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



http://www.springer.com/978-1-4419-7734-2

Web Mining and Social Networking
Techniques and Applications
Xu, G.; Zhang, Y.; Li, L.
2011, XVII, 210 p., Hardcover
ISBN: 978-1-4419-7734-2