources, and content has complicated Web data-management frameworks and practices.

The Web is actually a distributed global information resource containing a large spectrum of applications in which users interact with (or within) companies, organizations, governmental agencies, and educational or collaborative environments. The popularity of the Web originates from its potential to deliver readily dynamic, distributed, heterogeneous, and unstructured data all over the world. In this context, the Web is evolving at an alarming rate and is becoming increasingly chaotic without any specific, consistent organization. Therefore, the need of various Web data-management techniques and mechanisms has become obligatory toward providing information (that is actually useful to users) and improving information circulation and dissemination over the Web. Furthermore, new tools and techniques are needed to effectively manage these data since managing Web data with conventional tools is becoming almost impossible.

Efficient and effective Web data-management practices may form the basis for developing intelligent, personalized, and business-optimal Web services. Such enabling practices include Web data mining, scalable data warehousing, and preprocessing, sequence discovery, real-time processing, users and documents clustering and classification, user modeling, and evaluation models. These issues may provide valuable information about user preferences
and expectations, together with usage, content, and structural patterns as practiced over the Web.

As the demand for data and information management increases, there is also a critical need for effectively managing Web content. Specifically, Web data mining has a major effect on the performance of Web data accessing and querying. Therefore, new implementations (such as Web data clustering, Web data caching, and Web services) have emerged to manage the (continuously growing) number of documents, their dynamic content, and services under quality-of-service (QoS) guarantees. The term QoS refers to certain technical characteristics, such as performance, scalability, reliability, and speed. So, current Web data-management issues are enforced with specific capabilities to design new Web applications and improve Web data searching and workload balancing.

In this framework, the chapters of this book provide an overview of current research and development activities in the area of Web data management. Following our call for chapters in 2005, we received 25 chapter proposals. All chapters underwent a rigorous, double-blind refereeing process before final acceptance. Eventually, 12 chapters were accepted for inclusion in this book. This book brought together academic and industrial researchers and practitioners from many different countries, including Australia, Brazil, Canada, Cyprus, Czech Republic, France, Greece, Italy, Spain, and the USA. Authors’ research and industrial experience are reflected in their work and will certainly offer to readers in-depth knowledge of their areas of expertise.

**Organization of This Book**

The broad range of topics of the present book makes it an excellent reference on Web data management. The book is organized so that it could cover a wide range of audiences including undergraduate university students, postgraduate students, research engineers, and system developers. Because each chapter is self-contained, readers can focus on the topics that most interest them. Most of the chapters (if not all) in this book have great practical utility. Each chapter is accompanied by examples or case studies to show the applicability of the described techniques or methodologies. Each chapter opens with an abstract, which is the summary of the chapter, and an introduction, and then closes with a conclusion in order to give readers adequate background and knowledge for understanding the subject. All chapters also include an extensive list of references to allow readers to understand the subject thoroughly by not only studying the topic in depth, but also by referring to other works related to their topic of interest.

Overall, this book includes work in some of the most breaking topics, such as Web modeling, clustering over Web data, XML (extensible markup language) data management, Web services, and Web data prefetching and caching. A short outline of the material presented in each of the chapters of this book follows to help readers understand the chapter within their interests.

The book is divided into four major sections:

I. Web Data Mining
II. Content Management on the Web
III. Web Information Integration and Applications
IV. Web Services for Data Accessing
Section I deals with the issues concerned with Web data clustering and association rules, providing a pathway for current research and development activities in this area. This section is further divided into four chapters:

**Chapter I**, entitled “Data Clustering: From Documents to the Web” by Dušan Húsek, Jaroslav Pokorný, Hana Řezanková, and Václav Snášel, presents an overview of approaches to clustering in the Web environment. Particularly, this chapter deals with clustering Web search results, in which clustering search engines arrange the search results into groups around a common theme. Furthermore, some considerations are given concerning the justification of so many clustering algorithms and their application in the Web environment.

**Chapter II**, entitled “Clustering Web Information Sources” by Athena Vakali, George Pallis, and Lefteris Angelis, focuses on the topic of clustering information over the Web in an effort to provide an overview and survey on the theoretical background and the adopted practices of the most popular emerging and challenging clustering research efforts. An up-to-date survey of the existing clustering schemes is given to be of use for both researchers and practitioners interested in the area of Web data mining.

