Web Mining and Social Networking

Web Information Systems Engineering and Internet Technologies

Book Series

Series Editor: Yanchun Zhang, Victoria University, Australia

For more titles in this series, please visit
www.springer.com/series/6970

Semistructured Database Design by Tok Wang Ling, Mong Li Lee, Gillian Dobbie ISBN 0-378-23567-1
Web Content Delivery edited by Xueyan Tang, Jianliang Xu and Samuel T. Chanson ISBN 978-0-387-24356-6
Web Information Extraction and Integration by Marek Kowalkiewicz, Maria E. Orlowska, Tomasz Kaczmarek and Witold Abramowicz ISBN 978-0-387-72769-1 FORTHCOMING

Guandong Xu • Yanchun Zhang • Lin Li

Web Mining and Social Networking

Techniques and Applications

Springer

Guandong Xu
Centre for Applied Informatics
School of Engineering & Science
Victoria University
PO Box 14428, Melbourne
VIC 8001, Australia
Guandong.Xu@vu.edu.au

Yanchun Zhang
Centre for Applied Informatics
School of Engineering & Science
Victoria University
PO Box 14428, Melbourne
VIC 8001, Australia
Yanchun.Zhang@vu.edu.au

Lin Li
School of Computer Science & Technology
Wuhan University of Technology
Wuhan Hubei 430070
China
cathylilin@whut.edu.cn

ISBN 978-1-4419-7734-2 e-ISBN 978-1-4419-7735-9
DOI 10.1007/978-1-4419-7735-9
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010938217

Dedication to

To Feixue and Jack
From Guandong

To Jinli and Dana
From Yanchun

To Jie
From Lin

Preface

World Wide Web has become very popular in last decades and brought us a powerful platform to disseminate information and retrieve information as well as analyze information, and nowadays the Web has been known as a big data repository consisting of a variety of data types, as well as a knowledge base, in which informative Web knowledge is hidden. However, users are often facing the problems of information overload and drowning due to the significant and rapid growth in amount of information and the number of users. Particularly, Web users usually suffer from the difficulties in finding desirable and accurate information on the Web due to two problems of low precision and low recall caused by above reasons. For example, if a user wants to search for the desired information by utilizing a search engine such as Google, the search engine will provide not only Web contents related to the query topic, but also a large mount of irrelevant information (or called noisy pages), which results in difficulties for users to obtain their exactly needed information. Thus, these bring forward a great deal of challenges for Web researchers to address the challenging research issues of effective and efficient Web-based information management and retrieval.

Web Mining aims to discover the informative knowledge from massive data sources available on the Web by using data mining or machine learning approaches. Different from conventional data mining techniques, in which data models are usually in homogeneous and structured forms, Web mining approaches, instead, handle semi-structured or heterogeneous data representations, such as textual, hyperlink structure and usage information, to discover "nuggets" to improve the quality of services offered by various Web applications. Such applications cover a wide range of topics, including retrieving the desirable and related Web contents, mining and analyzing Web communities, user profiling, and customizing Web presentation according to users preference and so on. For example, Web recommendation and personalization is one kind of these applications in Web mining that focuses on identifying Web users and pages, collecting information with respect to users navigational preference or interests as well as adapting its service to satisfy users needs.

On the other hand, for the data on the Web, it has its own distinctive features from the data in conventional database management systems. Web data usually exhibits the

following characteristics: the data on the Web is huge in amount, distributed, heterogeneous, unstructured, and dynamic. To deal withe the heterogeneity and complexity characteristics of Web data, Web community has emerged as a new efficient Web data management means to model Web objects. Unlike the conventional database management, in which data models and schemas are well defined, Web community, which is a set of Web-based objects (documents and users) has its own logical structures. Web communities could be modeled as Web page groups, Web user clusters and co-clusters of Web pages and users. Web community construction is realized via various approaches on Web textual, linkage, usage, semantic or ontology-based analysis. Recently the research of Social Network Analysis in the Web has become a newly active topic due to the prevalence of Web 2.0 technologies, which results in an inter-disciplinary research area of Social Networking. Social networking refers to the process of capturing the social and societal characteristics of networked structures or communities over the Web. Social networking research involves in the combination of a variety of research paradigms, such as Web mining, Web communities, social network analysis and behavioral and cognitive modeling and so on.

