

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?	6
1.4. Summary of Chapters	8
1.5. How to Read this Book	11
Bibliographic Notes	12

Part I: Data Mining Foundations

2. Association Rules and Sequential Patterns	13
2.1. Basic Concepts of Association Rules	13
2.2. Apriori Algorithm	16
2.2.1. Frequent Itemset Generation	16
2.2.2. Association Rule Generation	20
2.3. Data Formats for Association Rule Mining	22
2.4. Mining with Multiple Minimum Supports	22
2.4.1. Extended Model	24
2.4.2. Mining Algorithm	26
2.4.3. Rule Generation	31
2.5. Mining Class Association Rules	32
2.5.1. Problem Definition	32
2.5.2. Mining Algorithm	34
2.5.3. Mining with Multiple Minimum Supports	37

VI Table of Contents

2.6. Basic Concepts of Sequential Patterns	37
2.7. Mining Sequential Patterns Based on GSP.....	39
2.7.1. GSP Algorithm	39
2.7.2. Mining with Multiple Minimum Supports	41
2.8. Mining Sequential Patterns Based on PrefixSpan	45
2.8.1. PrefixSpan Algorithm	46
2.8.2. Mining with Multiple Minimum Supports	48
2.9. Generating Rules from Sequential Patterns.....	49
2.9.1. Sequential Rules	50
2.9.2. Label Sequential Rules	50
2.9.3. Class Sequential Rules	51
Bibliographic Notes	52
3. Supervised Learning	55
3.1. Basic Concepts	55
3.2. Decision Tree Induction	59
3.2.1. Learning Algorithm	62
3.2.2. Impurity Function	63
3.2.3. Handling of Continuous Attributes	67
3.2.4. Some Other Issues	68
3.3. Classifier Evaluation	71
3.3.1. Evaluation Methods	71
3.3.2. Precision, Recall, F-score and Breakeven Point	73
3.4. Rule Induction	75
3.4.1. Sequential Covering	75
3.4.2. Rule Learning: Learn-One-Rule Function.....	78
3.4.3. Discussion	81
3.5. Classification Based on Associations	81
3.5.1. Classification Using Class Association Rules	82
3.5.2. Class Association Rules as Features	86
3.5.3. Classification Using Normal Association Rules	86
3.6. Naïve Bayesian Classification	87
3.7. Naïve Bayesian Text Classification	91
3.7.1. Probabilistic Framework	92
3.7.2. Naïve Bayesian Model	93
3.7.3. Discussion	96
3.8. Support Vector Machines	97
3.8.1. Linear SVM: Separable Case	99

3.8.2. Linear SVM: Non-Separable Case	105
3.8.3. Nonlinear SVM: Kernel Functions	108
3.9. K-Nearest Neighbor Learning	112
3.10. Ensemble of Classifiers	113
3.10.1. Bagging	114
3.10.2. Boosting	114
Bibliographic Notes	115
4. Unsupervised Learning	117
4.1. Basic Concepts	117
4.2. K-means Clustering	120
4.2.1. K-means Algorithm	120
4.2.2. Disk Version of the K-means Algorithm	123
4.2.3. Strengths and Weaknesses	124
4.3. Representation of Clusters	128
4.3.1. Common Ways of Representing Clusters	129
4.3.2. Clusters of Arbitrary Shapes	130
4.4. Hierarchical Clustering	131
4.4.1. Single-Link Method	133
4.4.2. Complete-Link Method	133
4.4.3. Average-Link Method	134
4.4.4. Strengths and Weaknesses	134
4.5. Distance Functions	135
4.5.1. Numeric Attributes	135
4.5.2. Binary and Nominal Attributes	136
4.5.3. Text Documents	138
4.6. Data Standardization	139
4.7. Handling of Mixed Attributes	141
4.8. Which Clustering Algorithm to Use?	143
4.9. Cluster Evaluation	143
4.10. Discovering Holes and Data Regions	146
Bibliographic Notes	149
5. Partially Supervised Learning	151
5.1. Learning from Labeled and Unlabeled Examples	151
5.1.1. EM Algorithm with Naïve Bayesian Classification	153

