

Table of Contents

1. Introduction	1
1.1. What is the World Wide Web?	1
1.2. A Brief History of the Web and the Internet	2
1.3. Web Data Mining	4
1.3.1. What is Data Mining?	6
1.3.2. What is Web Mining?	7
1.4. Summary of Chapters	8
1.5. How to Read this Book	11
Bibliographic Notes	12
Bibliography	13

Part I: Data Mining Foundations

2. Association Rules and Sequential Patterns	17
2.1. Basic Concepts of Association Rules	17
2.2. Apriori Algorithm	20
2.2.1. Frequent Itemset Generation	20
2.2.2. Association Rule Generation	24
2.3. Data Formats for Association Rule Mining	26
2.4. Mining with Multiple Minimum Supports	26
2.4.1. Extended Model	28
2.4.2. Mining Algorithm	30
2.4.3. Rule Generation	35
2.5. Mining Class Association Rules	36
2.5.1. Problem Definition	36
2.5.2. Mining Algorithm	38
2.5.3. Mining with Multiple Minimum Supports	41

2.6. Basic Concepts of Sequential Patterns	41
2.7. Mining Sequential Patterns Based on GSP.....	43
2.7.1. GSP Algorithm	43
2.7.2. Mining with Multiple Minimum Supports	45
2.8. Mining Sequential Patterns Based on PrefixSpan...	49
2.8.1. PrefixSpan Algorithm	50
2.8.2. Mining with Multiple Minimum Supports	52
2.9. Generating Rules from Sequential Patterns	53
2.9.1. Sequential Rules	54
2.9.2. Label Sequential Rules	54
2.9.3. Class Sequential Rules	55
Bibliographic Notes	56
Bibliography	58
3. Supervised Learning	63
3.1. Basic Concepts	63
3.2. Decision Tree Induction	67
3.2.1. Learning Algorithm	70
3.2.2. Impurity Function	71
3.2.3. Handling of Continuous Attributes	75
3.2.4. Some Other Issues	76
3.3. Classifier Evaluation	79
3.3.1. Evaluation Methods	79
3.3.2. Precision, Recall, F-score and Breakeven Point ..	81
3.3.3. Receiver Operating Characteristic Curve	83
3.3.4. Lift Curve	86
3.4. Rule Induction	87
3.4.1. Sequential Covering	87
3.4.2. Rule Learning: Learn-One-Rule Function	90
3.4.3. Discussion	93
3.5. Classification Based on Associations	93
3.5.1. Classification Using Class Association Rules ...	94
3.5.2. Class Association Rules as Features	98
3.5.3. Classification Using Normal Association Rules..	99
3.6. Naïve Bayesian Classification	100
3.7. Naïve Bayesian Text Classification	103

3.7.1. Probabilistic Framework	104
3.7.2. Naïve Bayesian Model	105
3.7.3. Discussion	108
3.8. Support Vector Machines	109
3.8.1. Linear SVM: Separable Case	111
3.8.2. Linear SVM: Non-Separable Case	117
3.8.3. Nonlinear SVM: Kernel Functions	120
3.9. K-Nearest Neighbor Learning	124
3.10. Ensemble of Classifiers	126
3.10.1. Bagging	126
3.10.2. Boosting	126
Bibliographic Notes	128
Bibliography	129
4. Unsupervised Learning	133
4.1. Basic Concepts	133
4.2. K-means Clustering	136
4.2.1. K-means Algorithm	136
4.2.2. Disk Version of the K-means Algorithm	139
4.2.3. Strengths and Weaknesses	140
4.3. Representation of Clusters	144
4.3.1. Common Ways of Representing Clusters	145
4.3.2. Clusters of Arbitrary Shapes	146
4.4. Hierarchical Clustering	147
4.4.1. Single-Link Method	149
4.4.2. Complete-Link Method	149
4.4.3. Average-Link Method	150
4.4.4. Strengths and Weaknesses	150
4.5. Distance Functions	151
4.5.1. Numeric Attributes	151
4.5.2. Binary and Nominal Attributes	152
4.5.3. Text Documents	154
4.6. Data Standardization	155
4.7. Handling of Mixed Attributes	157
4.8. Which Clustering Algorithm to Use?	159
4.9. Cluster Evaluation	159
4.10. Discovering Holes and Data Regions	162

Bibliographic Notes	165
Bibliography	166
5. Partially Supervised Learning	171
5.1. Learning from Labeled and Unlabeled Examples ·	171
5.1.1. EM Algorithm with Naïve Bayesian Classification	173
5.1.2. Co-Training	176
5.1.3. Self-Training	178
5.1.4. Transductive Support Vector Machines	179
5.1.5. Graph-Based Methods	180
5.1.6. Discussion	183
5.2. Learning from Positive and Unlabeled Examples ·	184
5.2.1. Applications of PU Learning	185
5.2.2. Theoretical Foundation	187
5.2.3. Building Classifiers: Two-Step Approach	190
5.2.4. Building Classifiers: Biased-SVM	197
5.2.5. Building Classifiers: Probability Estimation	199
5.2.6. Discussion	201
<i>Appendix: Derivation of EM for Naïve Bayesian</i> Classification	202
Bibliographic Notes	204
Bibliography	206