**Chapter III**, entitled “An Overview of Similarity Measures for Clustering XML Documents” by Giovanna Guerrini, Marco Mesiti, and Ismael Sanz, presents the most indicative research efforts for clustering XML documents relying on contents, structures, and link-related properties of XML documents. Specifically, the most relevant similarity measures are evaluated, providing a systematic comparison of all the presented measures that allows one to determine which measure applies in a particular context.

**Chapter IV**, entitled “Mining Association Rules from XML Documents” by Laura Irina Rusu, Wenny Rahayu, and David Taniar, presents some of the existing mining techniques for extracting association rules out of XML documents in the context of rapid changes in the Web knowledge-discovery area. Specifically, it presents the latest discoveries in the area of mining association rules from XML documents, both static and dynamic, in a well-structured manner, with examples and explanations so the reader will be able to easily identify the appropriate technique for his or her needs and replicate the algorithm in a development environment. At the same time, this chapter includes research work with a high level of usability, in which concepts and models are easy to be applied in real situations without imposing knowledge of any high-level mathematics concepts.

Section II presents interesting techniques for effectively managing the Web content. There are three chapters in this section:

**Chapter V**, entitled “Dynamically Generated Web Content: Research and Technology Practices” by Stavros Papastavrou, George Samaras, Paraskevas Evripidou, and Panos Chrysanthis, deals with the dynamic Web content technology, which is definitely one of the most emerging research areas due to the exponential increase in the information circulation and dissemination over the Web. This chapter covers past and present research approaches, practices, and available technologies that facilitate the extraction of information from Web databases and its dissemination to Web users.

**Chapter VI**, entitled “Caching on the Web” by Mehregan Mahdavi and Boualem Benatallah, studies Web caching techniques with focus on dynamic content. Caching is a key technique that addresses some of the performance issues in today’s Web-enabled applications. Deploying dynamic data, especially in an emerging class of Web applications, called Web
portals, makes caching even more interesting. This chapter discusses the limitations of caching in Web portals and studies a solution that addresses these limitations. The solution is based on the collaboration between the portal and its providers.

**Chapter VII**, entitled “Information-Theoretic Methods for Prediction in the Wireless and Wired Web” by Dimitrios Katsaros, presents information-theoretic techniques for discrete sequence prediction. It surveys, classifies, and compares the state-of-the-art solutions, suggesting routes for further research by discussing the critical issues and challenges of prediction in wired and wireless networks.

Section III is about information integration and Web-based applications and includes three chapters:

**Chapter VIII**, entitled “Designing and Mining Web Applications: A Conceptual Modeling Approach” by Rosa Meo and Maristella Matera, presents the usage of a modeling language, WebML, for the design of Web applications. It discusses the advantages of adopting conceptual modeling for the design and maintenance of a Web data-intensive application. Furthermore, it presents a case study about the analysis of the conceptual logs for testifying to the effectiveness of WebML and its conceptual modeling methods. The methodology of the analysis of Web logs is based on the data-mining paradigm of item sets and frequent patterns and makes full use of constraints on the conceptual logs’ content. Many interesting patterns are obtained, such as recurrent navigation paths, the most frequently visited page’s contents, and anomalies.

**Chapter IX**, entitled “Integrating Heterogeneous Data Sources in the Web” by Angelo Brayner, Marcelo Meirelles, and José de Aguiar Moraes Filho, describes an extension to the XQuery language, called MXQuery, which supports queries over several data sources and solves integration problems as semantic heterogeneity and incomplete information. The proposed language provides the necessary support for integrating a variable number of data sources with different degrees of autonomy. MXQuery solves problems of data integration, such as semantics heterogeneity, and copes with incomplete information. Furthermore, this chapter presents an architecture to process MXQuery queries over multiple heterogeneous databases available on the Web.

