

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

Editorial Board:

Robin Chen, AT&T

Umeshwar Dayal, HP

Arun Iyengar, IBM

Keith Jeffery, Rutherford Appleton Lab

Xiaohua Jia, City University of Hong Kong

Yahiko Kambayashi† Kyoto University

Masaru Kitsuregawa, Tokyo University

Qing Li, City University of Hong Kong

Philip Yu, IBM

Hongjun Lu, HKUST

John Mylopoulos, University of Toronto

Erich Neuhold, IPSI

Tamer Ozsu, Waterloo University

Maria Orlowska, DSTC

Gultekin Ozsoyoglu, Case Western Reserve University

Michael Papazoglou, Tilburg University

Marek Rusinkiewicz, Telcordia Technology

Stefano Spaccapietra, EPFL

Vijay Varadharajan, Macquarie University

Marianne Winslett, University of Illinois at Urbana-Champaign

Xiaofang Zhou, University of Queensland

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

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

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

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

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

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 amount 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 with 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 I Foundation

1	Introduction	3
1.1	Background	3
1.2	Data Mining and Web Mining	5
1.3	Web Community and Social Network Analysis	7
1.3.1	Characteristics of Web Data	7
1.3.2	Web Community	8
1.3.3	Social Networking	9
1.4	Summary of Chapters	10
1.5	Audience of This Book	11
2	Theoretical Backgrounds	13
2.1	Web Data Model	13
2.2	Textual, Linkage and Usage Expressions	14
2.3	Similarity Functions	16
2.3.1	Correlation-based Similarity	17
2.3.2	Cosine-Based Similarity	17
2.4	Eigenvector, Principal Eigenvector	17
2.5	Singular Value Decomposition (SVD) of Matrix	19
2.6	Tensor Expression and Decomposition	20
2.7	Information Retrieval Performance Evaluation Metrics	22
2.7.1	Performance measures	22
2.7.2	Web Recommendation Evaluation Metrics	24
2.8	Basic Concepts in Social Networks	25
2.8.1	Basic Metrics of Social Network	25
2.8.2	Social Network over the Web	26
3	Algorithms and Techniques	29
3.1	Association Rule Mining	29
3.1.1	Association Rule Mining Problem	29

3.1.2	Basic Algorithms for Association Rule Mining	31
3.1.3	Sequential Pattern Mining	36
3.2	Supervised Learning	46
3.2.1	Nearest Neighbor Classifiers	46
3.2.2	Decision Tree	46
3.2.3	Bayesian Classifiers	49
3.2.4	Neural Networks Classifier	50
3.3	Unsupervised Learning	52
3.3.1	The k -Means Algorithm	52
3.3.2	Hierarchical Clustering	53
3.3.3	Density based Clustering	55
3.4	Semi-supervised Learning	56
3.4.1	Self-Training	56
3.4.2	Co-Training	57
3.4.3	Generative Models	58
3.4.4	Graph based Methods	59
3.5	Markov Models	59
3.5.1	Regular Markov Models	60
3.5.2	Hidden Markov Models	61
3.6	K-Nearest-Neighboring	62
3.7	Content-based Recommendation	62
3.8	Collaborative Filtering Recommendation	63
3.8.1	Memory-based collaborative recommendation	63
3.8.2	Model-based Recommendation	64
3.9	Social Network Analysis	64
3.9.1	Detecting Community Structure in Networks	64
3.9.2	The Evolution of Social Networks	67

Part II Web Mining: Techniques and Applications

4	Web Content Mining	71
4.1	Vector Space Model	71
4.2	Web Search	73
4.2.1	Activities on Web archiving	73
4.2.2	Web Crawling	74
4.2.3	Personalized Web Search	76
4.3	Feature Enrichment of Short Texts	77
4.4	Latent Semantic Indexing	79
4.5	Automatic Topic Extraction from Web Documents	80
4.5.1	Topic Models	80
4.5.2	Topic Models for Web Documents	83
4.5.3	Inference and Parameter Estimation	84
4.6	Opinion Search and Opinion Spam	84
4.6.1	Opinion Search	85