Part II: Web Mining

6. Information Retrieval and Web Search	211
6.1. Basic Concepts of Information Retrieval	212
6.2. Information Retrieval Models	215
6.2.1. Boolean Model	216
6.2.2. Vector Space Model	217
6.2.3. Statistical Language Model	219
6.3. Relevance Feedback	220
6.4. Evaluation Measures	223

6.5. Text and Web Page Pre-Processing	227
6.5.1. Stopword Removal	227
6.5.2. Stemming	228
6.5.3. Other Pre-Processing Tasks for Text	228
6.5.4. Web Page Pre-Processing	229
6.5.5. Duplicate Detection	231
6.6. Inverted Index and Its Compression	232
6.6.1. Inverted Index	232
6.6.2. Search Using an Inverted Index	234
6.6.3. Index Construction	235
6.6.4. Index Compression	236
6.7. Latent Semantic Indexing	242
6.7.1. Singular Value Decomposition	243
6.7.2. Query and Retrieval	245
6.7.3. An Example	246
6.7.4. Discussion	249
6.8. Web Search	249
6.9. Meta-Search: Combining Multiple Rankings	252
6.9.1. Combination Using Similarity Scores	254
6.9.2. Combination Using Rank Positions	255
6.10. Web Spamming	257
6.10.1. Content Spamming	258
6.10.2. Link Spamming	259
6.10.3. Hiding Techniques	260
6.10.4. Combating Spam	261
Bibliographic Notes	263
Bibliography	264
7. Social Network Analysis	269
7.1. Social Network Analysis	270
7.1.1. Centrality	270
7.1.2. Prestige	273
7.2. Co-Citation and Bibliographic Coupling	275
7.2.1. Co-Citation	276
7.2.2. Bibliographic Coupling	277
7.3. PageRank	277
7.3.1. PageRank Algorithm	278
7.3.2. Strengths and Weaknesses of PageRank	285

7.3.3.	Timed PageRank and Recency Search	286
7.4.	HITS	288
7.4.1.	HITS Algorithm	289
7.4.2.	Finding Other Eigenvectors	291
7.4.3.	Relationships with Co-Citation and Bibliographic Coupling	292
7.4.4.	Strengths and Weaknesses of HITS	293
7.5.	Community Discovery	294
7.5.1.	Problem Definition	295
7.5.2.	Bipartite Core Communities	297
7.5.3.	Maximum Flow Communities	298
7.5.4.	Email Communities Based on Betweenness	301
7.5.5.	Overlapping Communities of Named Entities ..	303
	Bibliographic Notes	304
	Bibliography	305
8.	Web Crawling	311
8.1.	A Basic Crawler Algorithm	312
8.1.1.	Breadth-First Crawlers	313
8.1.2.	Preferential Crawlers	314
8.2.	Implementation Issues	315
8.2.1.	Fetching	315
8.2.2.	Parsing	316
8.2.3.	Stopword Removal and Stemming	318
8.2.4.	Link Extraction and Canonicalization	318
8.2.5.	Spider Traps	320
8.2.6.	Page Repository	321
8.2.7.	Concurrency	322
8.3.	Universal Crawlers	323
8.3.1.	Scalability	324
8.3.2.	Coverage vs. Freshness vs. Importance	326
8.4.	Focused Crawlers	327
8.5.	Topical Crawlers	330
8.5.1.	Topical Locality and Cues	332
8.5.2.	Best-First Variations	338
8.5.3.	Adaptation	341
8.6.	Evaluation	348
8.7.	Crawler Ethics and Conflicts	353

8.8. Some New Developments	356
Bibliographic Notes	358
Bibliography	359
9. Structured Data Extraction: Wrapper Generation ·	363
9.1 Preliminaries	364
9.1.1. Two Types of Data Rich Pages	364
9.1.2. Data Model	366
9.1.3. HTML Mark-Up Encoding of Data Instances	368
9.2. Wrapper Induction	370
9.2.1. Extraction from a Page	370
9.2.2. Learning Extraction Rules	373
9.2.3. Identifying Informative Examples	377
9.2.4. Wrapper Maintenance	378
9.3. Instance-Based Wrapper Learning	378
9.4. Automatic Wrapper Generation: Problems	381
9.4.1. Two Extraction Problems	382
9.4.2. Patterns as Regular Expressions	383
9.5. String Matching and Tree Matching	384
9.5.1. String Edit Distance	384
9.5.2. Tree Matching	386
9.6. Multiple Alignment	390
9.6.1. Center Star Method	390
9.6.2. Partial Tree Alignment	391
9.7. Building DOM Trees	396
9.8. Extraction Based on a Single List Page:	
Flat Data Records	397
9.8.1. Two Observations about Data Records	398
9.8.2. Mining Data Regions	399
9.8.3. Identifying Data Records in Data Regions	404
9.8.4. Data Item Alignment and Extraction	405
9.8.5. Making Use of Visual Information	406
9.8.6. Some Other Techniques	406
9.9. Extraction Based on a Single List Page:	
Nested Data Records	407
9.10. Extraction Based on Multiple Pages	413
9.10.1. Using Techniques in Previous Sections	413