**Chapter X**, entitled “E-Mail Mining: Emerging Techniques for E-Mail Management” by Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas, deals with e-mail mining. In particular, this chapter discusses how disciplines like machine learning and data mining can contribute to the solution of the problem by constructing intelligent techniques that automate e-mail managing tasks, and what advantages they hold over other conventional solutions. It also discusses the particularity of e-mail data and what special treatment e-mail requires. Some interesting e-mail mining applications like mail categorization, summarization, automatic answering, and spam filtering are also presented.

Section IV presents insights and perspectives for Web services and contains two chapters:

**Chapter XI**, entitled “Web Services: Technology Issues and Foundations” by Bernd Amann, Salima Benbernou, and Benjamin Nguyen, introduces the concept of service-oriented computing (SOC) on the Web and the current standards enabling the definition and publication of Web services. Moreover, this chapter illustrates the complexity of the Web-service composition problem and provides a representative overview of the existing approaches.
The chapter concludes with a short presentation of two research projects exploiting and extending the Web-service paradigm. Chapter XII, “Web-Services Management: Toward Efficient Web Data Access” by Farhana Zulkernine and Pat Martin, presents an overview and the state-of-the-art of various management approaches, models, and architectures for Web-services systems toward achieving QoS in Web data access. Moreover, it discusses the importance of autonomic or self-managing systems and provides an outline of the current research on autonomic Web services.

What Makes This Book Different

Several research efforts have already appeared in the area of Web data management, and this field seems to be of high importance for a wide academic and technical group due to the difficulties raised by the diversity of Web data structure and representation, information distribution, and communication and accessing costs. However, a dedicated book on important issues in Web data-management systems is still difficult to find. Most books are about either Web technology focusing on developing Web warehouses, or very specific areas such as Web modeling, Web mining, and Web replication.

This book provides a complete overview on important aspects in the Web data-management practice in order to be used either as a class textbook or as a complementary course text in a Web data-management course; in that case, its level is suitable for undergraduate- or graduate-level courses.

This book, therefore, different in that it covers an extensive range of topics, including related issues about Web modeling, Web mining, Web caching and replication, Web semantics, and the XML standard. Furthermore, the main advantage of this book is the integration of both theoretical and practical aspects in the Web data-management research area.

Intended Audience

*Web Data Management Practices: Emerging Techniques and Technologies* is intended for academic institutions and for working professionals, and for technical and non-technical readers. The broad range of topics in this book makes it a pathway for current research and development activities in the area of Web data management. The book is organized so that it could cover a wide range of audiences including undergraduate university students, postgraduate students, research engineers, and system developers.

Computer science instructors could use this book to teach Web data-management issues to senior undergraduate or postgraduate students. The chapters are organized such that they provide a great deal of flexibility; emphasis can be given to different chapters depending on the scope of the course and the instructor’s interests. Equivalently, computer-science students could use it in the context of a course or as a supplementary book for their independent study.

Computer-science researchers could also benefit from this book because it surveys a vast content of recent research in the area of Web data management. The research coverage is likely to benefit researchers and students from academia as well as industry. Moreover, this book is also ideal for researchers from other computer-science disciplines who wish to
get acquainted with this area and integrate it with their own fields. The general computer community will benefit from this book through its technical as well as practical overview of the area.

Finally, the chapters in this book can be used by Web application developers as a reference to use the correct techniques for modeling and designing Web services, as well as efficiently handling a huge amount of Web information.

---

**How to Read This Book**

The book as a whole is meant for anyone professionally interested in Web data-management techniques and who in some way wants to gain an understanding of how data-mining applications are implemented on the Web. The organization of the book has been carefully selected to help the reader. Each chapter may be studied separately or in conjunction with other chapters. Thus, it is not mandatory to study the topics in their order of appearance. If the reader wishes to perform an in-depth study of a particular subject, then he or she could focus on the corresponding section.

Although how the book is read largely depends on the personal interests of the reader, two possible paths are recommended. For readers who are interested in the Web applications and implementations, it is recommended to read Sections II, III, and IV. For the reader who wants to acquire a theoretical knowledge about Web data management issues, it is recommended to read Sections I and II, and at least “scan” Section IV.

---

**A Closing Remark**

The authors have made significant efforts to provide high-quality chapters despite space restrictions. The authors are well-known researchers in the area of Web data management, and they have already offered significant contributions to the literature. We hope that the reader will benefit from the works presented in this book.