

Preface

As we enter the second decade of the World Wide Web (WWW), the textual revolution has seen a tremendous change in the availability of online information. Finding information for just about any need has never been more automatic – just a keystroke or mouseclick away. While the digitalization and creation of textual materials continues at light speed, the ability to navigate, mine, or casually browse through documents too numerous to read (or print) lags far behind.

What approaches to text mining are available to efficiently organize, classify, label, and extract relevant information for today’s information-centric users? What algorithms and software should be used to detect emerging trends from both text streams and archives? These are just a few of the important questions addressed at the Text Mining Workshop held on April 13, 2002 in Arlington, VA. This workshop, the second in a series of annual workshops on text mining, was held on the third day of the Second SIAM International Conference on Data Mining (April 11–13, 2002).

With close to 60 applied mathematicians and computer scientists representing universities, industrial corporations, and government laboratories, the workshop featured both invited and contributed talks on important topics such as efficient methods for document clustering, synonym extraction, efficient vector space models and metalearning approaches for text retrieval, hot topic discovery from dirty text, and trend detection from both queries and documents. The workshop was sponsored by the Army High Performance Computing Research Center (AH-PCRC) – Laboratory for Advanced Computing, SPSS, Insightful Corporation, and Salford Systems.

Several of the invited and contributed papers presented at the 2002 Text Mining Workshop have been compiled and expanded for this volume. Collectively, they span several major topic areas in text mining:

- I. Clustering and Classification,
- II. Information Extraction and Retrieval, and
- III. Trend Detection.

In Part I (Clustering and Classification), Howland and Park present cluster-preserving dimension reduction methods for efficient text classification; Senellart and Blondel demonstrate thesaurus construction using similarity measures between

vertices in graphs; Frigui and Nasraoui discuss clustering and keyword weighting; and Dhillon, Kogan, and Nicholas illustrate how both feature selection and document clustering can be accomplished with reduced dimension vector space models.

In Part II (Information Extraction and Retrieval), Kobayashi and Aono demonstrate the importance of detecting and interpreting minor document clusters using a vector space model based on Principal Component Analysis (PCA) rather than the popular Latent Semantic Indexing (LSI) method; Castellanos demonstrates how important topics can be extracted from dirty text associated with search logs in the customer support domain; and Cornelson et al. describe an innovative approach to information retrieval based on metalearning in which several algorithms are applied to the same corpus.

In Part III (Trend Detection), Wang, Bownas, and Berry mine Web queries from a university website in order to expose the type and nature of query characteristics through time; and Kontostathis et al. formally evaluate available Emerging Trend Detection (ETD) systems and discuss future criteria for the development of effective industrial-strength ETD systems.

Each chapter of this volume is preceded by a brief chapter overview and concluded by a list of references cited in that chapter. A main bibliography of all references cited and a subject-level index are also provided at the end of the volume. This volume details state-of-the-art algorithms and software for text mining from both the academic and industrial perspectives. Familiarity or coursework (undergraduate-level) in vector calculus and linear algebra is needed for several of the chapters in Parts I and II. While many open research questions still remain, this collection serves as an important benchmark in the development of both current and future approaches to mining textual information.

Acknowledgments: The editor would like to thank Justin Giles, Kevin Heinrich, and Svetlana Mironova who were extremely helpful in proofreading many of the chapters of this volume. Justin Giles also did a phenomenal job in managing all the correspondences with the contributing authors.

Michael W. Berry
Knoxville, TN
December 2002

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