This book will systematically address the theories, techniques and applications that are involved in Web Mining, Social Networking, Web Personalization and Recommendation and Web Community Analysis topics. It covers the algorithmic and technical topics on Web mining, namely, Web Content Mining, Web linkage Mining and Web Usage Mining. As an application of Web mining, in particular, Web Personalization and Recommendation is intensively presented. Another main part discussed in this book is Web Community Analysis and Social Networking. All technical contents are structured and discussed together around the focuses of Web mining and Social Networking at three levels of theoretical background, algorithmic description and practical applications.

This book will start with a brief introduction on Information Retrieval and Web Data Management. For easily and better understanding the algorithms, techniques and prototypes that are described in the following sections, some mathematical notations and theoretical backgrounds are presented on the basis of *Information Retrieval* (IR), *Nature Language Processing*, *Data Mining* (DM), *Knowledge Discovery* (KD) and *Machine Learning* (ML) theories. Then the principles, and developed algorithms and systems on the research of Web Mining, Web Recommendation and Personalization, and Web Community and Social Network Analysis are presented in details in seven chapters. Moreover, this book will also focus on the applications of Web mining, such as how to utilize the knowledge mined from the aforementioned process for advanced Web applications. Particularly, the issues on how to incorporate Web mining into Web personalization and recommendation systems will be substantially addressed accordingly. Upon the informative Web knowledge discovered via Web mining, we then address Web community mining and social networking analysis to find the structural, organizational and temporal developments of Web communities as well as to reveal the societal sense of individuals or communities and its evolution over the Web by combining social network analysis. Finally, this book will summarize the main work mentioned regarding the techniques and applications of

Web mining, Web community and social network analysis, and outline the future directions and open questions in these areas.

This book is expected to benefit both research academia and industry communities, who are interested in the techniques and applications of Web search, Web data management, Web mining and Web recommendation as well as Web community and social network analysis, for either in-depth academic research and industrial development in related areas.

Aalborg, Melbourne, Wuhan
July 2010

Guandong Xu
Yanchun Zhang
Lin Li

Acknowledgements: We would like to first appreciate Springer Press for giving us an opportunity to make this book published in the Web Information Systems Engineering & Internet Technologies Book Series. During the book writing and final production, Melissa Fearon, Jennifer Maurer and Patrick Carr from Springer gave us numerous helpful guidances, feedbacks and assistances, which ensure the academic and presentation quality of the whole book. We also thank Priyanka Sharan and her team, who commit and oversee the production of the text of our book from manuscript to final printer files, providing several rounds of proofing, comments and corrections on the pages of cover, front matter as well as each chapter. Their dedicated work to the matters of style, organization, and coverage, as well as detailed comments on the subject matter of the book adds the decorative elegance of the book in addition to its academic value. To the extent that we have achieved our goals in writing this book, they deserve an important part of the credit.

Many colleagues and friends have assisted us technically in writing this book, especially researchers from Prof. Masaru Kitsuregawa's lab at University of Tokyo . Without their help, this book might not have become reality so smoothly. Our deepest gratitude goes to Dr. Zhenglu Yang, who was so kind to help write the most parts of Chapter 3, which is an essential chapter of the book. He is an expert in the this field. We are also very grateful to Dr. Somboonviwat Kulwadee, who largely helped in the writing of Section 4.5 of Chapter 4 on automatic topic extraction. Chapter 5 utilizes a large amount of research results from the doctoral thesis provided by her as well. Mr. Yanhui Gu helps to prepare the section of 8.2.

We are very grateful to many people who have given us comments, suggestions, and proof readings on the draft version of this book. Our great gratitude passes to Dr. Yanan Hao and Mr. Jiangang Ma for their careful proof readings, Mr. Rong Pan for reorganizing and sorting the bibliographic file.

