Contents

List of Contributors xi
Preface xiii

PART I TEXT EXTRACTION, CLASSIFICATION, AND CLUSTERING 1

1 Automatic keyword extraction from individual documents 3
  1.1 Introduction 3
     1.1.1 Keyword extraction methods 4
  1.2 Rapid automatic keyword extraction 5
     1.2.1 Candidate keywords 6
     1.2.2 Keyword scores 7
     1.2.3 Adjoining keywords 8
     1.2.4 Extracted keywords 8
  1.3 Benchmark evaluation 9
     1.3.1 Evaluating precision and recall 9
     1.3.2 Evaluating efficiency 10
  1.4 Stoplist generation 11
  1.5 Evaluation on news articles 15
     1.5.1 The MPQA Corpus 15
     1.5.2 Extracting keywords from news articles 15
  1.6 Summary 18
  1.7 Acknowledgements 19
  References 19

2 Algebraic techniques for multilingual document clustering 21
  2.1 Introduction 21
  2.2 Background 22
  2.3 Experimental setup 23
  2.4 Multilingual LSA 25
  2.5 Tucker1 method 27
5 Constrained clustering with \( k \)-means type algorithms

5.1 Introduction

5.2 Notations and classical \( k \)-means

5.3 Constrained \( k \)-means with Bregman divergences

5.3.1 Quadratic \( k \)-means with cannot-link constraints

5.3.2 Elimination of must-link constraints

5.3.3 Clustering with Bregman divergences

5.4 Constrained \( \text{smoka} \) type clustering

5.5 Constrained spherical \( k \)-means

5.5.1 Spherical \( k \)-means with cannot-link constraints only

5.5.2 Spherical \( k \)-means with cannot-link and must-link constraints

5.6 Numerical experiments

5.6.1 Quadratic \( k \)-means

5.6.2 Spherical \( k \)-means

5.7 Conclusion

References

PART II ANOMALY AND TREND DETECTION

6 Survey of text visualization techniques

6.1 Visualization in text analysis

6.2 Tag clouds

6.3 Authorship and change tracking

6.4 Data exploration and the search for novel patterns

6.5 Sentiment tracking

6.6 Visual analytics and FutureLens

6.7 Scenario discovery

6.7.1 Scenarios

6.7.2 Evaluating solutions

6.8 Earlier prototype

6.9 Features of FutureLens

6.10 Scenario discovery example: bioterrorism

6.11 Scenario discovery example: drug trafficking

6.12 Future work

References

7 Adaptive threshold setting for novelty mining

7.1 Introduction

7.2 Adaptive threshold setting in novelty mining
8 Text mining and cybercrime
  8.1 Introduction 149
  8.2 Current research in Internet predation and cyberbullying
    8.2.1 Capturing IM and IRC chat 151
    8.2.2 Current collections for use in analysis 152
    8.2.3 Analysis of IM and IRC chat 153
    8.2.4 Internet predation detection 153
    8.2.5 Cyberbullying detection 158
    8.2.6 Legal issues 159
  8.3 Commercial software for monitoring chat 159
  8.4 Conclusions and future directions 161
  8.5 Acknowledgements 162
  References 162

PART III TEXT STREAMS 165

9 Events and trends in text streams
  9.1 Introduction 167
  9.2 Text streams 169
  9.3 Feature extraction and data reduction 170
  9.4 Event detection 171
  9.5 Trend detection 174
  9.6 Event and trend descriptions 176
  9.7 Discussion 180
  9.8 Summary 181
  9.9 Acknowledgements 181
  References 181

10 Embedding semantics in LDA topic models
  10.1 Introduction 183
  10.2 Background 184