

Contents

Preface	xi
Contributors	xiii
I Clustering and Classification	1
1 Cluster-Preserving Dimension Reduction Methods for Efficient Classification of Text Data	3
Peg Howland and Haesun Park	
1.1 Introduction	3
1.2 Dimension Reduction in the Vector Space Model	4
1.3 A Method Based on an Orthogonal Basis of Centroids	5
1.3.1 Relationship to a Method from Factor Analysis	7
1.4 Discriminant Analysis and Its Extension for Text Data	8
1.4.1 Generalized Singular Value Decomposition	10
1.4.2 Extension of Discriminant Analysis	11
1.4.3 Equivalence for Various S_1 and S_2	14
1.5 Trace Optimization Using an Orthogonal Basis of Centroids . .	16
1.6 Document Classification Experiments	17
1.7 Conclusion	19
References	22
2 Automatic Discovery of Similar Words	25
Pierre P. Senellart and Vincent D. Blondel	
2.1 Introduction	25
2.2 Discovery of Similar Words from a Large Corpus	26
2.2.1 A Document Vector Space Model	27
2.2.2 A Thesaurus of Infrequent Words	28
2.2.3 The SEXTANT System	29
2.2.4 How to Deal with the Web	32
2.3 Discovery of Similar Words in a Dictionary	33

7.5	Implementation	166
7.6	Experimental Results	166
7.7	Further Work	167
7.8	Summary and Conclusion	168
	References	168

III Trend Detection 171

8 Trend and Behavior Detection from Web Queries 173

Peiling Wang, Jennifer Bownas, and Michael W. Berry

8.1	Introduction	173
8.2	Query Data and Analysis	174
	8.2.1 Descriptive Statistics of Web Queries	175
	8.2.2 Trend Analysis of Web Searching	176
8.3	Zipf's Law	178
	8.3.1 Natural Logarithm Transformations	178
	8.3.2 Piecewise Trendlines	179
8.4	Vocabulary Growth	179
8.5	Conclusions and Further Studies	181
	References	182

9 A Survey of Emerging Trend Detection in Textual Data Mining 185

April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps

9.1	Introduction	186
9.2	ETD Systems	187
	9.2.1 Technology Opportunities Analysis (TOA)	189
	9.2.2 CIMEL: Constructive, Collaborative Inquiry-Based Multimedia E-Learning	191
	9.2.3 TimeMines	195
	9.2.4 New Event Detection	199
	9.2.5 ThemeRiver™	201
	9.2.6 PatentMiner	204
	9.2.7 HDDI™	207
	9.2.8 Other Related Work	211
9.3	Commercial Software Overview	212
	9.3.1 Autonomy	212
	9.3.2 SPSS LexiQuest	212
	9.3.3 ClearForest	213
9.4	Conclusions and Future Work	214
9.5	Industrial Counterpoint: Is ETD Useful? Dr. Daniel J. Phelps, Leader, Information Mining Group, Eastman Kodak	215
	References	219

Survey of Text Mining

Clustering, Classification, and Retrieval

Berry, M.W. (Ed.)

2004, XVII, 244 p. 46 illus., Hardcover

ISBN: 978-0-387-95563-6