Last but not the least, Guandong Xu thanks his family for many hours they have let him spend working on this book, and hopes he will have a bit more free time on weekends next year. Yanchun Zhang thanks his family for their patient support through the writing of this book. Lin Li would like to thank her parents, family, and friends for their support while writing this book.

Contents

Part III Social Networking and Web Recommendation: Techniques and Applications

Part I

Foundation

1

Introduction

1.1 Background

With the dramatically quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. Due to the huge, diverse, dynamic and unstructured nature in Web data, Web data research has encountered a lot of challenges, such as heterogeneous structure, distributed residence and scalability issues etc. As a result, Web users are always drowning in an "ocean" of information and facing the problem of information overload when interacting with the Web, for example. Typically, the following problems are often encountered in Web related researches and applications:

(1). Finding relevant information: To find specific information on the Web, a user often either browses Web documents directly or uses a search engine as a search assistant. When the user utilizes a search engine to locate information, he or she often enters one or several keywords as a query, then search engine returns a list of ranked pages based on the relevance to the query. However, there are usually two major concerns associated with the query-based Web search [140]. The first problem is low precision, which is caused by a lot of irrelevant pages returned by search engines. The second problem is low recall, which is due to lack of capability of indexing all Web pages available on the Internet. This causes the difficulty in locating the unindexed information that is actually relevant. How to find more relevant pages to the query, thus, is becoming a popular topic in Web data management in last decade [274].

(2). Finding needed information: Since most of search engines perform in a query-triggered way that is mainly on a basis of one keyword or several keywords entered. Sometimes the results returned by the search engine are not exactly matched with what a user really needs due to the fact of existence of homograph. For example, when one user with information technology background wishes to search for information with respect to "Python" programming language, he/she might be presented with the information of creatural python, one kind of snake rather than programming language, given entering only one "python" word as the query. In other words, semantics of Web data [97] is rarely taken into account in the context of Web search.

G. Xu et al., *Web Mining and Social Networking,*
DOI 10.1007/978-1-4419-7735-9_1, © Springer Science+Business Media, LLC 2011

(3). Learning useful knowledge: With traditional Web search service, query results relevant to query input are returned to Web users in a ranked list of pages. In some cases, we are interested in not only browsing the returned collection of Web pages, but also extracting potentially useful knowledge out of them (data mining oriented). More interestingly, more studies [56, 46, 58] have been conducted on how to utilize the Web as a knowledge base for decision making or knowledge discovery recently.

(4). Recommendation/personalization of information: While a user is interacting with Web, there is a wide diversity of the user's navigational preference, which results in needing different contents and presentations of information. To improve the Internet service quality and increase the user click rate on a specific website, thus, it is necessary for Web developers or designers to know what the user really wants to do, to predict which pages the user would be potentially interested in, and to present the customized Web pages to the user by learning user navigational pattern knowledge [97, 206, 183].

(5). Web communities and social networking: Opposite to traditional data schema in database management systems, Web objects exhibit totally different characteristics and management strategy [274]. Existence of inherent associations amongst Web objects is an important and distinct phenomenon on the Web. Such kind of relationships can be modeled as a graphic expression, where nodes denote the Web objects and edges represent the linking or collaboration between nodes. In these cases, Web community is proposed to deal with Web data, and in some extent, is extended to the applications of social networking.

Above problems greatly suffer the existing search engines and other Web applications, and hereby produce more demands for Web data and knowledge research. A variety of efforts have been contributed to deal with these difficulties by developing advanced computational intelligent techniques or algorithms from different research domains, such as database, data mining, machine learning, information retrieval and knowledge management, etc. Therefore, the evolution of Web has put forward a great deal of challenges to Web researchers and engineers on innovative Web-based data management strategy and effective Web application development.

