Preface

As we enter the third 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 28, 2007, in Minneapolis, MN. This workshop, the fifth in a series of annual workshops on text mining, was held on the final day of the Seventh SIAM International Conference on Data Mining (April 26–28, 2007).

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 the application of techniques of machine learning in conjunction with natural language processing, information extraction and algebraic/mathematical approaches to computational information retrieval. The workshop's program also included an Anomaly Detection/Text Mining competition. NASA Ames Research Center of Moffett Field, CA, and SAS Institute Inc. of Cary, NC, sponsored the workshop.

Most of the invited and contributed papers presented at the 2007 Text Mining Workshop have been compiled and expanded for this volume. Several others are revised papers from the first edition of the book. Collectively, they span several major topic areas in text mining:

I. Clustering,
II. Document retrieval and representation,
III. Email surveillance and filtering, and
IV. Anomaly detection.

In Part I (Clustering), Howland and Park update their work on cluster-preserving dimension reduction methods for efficient text classification. Likewise, Senellart and Blondel revisit thesaurus construction using similarity measures between vertices in graphs. Both of these chapter were part of the first edition of this book (based on a SIAM text mining workshop held in April 2002). The next three chapters are completely new contributions. Zeimpekis and Gallopoulos implement and evaluate several clustering schemes that combine partitioning and hierarchical algorithms. Kogan, Nicholas, and Wiacek look at the hybrid clustering of large, high-dimensional data. AlSumait and Domeniconi round out this topic area with an examination of local semantic kernels for the clustering of text documents.

In Part II (Document Retrieval and Representation), Kobayashi and Aono revise their first edition chapter on 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. This is followed by Xia, Xing, Qi, and Li's chapter on applications of semidefinite programming in XML document classification.

In Part III (Email Surveillance and Filtering), Bader, Berry, and Browne take advantage of the Enron email dataset to look at topic detection over time using PARAFAC and multilinear algebra. Gansterer, Janecek, and Neumayer examine the use of latent semantic indexing to combat email spam.

In Part IV (Anomaly Detection), researchers from the NASA Ames Research Center share approaches to anomaly detection. These techniques were actually entries in a competition. held as part of the workshop. The top three finishers in the competition were: Cyril Goutte of NRC Canada, Edward G. Allan, Michael R. Horvath, Christopher V. Kopek, Brian T. Lamb, and Thomas S. Whaples of Wake Forest University (Michael W. Berry of the University of Tennessee was their advisor), and an international group from the Middle East led by Mostafa Keikha. Each chapter provides an explanation of its approach to the contest.

This volume details the 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. 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 editors would like to thank Murray Browne of the University of Tennessee and Catherine Brett of Springer UK in coordinating the management of manuscripts among the authors, editors, and the publisher.

Michael W. Berry and Malu Castellanos
Knoxville, TN and Palo Alto, CA
August 2007

http://www.springer.com/978-1-84800-045-2

Survey of Text Mining II
Clustering, Classification, and Retrieval
Berry, M.W.; CASTELLANOS, M. (Eds.)
2008, XVI, 240 p., Hardcover
ISBN: 978-1-84800-045-2