

Text mining is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval, and knowledge management. Similarly, link detection – a rapidly evolving approach to the analysis of text that shares and builds on many of the key elements of text mining – also provides new tools for people to better leverage their burgeoning textual data resources. Link detection relies on a process of building up networks of interconnected objects through various relationships in order to discover patterns and trends. The main tasks of link detection are to extract, discover, and link together sparse evidence from vast amounts of data sources, to represent and evaluate the significance of the related evidence, and to learn patterns to guide the extraction, discovery, and linkage of entities.

The Text Mining Handbook presents a comprehensive discussion of the state of the art in text mining and link detection. In addition to providing an in-depth examination of core text mining and link detection algorithms and operations, the work examines advanced preprocessing techniques, knowledge representation considerations, and visualization approaches. Finally, the book explores current real-world, mission-critical applications of text mining and link detection in such varied fields as corporate finance business intelligence, genomics research, and counterterrorism activities.

Ronen Feldman is an Associate Professor and the head of the Information Systems department at the Business School of the Hebrew University in Jerusalem. Professor Feldman received his B.Sc. in Math, Physics and Computer Science (specializing in Machine Learning) from the Hebrew University and his Ph.D. in Computer Science from Cornell University in NY. He was an Adjunct Professor at NYU Stern Business School. Professor Feldman is the Chief Scientist of Digital Trowel, an Israeli company specializing in development of text mining tools and applications. A pioneer in the areas of machine learning, data mining, and unstructured data management, he has authored or co-authored more than 100 published articles and conference papers in these areas.

James Sanger is a venture capitalist, applied technologist, and recognized industry expert in the areas of commercial data solutions, Internet applications, and IT security products. He is a partner at ABS Ventures, an independent venture firm founded in 1982 and originally associated with technology banking leader Alex. Brown and Sons. Immediately before joining ABS Ventures, Mr. Sanger was a Managing Director in the New York offices of DB Capital Venture Partners, the global venture capital arm of Deutsche Bank. Mr. Sanger has been a board member of several thought-leading technology companies, including Inxight Software, Gomez Inc., and ClearForest, Inc; he has also served as an official observer to the boards of AlphaBlox (acquired by IBM in 2004), Intralinks, and Imagine Software and as a member of the Technical Advisory Board of Qualys, Inc.

"This book is definitely worth having in your book shelf as a handy reference."

– IAPR Newsletter

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-0-521-83657-9



9 780521 836579 >

I. Introduction to Text Mining	1
I.1 Defining Text Mining	1
I.2 General Architecture of Text Mining Systems	13
II. Core Text Mining Operations	19
II.1 Core Text Mining Operations	19
II.2 Using Background Knowledge for Text Mining	41
II.3 Text Mining Query Languages	51
III. Text Mining Preprocessing Techniques	57
III.1 Task-Oriented Approaches	58
III.2 Further Reading	62
IV. Categorization	64
IV.1 Applications of Text Categorization	65
IV.2 Definition of the Problem	66
IV.3 Document Representation	68
IV.4 Knowledge Engineering Approach to TC	70
IV.5 Machine Learning Approach to TC	70
IV.6 Using Unlabeled Data to Improve Classification	78
IV.7 Evaluation of Text Classifiers	79
IV.8 Citations and Notes	80
V. Clustering	82
V.1 Clustering Tasks in Text Analysis	82
V.2 The General Clustering Problem	84
V.3 Clustering Algorithms	85
V.4 Clustering of Textual Data	88
V.5 Citations and Notes	92

VI. Information Extraction	94
VI.1 Introduction to Information Extraction	94
VI.2 Historical Evolution of IE: The Message Understanding Conferences and Tipster	96
VI.3 IE Examples	101
VI.4 Architecture of IE Systems	104
VI.5 Anaphora Resolution	109
VI.6 Inductive Algorithms for IE	119
VI.7 Structural IE	122
VI.8 Further Reading	129
VII. Probabilistic Models for Information Extraction	131
VII.1 Hidden Markov Models	131
VII.2 Stochastic Context-Free Grammars	137
VII.3 Maximal Entropy Modeling	138
VII.4 Maximal Entropy Markov Models	140
VII.5 Conditional Random Fields	142
VII.6 Further Reading	145
VIII. Preprocessing Applications Using Probabilistic and Hybrid Approaches	146
VIII.1 Applications of HMM to Textual Analysis	146
VIII.2 Using MEMM for Information Extraction	152
VIII.3 Applications of CRFs to Textual Analysis	153
VIII.4 TEG: Using SCFG Rules for Hybrid Statistical-Knowledge-Based IE	155
VIII.5 Bootstrapping	166
VIII.6 Further Reading	175
IX. Presentation-Layer Considerations for Browsing and Query Refinement	177
IX.1 Browsing	177
IX.2 Accessing Constraints and Simple Specification Filters at the Presentation Layer	185
IX.3 Accessing the Underlying Query Language	186
IX.4 Citations and Notes	187
X. Visualization Approaches	189
X.1 Introduction	189
X.2 Architectural Considerations	192
X.3 Common Visualization Approaches for Text Mining	194
X.4 Visualization Techniques in Link Analysis	225
X.5 Real-World Example: The Document Explorer System	235
XI. Link Analysis	242
XI.1 Preliminaries	242

XI.2 Automatic Layout of Networks	244
XI.3 Paths and Cycles in Graphs	248
XI.4 Centrality	249
XI.5 Partitioning of Networks	257
XI.6 Pattern Matching in Networks	270
XI.7 Software Packages for Link Analysis	271
XI.8 Citations and Notes	272

XII. Text Mining Applications

273

XII.1 General Considerations	274
XII.2 Corporate Finance: Mining Industry Literature for Business Intelligence	279
XII.3 A “Horizontal” Text Mining Application: Patent Analysis Solution Leveraging a Commercial Text Analytics Platform	295
XII.4 Life Sciences Research: Mining Biological Pathway Information with GeneWays	307

Appendix A: DIAL: A Dedicated Information Extraction Language for Text Mining

315

A.1 What Is the DIAL Language?	315
A.2 Information Extraction in the DIAL Environment	316
A.3 Text Tokenization	318
A.4 Concept and Rule Structure	318
A.5 Pattern Matching	320
A.6 Pattern Elements	321
A.7 Rule Constraints	325
A.8 Concept Guards	326
A.9 Complete DIAL Examples	327

Bibliography

335

Index

389