Web search engine technology [196] has emerged to carter for the rapid growth and exponential flux of Web data on the Internet, to help Web users find desired information, and has resulted in various commercial Web search engines available online such as Yahoo!, Google, AltaVista, Baidu and so on. Search engines can be categorized into two types: one is general-purpose search engines and the other is specific-purpose search engines. The general-purpose search engines, for example, the well-known Google search engine, try to retrieve as many Web pages available on the Internet that is relevant to the query as possible to Web users. The returned Web pages to user are ranked in a sequence according to their relevant weights to the query, and the satisfaction to the search results from users is dependent on how quickly and how accurately users can find the desired information. The specific-purpose search engines, on the other hand, aim at searching those Web pages for a specific task or an identified community. For example, Google Scholar and DBLP are two representatives of the specific-purpose search engines. The former is a search en-

gine for searching academic papers or books as well as their citation information for different disciplines, while the latter is designed for a specific researcher community, i.e. computer science, to provide various research information regarding conferences or journals in computer science domain, such as conference website, abstracts or full text of papers published in computer science journals or conference proceedings. DBLP has become a helpful and practicable tool for researchers or engineers in computer science area to find the needed literature easily, or for authorities to assess the track record of one researcher objectively. No matter which type the search engine is, each search engine owns a background text database, which is indexed by a set of keywords extracted from collected documents. To satisfy higher recall and accuracy rate of the search, Web search engines are requested to provide an efficient and effective mechanism to collect and manage the Web data, and the capabilities to match user queries with the background indexing database quickly and rank the returned Web contents in an efficient way that Web user can locate the desired Web pages in a short time via clicking a few hyperlinks. To achieve these aims, a variety of algorithms or strategies are involved in handling the above mentioned tasks [196, 77, 40, 112, 133], which lead to a hot and popular topic in the context of Web-based research, i.e. Web data management.

1.2 Data Mining and Web Mining

Data mining is proposed recently as a useful approach in the domain of data engineering and knowledge discovery [213]. Basically, data mining refers to extracting informative knowledge from a large amount of data, which could be expressed in different data types, such as transaction data in e-commerce applications or genetic expressions in bioinformatics research domain. No matter which type of data it is, the main purpose of data mining is discovering hidden or unseen knowledge, normally in the forms of patterns, from available data repository. Association rule mining, sequential pattern mining, supervised learning and unsupervised learning algorithms are commonly used and well studied data mining approaches in last decades [213].

Nowadays data mining has attracted more and more attentions from academia and industries, and a great amount of progresses have been achieved in many applications. In the last decade, data mining has been successfully introduced into the research of Web data management, in which a board range of Web objects including Web documents, Web linkage structures, Web user transactions, Web semantics become the mined targets. Obviously, the informative knowledge mined from various types of Web data can provide us help in discovering and understanding the intrinsic relationships among various Web objects, in turn, will be utilized to benefit the improvement of Web data management [58, 106, 39, 10, 145, 149, 167].

As known above, the Web is a big data repository and source consisting of a variety of data types as well as a large amount of unseen informative knowledge, which can be discovered via a wide range of data mining or machine learning paradigms. All these kinds of techniques are based on intelligent computing approaches, or so-

called computational intelligence, which are widely used in the research of database, data mining, machine learning, and information retrieval and so on.

Web (data) mining is one of the intelligent computing techniques in the context of Web data management. In general, Web mining is the means of utilizing data mining methods to induce and extract useful information from Web data information. Web mining research has attracted a variety of academics and engineers from database management, information retrieval, artificial intelligence research areas, especially from data mining, knowledge discovery, and machine learning etc. Basically, Web mining could be classified into three categories based on the mining goals, which determine the part of Web to be mined: Web content mining, Web structure mining, and Web usage mining [234, 140]. Web content mining tries to discover valuable information from Web contents (i.e. Web documents). Generally, Web content is mainly referred to textual objects, thus, it is also alternatively termed as text mining sometimes [50]. Web structure mining involves in modeling Web sites in terms of linking structures. The mutual linkage information obtained could, in turn, be used to construct Web page communities or find relevant pages based on the similarity or relevance between two Web pages. A successful application addressing this topic is finding relevant Web pages through linkage analysis [120, 137, 67, 234, 184, 174]. Web usage mining tries to reveal the underlying access patterns from Web transaction or user session data that recorded in Web log files [238, 99]. Generally, Web users are usually performing their interest-driven visits by clicking one or more functional Web objects. They may exhibit different types of access interests associated with their navigational tasks during their surfing periods. Thus, employing data mining techniques on the observed usage data may lead to finding underlying usage pattern. In addition, capturing Web user access interest or pattern can, not only provide help for better understanding user navigational behavior, but also for efficiently improving Web site structure or design. This, furthermore, can be utilized to recommend or predict Web contents tailored and personalized to Web users who can benefit from obtaining more preferred information and reducing waiting time [146, 119].

