

Web Communities

Yanchun Zhang · Jeffrey Xu Yu · Jingyu Hou

Web Communities

Analysis and Construction

With 28 Figures

Authors

Yanchun Zhang

School of Computer Science and Mathematics
Victoria University of Technology
Ballarat Road, Footscray
PO Box 14428
MC 8001, Melbourne City, Australia
yzhang@csm.vu.edu.au

Jeffrey Xu Yu

Dept. of Systems Engineering and Engineering Management
Chinese University of Hong Kong
Shatin, N.T., Hong Kong, China
yu@se.cuhk.edu.hk

Jingyu Hou

School of Information Technology
Deakin University
Burwood, Victoria 3125, Australia
jingyu@deakin.edu.au

Library of Congress Control Number: 2005936102

ACM Computing Classification (1998): H.3, H.5

ISBN-10 3-540-27737-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-27737-8 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typeset by the authors using a Springer $\text{\TeX}$ macro package
Production: LE- $\text{\TeX}$ Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design: KunkelLopka Werbeagentur, Heidelberg

Printed on acid-free paper 45/3142/YL - 5 4 3 2 1 0

Dedication

To Jinli and Dana
From Yanchun

To Hannah, Michael and Stephen
From Jeffrey

To Huiming, Mingxi and Mingyi
From Jingyu

Contents

Preface.....	XI
1 Introduction.....	1
1.1 Background.....	1
1.2 Web Community	4
1.3 Outline of the Book	5
1.4 Audience of the Book.....	6
2 Preliminaries	7
2.1 Matrix Expression of Hyperlinks	7
2.2 Eigenvalue and Eigenvector of the Matrix	9
2.3 Matrix Norms and the Lipschitz Continuous Function	10
2.4 Singular Value Decomposition (SVD) of a Matrix	11
2.5 Similarity in Vector Space Models.....	14
2.6 Graph Theory Basics	14
2.7 Introduction to the Markov Model	15
3 HITS and Related Algorithms.....	17
3.1 Original HITS.....	17
3.2 The Stability Issues	20
3.3 Randomized HITS.....	22
3.4 Subspace HITS.....	23
3.5 Weighted HITS.....	24
3.6 The Vector Space Model (VSM).....	27
3.7 Cover Density Ranking (CDR)	29
3.8 In-depth Analysis of HITS.....	31
3.9 HITS Improvement	35
3.10 Noise Page Elimination Algorithm Based on SVD.....	38
3.11 SALSA (Stochastic algorithm)	43
4 PageRank Related Algorithms.....	49
4.1 The Original PageRank Algorithm.....	49
4.2 Probabilistic Combination of Link and Content Information	53
4.3 Topic-Sensitive PageRank	56

4.4	Quadratic Extrapolation.....	58
4.5	Exploring the Block Structure of the Web for Computing PageRank	60
4.6	Web Page Scoring Systems (WPSS)	64
4.7	The Voting Model	71
4.8	Using Non-Affiliated Experts to Rank Popular Topics	75
4.9	A Latent Linkage Information (LLI) Algorithm	79
5	Affinity and Co-Citation Analysis Approaches	85
5.1	Web Page Similarity Measurement	85
5.1.1	Page Source Construction	85
5.1.2	Page Weight Definition.....	87
5.1.3	Page Correlation Matrix.....	89
5.1.4	Page Similarity	92
5.2	Hierarchical Web Page Clustering	95
5.3	Matrix-Based Clustering Algorithms	97
5.3.1	Similarity Matrix Permutation.....	97
5.3.2	Clustering Algorithm from a Matrix Partition.....	99
5.3.3	Cluster-Overlapping Algorithm.....	101
5.4	Co-Citation Algorithms	104
5.4.1	Citation and Co-Citation Analysis.....	104
5.4.2	Extended Co-Citation Algorithms	106
6	Building a Web Community	111
6.1	Web Community	111
6.2	Small World Phenomenon on the Web	113
6.3	Trawling the Web.....	115
6.3.1	Finding Web Communities Based on Complete Directed Bipartite Graphs	117
6.4	From Complete Bipartite Graph to Dense Directed Bipartite Graph.....	118
6.4.1	The Algorithm.....	119
6.5	Maximum Flow Approaches.....	123
6.5.1	Maximum Flow and Minimum Cut.....	124
6.5.2	FLG Approach.....	125
6.5.3	IK Approach	129
6.6	Web Community Charts	133
6.6.1	The Algorithm	135
6.7	From Web Community Chart to Web Community Evolution ...	138
6.8	Uniqueness of a Web Community.....	141
7	Web Community Related Techniques	145

7.1	Web Community and Web Usage Mining.....	145
7.2	Discovering Web Communities Using Co-occurrence.....	147
7.3	Finding High-Level Web Communities	149
7.4	Web Community and Formal Concept Analysis.....	151
7.4.1	Formal Concept Analysis.....	152
7.4.2	From Concepts to Web Communities.....	152
7.5	Generating Web Graphs with Embedded Web Communities....	155
7.6	Modeling Web Communities Using Graph Grammars	157
7.7	Geographical Scopes of Web Resources	158
7.7.1	Two Conditions: Fraction and Uniformity.....	159
7.7.2	Geographical Scope Estimation	161
7.8	Discovering Unexpected Information from Competitors	161
7.9	Probabilistic Latent Semantic Analysis Approach	164
7.9.1	Usage Data and the PLSA Model	165
7.9.2	Discovering Usage-Based Web Page Categories	167
8	Conclusions.....	169
8.1	Summary.....	169
8.2	Future Directions.....	171
	References	173
	Index.....	181
	About the Authors.....	185

Preface

The rapid development of Web technology has made the World Wide Web an important and popular application platform for disseminating and searching information as well as conducting business. However, due to the lack of uniform schema for Web documents, the low precision of most search engines and the information explosion on the World Wide Web, the user is often flooded with a huge amount of information.

Unlike the conventional database management in which data models and schemas are defined, the 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. According to the practical requirements and concerned situations, the Web community would appear as different formats.

This book addresses the construction and analysis of various Web communities based on information available from Web, such as Web document content, hyperlinks, semantics and user access logs. Web community applications are another aspect emphasized in this book. Before presenting various algorithms, some preliminaries are provided for better understanding of the materials. Representative algorithms for constructing and analysing various Web communities are then presented and discussed. These algorithms, as well as their discussions, lead to various applications that are also presented in this book. Finally, this book summarizes the main work in Web community research and discusses future research in this area.

Acknowledgements

Our special thanks go to Mr. Guandong Xu and Mr. Yanan Hao for their assistance in preparing manuscripts of the book.