4.6.2	Opinion Spam	86
5	Web Linkage Mining	89
5.1	Web Search and Hyperlink	89
5.2	Co-citation and Bibliographic Coupling	90
5.2.1	Co-citation	90
5.2.2	Bibliographic Coupling	90
5.3	PageRank and HITS Algorithms	91
5.3.1	PageRank	91
5.3.2	HITS	93
5.4	Web Community Discovery	95
5.4.1	Bipartite Cores as Communities	96
5.4.2	Network Flow/Cut-based Notions of Communities	97
5.4.3	Web Community Chart	97
5.5	Web Graph Measurement and Modeling	100
5.5.1	Graph Terminologies	101
5.5.2	Power-law Distribution	101
5.5.3	Power-law Connectivity of the Web Graph	101
5.5.4	Bow-tie Structure of the Web Graph	102
5.6	Using Link Information for Web Page Classification	102
5.6.1	Using Web Structure for Classifying and Describing Web Pages	103
5.6.2	Using Implicit and Explicit Links for Web Page Classification	105
6	Web Usage Mining	109
6.1	Modeling Web User Interests using Clustering	109
6.1.1	Measuring Similarity of Interest for Clustering Web Users ..	109
6.1.2	Clustering Web Users using Latent Semantic Indexing	115
6.2	Web Usage Mining using Probabilistic Latent Semantic Analysis ..	118
6.2.1	Probabilistic Latent Semantic Analysis Model	118
6.2.2	Constructing User Access Pattern and Identifying Latent Factor with PLSA	120
6.3	Finding User Access Pattern via Latent Dirichlet Allocation Model .	124
6.3.1	Latent Dirichlet Allocation Model	124
6.3.2	Modeling User Navigational Task via LDA	128
6.4	Co-Clustering Analysis of weblogs using Bipartite Spectral Projection Approach	130
6.4.1	Problem Formulation	131
6.4.2	An Example of Usage Bipartite Graph	132
6.4.3	Clustering User Sessions and Web Pages	132
6.5	Web Usage Mining Applications	133
6.5.1	Mining Web Logs to Improve Website Organization	134
6.5.2	Clustering User Queries from Web logs for Related Query ..	137
6.5.3	Using Ontology-Based User Preferences to Improve Web Search	141

Part III Social Networking and Web Recommendation: Techniques and Applications

7	Extracting and Analyzing Web Social Networks	145
7.1	Extracting Evolution of Web Community from a Series of Web Archive	145
7.1.1	Types of Changes	146
7.1.2	Evolution Metrics	146
7.1.3	Web Archives and Graphs	148
7.1.4	Evolution of Web Community Charts	148
7.2	Temporal Analysis on Semantic Graph using Three-Way Tensor Decomposition	153
7.2.1	Background	153
7.2.2	Algorithms	155
7.2.3	Examples of Formed Community	156
7.3	Analysis of Communities and Their Evolutions in Dynamic Networks	157
7.3.1	Motivation	158
7.3.2	Problem Formulation	159
7.3.3	Algorithm	160
7.3.4	Community Discovery Examples	161
7.4	Socio-Sense: A System for Analyzing the Societal Behavior from Web Archive	161
7.4.1	System Overview	163
7.4.2	Web Structural Analysis	163
7.4.3	Web Temporal Analysis	165
7.4.4	Consumer Behavior Analysis	166
8	Web Mining and Recommendation Systems	169
8.1	User-based and Item-based Collaborative Filtering Recommender Systems	169
8.1.1	User-based Collaborative Filtering	170
8.1.2	Item-based Collaborative Filtering Algorithm	171
8.1.3	Performance Evaluation	174
8.2	A Hybrid User-based and Item-based Web Recommendation System	175
8.2.1	Problem Domain	175
8.2.2	Hybrid User and Item-based Approach	176
8.2.3	Experimental Observations	178
8.3	User Profiling for Web Recommendation Based on PLSA and LDA Model	178
8.3.1	Recommendation Algorithm based on PLSA Model	178
8.3.2	Recommendation Algorithm Based on LDA Model	181
8.4	Combing Long-Term Web Achieves and Logs for Web Query Recommendation	183

8.5	Combinational CF Approach for Personalized Community	
	Recommendation.....	185
8.5.1	CCF: Combinational Collaborative Filtering.....	186
8.5.2	C-U and C-D Baseline Models.....	186
8.5.3	CCF Model.....	187
9	Conclusions	189
9.1	Summary.....	189
9.2	Future Directions.....	191
	References	195