Discovering the latent semantic space from Web data by using statistical learning algorithms is another recently emerging research topic in Web knowledge discovery. Similar to semantic Web, semantic Web mining is considered as a new branch of Web mining research [121]. The abstract Web semantics along with other intuitive Web data forms, such as Web textual, linkage and usage information constitute a multidimensional and comprehensive data space for Web data analysis.

By using Web mining techniques, Web research academia has achieved substantial success in Web research areas, such as retrieving the desirable and related information [184], creating good quality Web community [137, 274], extracting informative knowledge out of available information [223], capturing underlying usage pattern from Web observation data [140], recommending or recommending user customized information to offer better Internet service [238], and furthermore mining valuable business information from the common or individual customers' navigational behavior as well [146].

Although much work has been done in Web-based data management and a great amount of achievements have been made so far, there still remain many open research

problems to be solved in this area due to the fact of the distinctive characteristics of Web data, the complexity of Web data model, the diversity of various Web applications, the progress made in related research areas and the increased demands from Web users. How to efficiently and effectively address Web-based data management by using more advanced data processing techniques, thus, is becoming an active research topic that is full of many challenges.

1.3 Web Community and Social Network Analysis

1.3.1 Characteristics of Web Data

For the data on the Web, it has its own distinctive features from the data in conventional database management systems. Web data usually exhibits the following characteristics:

- The data on the Web is huge in amount. Currently, it is hard to estimate the exact data volume available on the Internet due to the exponential growth of Web data every day. For example, in 1994, one of the first Web search engines, the World Wide Web Worm (WWWW) had an index of 110,000 Web pages and Web accessible documents. As of November, 1997, the top search engines claim to index from 2 million (WebCrawler) to 100 million Web documents. The enormous volume of data on the Web makes it difficult to well handle Web data via traditional database techniques.
- The data on the Web is distributed and heterogeneous. Due to the essential property of Web being an interconnection of various nodes over the Internet, Web data is usually distributed across a wide range of computers or servers, which are located at different places around the world. Meanwhile, Web data is often exhibiting the intrinsic nature of multimedia, that is, in addition to textual information, which is mostly used to express contents; many other types of Web data, such as images, audio files and video slips are often included in a Web page. It requires the developed techniques for Web data processing with the ability of dealing with heterogeneity of multimedia data.
- The data on the Web is unstructured. There are, so far, no rigid and uniform data structures or schemas that Web pages should strictly follow, that are common requirements in conventional database management. Instead, Web designers are able to arbitrarily organize related information on the Web together in their own ways, as long as the information arrangement meets the basic layout requirements of Web documents, such as HTML format. Although Web pages in well-defined HTML format could contain some preliminary Web data structures, e.g. tags or anchors, these structural components, however, can primarily benefit the presentation quality of Web documents rather than reveal the semantics contained in Web documents. As a result, there is an increasing requirement to better deal with the unstructured nature of Web documents and extract the mutual relationships hidden in Web data for facilitating users to locate needed Web information or service.

- The data on the Web is dynamic. The implicit and explicit structure of Web data is updated frequently. Especially, due to different applications of Web-based data management systems, a variety of presentations of Web documents will be generated while contents resided in databases update. And dangling links and relocation problems will be produced when domain or file names change or disappear. This feature leads to frequent schema modifications of Web documents, which often suffer traditional information retrieval.

