

# Contents

1		
An Introduction to Text Mining		1
<i>Charu C. Aggarwal and ChengXiang Zhai</i>		
1. Introduction		1
2. Algorithms for Text Mining		4
3. Future Directions		8
References		10
2		
Information Extraction from Text		11
<i>Jing Jiang</i>		
1. Introduction		11
2. Named Entity Recognition		15
2.1 Rule-based Approach		16
2.2 Statistical Learning Approach		17
3. Relation Extraction		22
3.1 Feature-based Classification		23
3.2 Kernel Methods		26
3.3 Weakly Supervised Learning Methods		29
4. Unsupervised Information Extraction		30
4.1 Relation Discovery and Template Induction		31
4.2 Open Information Extraction		32
5. Evaluation		33
6. Conclusions and Summary		34
References		35
3		
A Survey of Text Summarization Techniques		43
<i>Ani Nenkova and Kathleen McKeown</i>		
1. How do Extractive Summarizers Work?		44
2. Topic Representation Approaches		46
2.1 Topic Words		46
2.2 Frequency-driven Approaches		48
2.3 Latent Semantic Analysis		52
2.4 Bayesian Topic Models		53
2.5 Sentence Clustering and Domain-dependent Topics		55
3. Influence of Context		56
3.1 Web Summarization		57
3.2 Summarization of Scientific Articles		58

3.3	Query-focused Summarization	58
3.4	Email Summarization	59
4.	Indicator Representations and Machine Learning for Summarization	60
4.1	Graph Methods for Sentence Importance	60
4.2	Machine Learning for Summarization	62
5.	Selecting Summary Sentences	64
5.1	Greedy Approaches: Maximal Marginal Relevance	64
5.2	Global Summary Selection	65
6.	Conclusion	66
	References	66
4		
	A Survey of Text Clustering Algorithms	77
	<i>Charu C. Aggarwal and ChengXiang Zhai</i>	
1.	Introduction	77
2.	Feature Selection and Transformation Methods for Text Clustering	81
2.1	Feature Selection Methods	81
2.2	LSI-based Methods	84
2.3	Non-negative Matrix Factorization	86
3.	Distance-based Clustering Algorithms	89
3.1	Agglomerative and Hierarchical Clustering Algorithms	90
3.2	Distance-based Partitioning Algorithms	92
3.3	A Hybrid Approach: The Scatter-Gather Method	94
4.	Word and Phrase-based Clustering	99
4.1	Clustering with Frequent Word Patterns	100
4.2	Leveraging Word Clusters for Document Clusters	102
4.3	Co-clustering Words and Documents	103
4.4	Clustering with Frequent Phrases	105
5.	Probabilistic Document Clustering and Topic Models	107
6.	Online Clustering with Text Streams	110
7.	Clustering Text in Networks	115
8.	Semi-Supervised Clustering	118
9.	Conclusions and Summary	120
	References	121
5		
	Dimensionality Reduction and Topic Modeling	129
	<i>Steven P. Crain, Ke Zhou, Shuang-Hong Yang and Hongyuan Zha</i>	
1.	Introduction	130
1.1	The Relationship Between Clustering, Dimension Reduction and Topic Modeling	131
1.2	Notation and Concepts	132
2.	Latent Semantic Indexing	133
2.1	The Procedure of Latent Semantic Indexing	134
2.2	Implementation Issues	135
2.3	Analysis	137
3.	Topic Models and Dimension Reduction	139
3.1	Probabilistic Latent Semantic Indexing	140
3.2	Latent Dirichlet Allocation	142
4.	Interpretation and Evaluation	148

4.1	Interpretation	148
4.2	Evaluation	149
4.3	Parameter Selection	150
4.4	Dimension Reduction	150
5.	Beyond Latent Dirichlet Allocation	151
5.1	Scalability	151
5.2	Dynamic Data	151
5.3	Networked Data	152
5.4	Adapting Topic Models to Applications	154
6.	Conclusion	155
	References	156
6		
	A Survey of Text Classification Algorithms	163
	<i>Charu C. Aggarwal and ChengXiang Zhai</i>	
1.	Introduction	163
2.	Feature Selection for Text Classification	167
2.1	Gini Index	168
2.2	Information Gain	169
2.3	Mutual Information	169
2.4	$\chi^2$ -Statistic	170
2.5	Feature Transformation Methods: Supervised LSI	171
2.6	Supervised Clustering for Dimensionality Reduction	172
2.7	Linear Discriminant Analysis	173
2.8	Generalized Singular Value Decomposition	175
2.9	Interaction of Feature Selection with Classification	175
3.	Decision Tree Classifiers	176
4.	Rule-based Classifiers	178
5.	Probabilistic and Naive Bayes Classifiers	181
5.1	Bernoulli Multivariate Model	183
5.2	Multinomial Distribution	188
5.3	Mixture Modeling for Text Classification	190
6.	Linear Classifiers	193
6.1	SVM Classifiers	194
6.2	Regression-Based Classifiers	196
6.3	Neural Network Classifiers	197
6.4	Some Observations about Linear Classifiers	199
7.	Proximity-based Classifiers	200
8.	Classification of Linked and Web Data	203
9.	Meta-Algorithms for Text Classification	209
9.1	Classifier Ensemble Learning	209
9.2	Data Centered Methods: Boosting and Bagging	210
9.3	Optimizing Specific Measures of Accuracy	211
10.	Conclusions and Summary	213
	References	213
7		
	Transfer Learning for Text Mining	223
	<i>Weike Pan, Erheng Zhong and Qiang Yang</i>	
1.	Introduction	224
2.	Transfer Learning in Text Classification	225
2.1	Cross Domain Text Classification	225