XVIII Table of Contents

9.10.2. RoadRunner Algorithm	414
9.11. Some Other Issues	415
9.11.1. Extraction from Other Pages	416
9.11.2. Disjunction or Optional	416
9.11.3. A Set Type or a Tuple Type	417
9.11.4. Labeling and Integration	418
9.11.5. Domain Specific Extraction	418
9.12. Discussion	419
Bibliographic Notes	419
Bibliography	421
10. Information Integration	425
10.1. Introduction to Schema Matching	426
10.2. Pre-Processing for Schema Matching	428
10.3. Schema-Level Matching	429
10.3.1. Linguistic Approaches	429
10.3.2. Constraint Based Approaches	430
10.4. Domain and Instance-Level Matching	431
10.5. Combining Similarities	434
10.6. 1:m Match	435
10.7. Some Other Issues	436
10.7.1. Reuse of Previous Match Results	436
10.7.2. Matching a Large Number of Schemas	437
10.7.3. Schema Match Results	437
10.7.4. User Interactions	438
10.8. Integration of Web Query Interfaces	438
10.8.1. A Clustering Based Approach	441
10.8.2. A Correlation Based Approach	444
10.8.3. An Instance Based Approach	447
10.9. Constructing a Unified Global Query Interface ...	450
10.9.1. Structural Appropriateness and the Merge Algorithm	451
10.9.2. Lexical Appropriateness	453
10.9.3. Instance Appropriateness	454
Bibliographic Notes	454
Bibliography	455

11. Opinion Mining and Sentiment Analysis	459
11.1. The Problem of Opinion Mining	460
11.1.1. Problem Definitions	460
11.1.2. Aspect-Based Opinion Summary	467
11.2. Document Sentiment Classification	469
11.2.1. Classification Based on Supervised Learning	470
11.2.2. Classification Based on Unsupervised Learning	472
11.3. Sentence Subjectivity and Sentiment Classification	474
11.4. Opinion Lexicon Expansion	477
11.5. Aspect-Based Opinion Mining	480
11.5.1. Aspect Sentiment Classification	481
11.5.2. Basic Rules of Opinions	483
11.5.3. Aspect Extraction	486
11.5.4. Simultaneous Opinion Lexicon Expansion and Aspect Extraction	490
11.6. Mining Comparative Opinions	493
11.6.1. Problem Definitions	493
11.6.2. Identification of Comparative Sentences	495
11.6.3. Identification of Preferred Entities	496
11.7. Some Other Problems	498
11.8. Opinion Search and Retrieval	503
11.9. Opinion Spam Detection	506
11.9.1. Types of Spam and Spammers	506
11.9.2. Hiding Techniques	508
11.9.3. Spam Detection Based on Supervised Learning	509
11.9.4. Spam Detection Based on Abnormal Behaviors	511
11.9.5. Group Spam Detection.....	513
11.10. Utility of Reviews	514
Bibliographic Notes	515
Bibliography	517
12. Web Usage Mining	527
12.1. Data Collection and Pre-Processing	528

12.1.1. Sources and Types of Data	530
12.1.2. Key Elements of Web Usage Data Pre-Processing	533
12.2. Data Modeling for Web Usage Mining	540
12.3. Discovery and Analysis of Web Usage Patterns .	544
12.3.1. Session and Visitor Analysis	544
12.3.2. Cluster Analysis and Visitor Segmentation ...	545
12.3.3. Association and Correlation Analysis	549
12.3.4. Analysis of Sequential and Navigational Patterns	550
2.3.5. Classification and Prediction based on Web User Transactions	554
12.4. Recommender Systems and Collaborative Filtering	555
12.4.1. The Recommendation Problem	556
12.4.2. Content-Based Recommendation	557
12.4.3. Collaborative Filtering: <i>K</i> -Nearest Neighbor (KNN)	559
12.4.4. Collaborative Filtering: Using Association Rules	561
12.4.5. Collaborative Filtering: Matrix Factorization ...	565
12.5. Query Log Mining	571
12.5.1. Data Sources, Characteristics, and Challenges..	573
12.5.2. Query Log Data Preparation	574
12.5.3. Query Log Data Models	577
12.5.4. Query Log Feature Extraction	582
12.5.5. Query Log Mining Applications	583
12.5.6. Query Log Mining Methods	586
12.6. Computational Advertising	589
12.7. Discussion and Outlook	593
Bibliographic Notes	593
Bibliography	594
Subject Index	605



<http://www.springer.com/978-3-642-19459-7>

Web Data Mining

Exploring Hyperlinks, Contents, and Usage Data

Liu, B.

2011, XX, 624 p., Hardcover

ISBN: 978-3-642-19459-7