VIII Table of Contents

5.1.2. Co-Training	156
5.1.3. Self-Training	158
5.1.4. Transductive Support Vector Machines	159
5.1.5. Graph-Based Methods	160
5.1.6. Discussion	164
5.2. Learning from Positive and Unlabeled Examples	165
5.2.1. Applications of PU Learning	165
5.2.2. Theoretical Foundation	168
5.2.3. Building Classifiers: Two-Step Approach	169
5.2.4. Building Classifiers: Direct Approach	175
5.2.5. Discussion	178
<i>Appendix: Derivation of EM for Naïve Bayesian Classification</i> ..	179
Bibliographic Notes	181

Part II: Web Mining

6. Information Retrieval and Web Search	183
6.1. Basic Concepts of Information Retrieval	184
6.2. Information Retrieval Models	187
6.2.1. Boolean Model	188
6.2.2. Vector Space Model	188
6.2.3. Statistical Language Model	191
6.3. Relevance Feedback	192
6.4. Evaluation Measures	195
6.5. Text and Web Page Pre-Processing	199
6.5.1. Stopword Removal	199
6.5.2. Stemming	200
6.5.3. Other Pre-Processing Tasks for Text	200
6.5.4. Web Page Pre-Processing	201
6.5.5. Duplicate Detection	203
6.6. Inverted Index and Its Compression	204
6.6.1. Inverted Index	204
6.6.2. Search Using an Inverted Index	206
6.6.3. Index Construction	207
6.6.4. Index Compression	209

6.7. Latent Semantic Indexing	215
6.7.1. Singular Value Decomposition	215
6.7.2. Query and Retrieval	218
6.7.3. An Example	219
6.7.4. Discussion	221
6.8. Web Search	222
6.9. Meta-Search: Combining Multiple Rankings	225
6.9.1. Combination Using Similarity Scores	226
6.9.2. Combination Using Rank Positions	227
6.10. Web Spamming	229
6.10.1. Content Spamming	230
6.10.2. Link Spamming	231
6.10.3. Hiding Techniques	233
6.10.4. Combating Spam	234
Bibliographic Notes	235
7. Link Analysis	237
7.1. Social Network Analysis	238
7.1.1. Centrality	238
7.1.2. Prestige	241
7.2. Co-Citation and Bibliographic Coupling	243
7.2.1. Co-Citation	244
7.2.2. Bibliographic Coupling	245
7.3. PageRank	245
7.3.1. PageRank Algorithm	246
7.3.2. Strengths and Weaknesses of PageRank	253
7.3.3. Timed PageRank	254
7.4. HITS	255
7.4.1. HITS Algorithm	256
7.4.2. Finding Other Eigenvectors	259
7.4.3. Relationships with Co-Citation and Bibliographic Coupling	259
7.4.4. Strengths and Weaknesses of HITS	260
7.5. Community Discovery	261
7.5.1. Problem Definition	262
7.5.2. Bipartite Core Communities	264
7.5.3. Maximum Flow Communities	265
7.5.4. Email Communities Based on Betweenness	268
7.5.5. Overlapping Communities of Named Entities	270

Bibliographic Notes	271
8. Web Crawling	273
8.1. A Basic Crawler Algorithm	274
8.1.1. Breadth-First Crawlers	275
8.1.2. Preferential Crawlers	276
8.2. Implementation Issues	277
8.2.1. Fetching	277
8.2.2. Parsing	278
8.2.3. Stopword Removal and Stemming	280
8.2.4. Link Extraction and Canonicalization	280
8.2.5. Spider Traps	282
8.2.6. Page Repository	283
8.2.7. Concurrency	284
8.3. Universal Crawlers	285
8.3.1. Scalability	286
8.3.2. Coverage vs Freshness vs Importance	288
8.4. Focused Crawlers	289
8.5. Topical Crawlers	292
8.5.1. Topical Locality and Cues	294
8.5.2. Best-First Variations	300
8.5.3. Adaptation	303
8.6. Evaluation	310
8.7. Crawler Ethics and Conflicts	315
8.8. Some New Developments	318
Bibliographic Notes	320
9. Structured Data Extraction: Wrapper Generation ·	323
9.1 Preliminaries	324
9.1.1. Two Types of Data Rich Pages	324
9.1.2. Data Model	326
9.1.3. HTML Mark-Up Encoding of Data Instances	328
9.2. Wrapper Induction	330
9.2.1. Extraction from a Page	330
9.2.2. Learning Extraction Rules	333
9.2.3. Identifying Informative Examples	337
9.2.4. Wrapper Maintenance	338