The aforementioned features indicate that Web data is a specific type of data different from the data resided in traditional database systems. As a result, there is an increasing demand to develop more advanced techniques to address Web information search and data management. The recently emerging Web community technology is a representative of new technical concepts that efficiently tackles the Web-based data management.

1.3.2 Web Community

Theoretically, Web Community is defined as an aggregation of Web objects in terms of Web pages or users, in which each object is "losely" related to the other under a certain distance space. Unlike the conventional database management in which data models and schemas are defined, a Web community, which is a set of Web-based objects (documents and users) that has its own logical structures, is another effective and efficient approach to reorganize Web-based objects, support information retrieval and implement various applications. Therefore, community centered Web data management systems provide more capabilities than database-centered ones in Web-based data management.

So far a large amount of research efforts have been contributed to the research of Web Community, and a great deal of successes have been achieved accordingly. According to the aims and purposes, these studies and developments are mainly about two aspects of Web data management, that is, how to accurately find the needed information on the Internet, i.e. Web information search, and how to efficiently and effectively manage and utilize the informative knowledge mined from the massive data on the Internet, i.e. Web data/knowledge management. For example, finding Web communities from a collected data source via linkage analysis is an active and hot topic in Web search and information filtering areas. In this case, a Web community is a Web page group, within which all members share similar hyperlink topology to a specific Web page. These discovered Web communities might be able to help users to find Web pages which are related to the query page in terms of hyperlink structures. In the scenario of e-commerce, market basket analysis is a very popular research problem in data mining, which aims to analyze the customer's behavior pattern during the online shopping process. Web usage mining through analyzing Web log files is proposed as an efficient analytical tool for business organizations to investigate various types of user navigational pattern of how customers access the e-commerce website. Here the Web communities expressed as categories of Web users represent the different customers' shopping behavior types.

1.3.3 Social Networking

Recently, with the popularity and development of innovative Web technologies, for example, semantic Web or Web 2.0, more and more advanced Web data based services and applications are emerging for Web users to easily generate and distribute Web contents, and conveniently share information in a collaborative environment. The core component of the second generation Web is Web-based communities and hosted services, such as social networking sites, wikis and folksonomies, which are characterized by the features of open-communication, decentralization of authority, and freedom to share and self-manage. These newly enhanced Web functionalities make it possible for Web users to share and locate the needed Web contents easily, to collaborate and interact with each other socially, and to realize knowledge utilization and management freely on the Web. For example, the social Web hosted service like *Myspace* and *Facebook* are becoming a global and influential information sharing and exchanging platform and data source in the world. As a result, Social Networks is becoming a newly emerging research topic in Web research although this term has appeared in social science, especially psychology in several decades ago. A social network is a representative of relationships existing within a community [276]. Social Networking provide us a useful means to study the mutual relationships and networked structures, often derived and expressed by collaborations amongst community peers or nodes, through theories developed in social network analysis and social computing [81, 117].

As we discussed, Web community analysis is to discover the aggregations of Web pages, users as well as co-clusters of Web objects. As a result, Web communities are always modeled as groups of pages and users, which can also be represented by various graphic expressions, for example, here the nodes denote the users, while the lines stand for the relationships between two users, such as pages commonly visited by these two users or email communications between senders and receivers. In other words, a Web community could be modeled as a network of users exchanging information or exhibiting common interest, that is, a social network. In this sense, the gap between Web community analysis and social network analysis is becoming closer and closer, many concepts and techniques used and developed in one area could be extended into the research area of the other.

In summary, with the prevalence and maturity of Web 2.0 technologies, the Web is becoming a useful platform and an influential source of data for individuals to share their information and express their opinions, and the collaboration or linking between various Web users is knitting as a community-centered social networking over the Web. From this viewing point, how to extend the current Web community analysis to a very massive data source to investigate the social behavior pattern or evaluation, or how to introduce the achievements from traditional social network analysis into Web data management to better interpret and understand the knowledge discovered, is bringing forward a huge amount of challenges that Web researchers and engineers have to face. Linking the two distinctive research areas, but with immanent underlying connection, and complementing the respective research strengths