2.2	Instance-based Transfer	231
2.3	Cross-Domain Ensemble Learning	232
2.4	Feature-based Transfer Learning for Document Classification	235
3.	Heterogeneous Transfer Learning	239
3.1	Heterogeneous Feature Space	241
3.2	Heterogeneous Label Space	243
3.3	Summary	244
4.	Discussion	245
5.	Conclusions	246
	References	247
8		
	Probabilistic Models for Text Mining	259
	<i>Yizhou Sun, Hongbo Deng and Jiawei Han</i>	
1.	Introduction	260
2.	Mixture Models	261
2.1	General Mixture Model Framework	262
2.2	Variations and Applications	263
2.3	The Learning Algorithms	266
3.	Stochastic Processes in Bayesian Nonparametric Models	269
3.1	Chinese Restaurant Process	269
3.2	Dirichlet Process	270
3.3	Pitman-Yor Process	274
3.4	Others	275
4.	Graphical Models	275
4.1	Bayesian Networks	276
4.2	Hidden Markov Models	278
4.3	Markov Random Fields	282
4.4	Conditional Random Fields	285
4.5	Other Models	286
5.	Probabilistic Models with Constraints	287
6.	Parallel Learning Algorithms	288
7.	Conclusions	289
	References	290
9		
	Mining Text Streams	297
	<i>Charu C. Aggarwal</i>	
1.	Introduction	297
2.	Clustering Text Streams	299
2.1	Topic Detection and Tracking in Text Streams	307
3.	Classification of Text Streams	312
4.	Evolution Analysis in Text Streams	316
5.	Conclusions	317
	References	318
10		
	Translingual Mining from Text Data	323
	<i>Jian-Yun Nie, Jianfeng Gao and Guihong Cao</i>	
1.	Introduction	324
2.	Traditional Translingual Text Mining – Machine Translation	325

2.1	SMT and Generative Translation Models	325
2.2	Word-Based Models	327
2.3	Phrase-Based Models	329
2.4	Syntax-Based Models	333
3.	Automatic Mining of Parallel texts	336
3.1	Using Web structure	337
3.2	Matching parallel pages	339
4.	Using Translation Models in CLIR	341
5.	Collecting and Exploiting Comparable Texts	344
6.	Selecting Parallel Sentences, Phrases and Translation Words	347
7.	Mining Translingual Relations From Monolingual Texts	349
8.	Mining using hyperlinks	351
9.	Conclusions and Discussions	353
	References	354
11		
	Text Mining in Multimedia	361
	<i>Zheng-Jun Zha, Meng Wang, Jialie Shen and Tat-Seng Chua</i>	
1.	Introduction	362
2.	Surrounding Text Mining	364
3.	Tag Mining	366
3.1	Tag Ranking	366
3.2	Tag Refinement	367
3.3	Tag Information Enrichment	369
4.	Joint Text and Visual Content Mining	370
4.1	Visual Re-ranking	371
5.	Cross Text and Visual Content Mining	374
6.	Summary and Open Issues	377
	References	379
12		
	Text Analytics in Social Media	385
	<i>Xia Hu and Huan Liu</i>	
1.	Introduction	385
2.	Distinct Aspects of Text in Social Media	388
2.1	A General Framework for Text Analytics	388
2.2	Time Sensitivity	390
2.3	Short Length	391
2.4	Unstructured Phrases	392
2.5	Abundant Information	393
3.	Applying Text Analytics to Social Media	393
3.1	Event Detection	393
3.2	Collaborative Question Answering	395
3.3	Social Tagging	397
3.4	Bridging the Semantic Gap	398
3.5	Exploiting the Power of Abundant Information	399
3.6	Related Efforts	401
4.	An Illustrative Example	402
4.1	Seed Phrase Extraction	402
4.2	Semantic Feature Generation	404
4.3	Feature Space Construction	406
5.	Conclusion and Future Work	407
	References	408

13

A Survey of Opinion Mining and Sentiment Analysis	415
<i>Bing Liu and Lei Zhang</i>	
1. The Problem of Opinion Mining	416
1.1 Opinion Definition	416
1.2 Aspect-Based Opinion Summary	420
2. Document Sentiment Classification	422
2.1 Classification based on Supervised Learning	422
2.2 Classification based on Unsupervised Learning	424
3. Sentence Subjectivity and Sentiment Classification	426
4. Opinion Lexicon Expansion	429
4.1 Dictionary based approach	429
4.2 Corpus-based approach and sentiment consistency	430
5. Aspect-Based Sentiment Analysis	432
5.1 Aspect Sentiment Classification	433
5.2 Basic Rules of Opinions	434
5.3 Aspect Extraction	438
5.4 Simultaneous Opinion Lexicon Expansion and Aspect Extraction	440
6. Mining Comparative Opinions	441
7. Some Other Problems	444
8. Opinion Spam Detection	447
8.1 Spam Detection Based on Supervised Learning	448
8.2 Spam Detection Based on Abnormal Behaviors	449
8.3 Group Spam Detection	450
9. Utility of Reviews	451
10. Conclusions	452
References	453

14

Biomedical Text Mining: A Survey of Recent Progress	465
<i>Matthew S. Simpson and Dina Demner-Fushman</i>	
1. Introduction	466
2. Resources for Biomedical Text Mining	467
2.1 Corpora	467
2.2 Annotation	469
2.3 Knowledge Sources	470
2.4 Supporting Tools	471
3. Information Extraction	472
3.1 Named Entity Recognition	473
3.2 Relation Extraction	478
3.3 Event Extraction	482
4. Summarization	484
5. Question Answering	488
5.1 Medical Question Answering	489
5.2 Biological Question Answering	491
6. Literature-Based Discovery	492
7. Conclusion	495
References	496

Index

519