9.3. Instance-Based Wrapper Learning	338
9.4. Automatic Wrapper Generation: Problems	341
9.4.1. Two Extraction Problems	342
9.4.2. Patterns as Regular Expressions	343
9.5. String Matching and Tree Matching	344
9.5.1. String Edit Distance	344
9.5.2. Tree Matching	346
9.6. Multiple Alignment	350
9.6.1. Center Star Method	350
9.6.2. Partial Tree Alignment	351
9.7. Building DOM Trees	356
9.8. Extraction Based on a Single List Page:	
Flat Data Records	357
9.8.1. Two Observations about Data Records	358
9.8.2. Mining Data Regions	359
9.8.3. Identifying Data Records in Data Regions	364
9.8.4. Data Item Alignment and Extraction	365
9.8.5. Making Use of Visual Information	366
9.8.6. Some Other Techniques	366
9.9. Extraction Based on a Single List Page:	
Nested Data Records	367
9.10. Extraction Based on Multiple Pages	373
9.10.1. Using Techniques in Previous Sections	373
9.10.2. RoadRunner Algorithm	374
9.11. Some Other Issues	375
9.11.1. Extraction from Other Pages	375
9.11.2. Disjunction or Optional	376
9.11.3. A Set Type or a Tuple Type	377
9.11.4. Labeling and Integration	378
9.11.5. Domain Specific Extraction	378
9.12. Discussion	379
Bibliographic Notes	379
10. Information Integration	381
10.1. Introduction to Schema Matching	382
10.2. Pre-Processing for Schema Matching	384
10.3. Schema-Level Match	385

XII Table of Contents

10.3.1. Linguistic Approaches	385
10.3.2. Constraint Based Approaches	386
10.4. Domain and Instance-Level Matching	387
10.5. Combining Similarities	390
10.6. 1:m Match	391
10.7. Some Other Issues	392
10.7.1. Reuse of Previous Match Results	392
10.7.2. Matching a Large Number of Schemas	393
10.7.3. Schema Match Results	393
10.7.4. User Interactions	394
10.8. Integration of Web Query Interfaces	394
10.8.1. A Clustering Based Approach	397
10.8.2. A Correlation Based Approach	400
10.8.3. An Instance Based Approach	403
10.9. Constructing a Unified Global Query Interface	406
10.9.1. Structural Appropriateness and the Merge Algorithm	406
10.9.2. Lexical Appropriateness	408
10.9.3. Instance Appropriateness	409
Bibliographic Notes	410
11. Opinion Mining	411
11.1. Sentiment Classification	412
11.1.1. Classification Based on Sentiment Phrases	413
11.1.2. Classification Using Text Classification Methods	415
11.1.3. Classification Using a Score Function	416
11.2. Feature-Based Opinion Mining and Summarization	417
11.2.1. Problem Definition	418
11.2.2. Object Feature Extraction	424
11.2.3. Feature Extraction from Pros and Cons of Format 1	425
11.2.4. Feature Extraction from Reviews of of Formats 2 and 3	429
11.2.5. Opinion Orientation Classification	430
11.3. Comparative Sentence and Relation Mining	432
11.3.1. Problem Definition	433
11.3.2. Identification of Gradable Comparative Sentences	435

11.3.3. Extraction of Comparative Relations	437
11.4. Opinion Search	439
11.5. Opinion Spam	441
11.5.1. Objectives and Actions of Opinion Spamming	441
11.5.2. Types of Spam and Spammers	442
11.5.3. Hiding Techniques	443
11.5.4. Spam Detection	444
Bibliographic Notes	446
12. Web Usage Mining	449
12.1. Data Collection and Pre-Processing	450
12.1.1 Sources and Types of Data	452
12.1.2 Key Elements of Web Usage Data Pre-Processing	455
12.2 Data Modeling for Web Usage Mining	462
12.3 Discovery and Analysis of Web Usage Patterns	466
12.3.1. Session and Visitor Analysis	466
12.3.2. Cluster Analysis and Visitor Segmentation	467
12.3.3 Association and Correlation Analysis	471
12.3.4 Analysis of Sequential and Navigational Patterns	475
12.3.5. Classification and Prediction Based on Web User Transactions	479
12.4. Discussion and Outlook	482
Bibliographic Notes	482
References	485
Index	517