Web Information Systems Engineering and Internet Technologies Book Series

Guandong Xu
Yanchun Zhang
Lin Li

Web Mining and Social Networking

Techniques and Applications

Springer

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