Survey of Text Mining II

Michael W. Berry • Malu Castellanos
Editors

Survey of Text Mining II

Clustering, Classification, and Retrieval

Springer

Michael W. Berry, BS, MS, PhD
Department of Computer Science
University of Tennessee, USA

Malu Castellanos, PhD
Hewlett-Packard Laboratories
Palo Alto, California, USA

ISBN 978-1-84800-045-2 e-ISBN 978-1-84800-046-9
DOI: 10.1007/978-1-84800-046-9

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2007935209

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

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

Contents

Contributors

Edward G. Allan
Department of Computer Science
Wake Forest University
P.O. Box 7311
Winston-Salem, NC 27109
Email: allaeg3@wfu.edu

Loulwah Alsumait
Department of Computer Science
George Mason University
4400 University Drive MSN 4A4
Fairfax, VA 22030
Email: lalsumai@gmu.edu

Masaki Aono
Department of Information and Computer Sciences, C-511
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho
Toyohashi-shi, Aichi 441-8580
Japan
Email: aono@ics.tut.ac.jp

Brett W. Bader
Sandia National Laboratories
Applied Computational Methods Department
P.O. Box 5800
Albuquerque, NM 87185-1318
Email: bwbader@sandia.gov
Homepage: http://www.cs.sandia.gov/~bwbader

Michael W. Berry
Department of Electrical Engineering and Computer Science
University of Tennessee
203 Claxton Complex
Knoxville, TN 37996-3450
Email: berry@eecs.utk.edu
Homepage: http://www.cs.utk.edu/~berry

Vincent D. Blondel
Division of Applied Mathematics
Université de Louvain
4, Avenue Georges Lemaître
B-1348 Louvain-la-neuve
Belgium
Email: blondel@inma.ucl.ac.be
Homepage: http://www.inma.ucl.ac.be/~blondel

Murray Browne
Department of Electrical Engineering and Computer Science
University of Tennessee
203 Claxton Complex
Knoxville, TN 37996-3450
Email: mbrowne@eecs.utk.edu

Malú Castellanos
IETL Department
Hewlett-Packard Laboratories
1501 Page Mill Road MS-1148
Palo Alto, CA 94304
Email: malu.castellanos@hp.com

Pat Castle
Intelligent Systems Division
NASA Ames Research Center
Moffett Field, CA 94035
Email: pcastle@email.arc.nasa.gov

Santanu Das
Intelligent Systems Division
NASA Ames Research Center
Moffett Field, CA 94035
Email: sdas@email.arc.nasa.gov

Carlotta Domeniconi
Department of Computer Science
George Mason University
4400 University Drive MSN 4A4
Fairfax, VA 22030
Email: carlotta@ise.gmu.edu
Homepage: http://www.ise.gmu.edu/~carlotta

Efstratios Gallopoulos
Department of Computer Engineering and Informatics
University of Patras
26500 Patras
Greece
Email: stratis@hpclab.ceid.upatras.gr
Homepage: http://scgroup.hpclab.ceid.upatras.gr/faculty/stratis/stratise.html

Wilfried N. Gansterer
Research Lab for Computational Technologies and Applications
University of Vienna
Lenaugasse 2/8
A - 1080 Vienna
Austria
Email: wilfried.gansterer@univie.ac.at

Cyril Goutte
Interactive Language Technologies
NRC Institute for Information Technology
283 Boulevard Alexandre Taché
Gatineau, QC J8X 3X7
Canada
Email: cyril.goutte@nrc-cnrc.gc.ca
Homepage: http://iit-iti.nrc-cnrc.gc.ca/personnel/goutte_cyril_f.html

Michael R. Horvath
Department of Computer Science
Wake Forest University
P.O. Box 7311
Winston-Salem, NC 27109
Email: horvmr5@wfu.edu

Peg Howland
Department of Mathematics and Statistics
Utah State University
3900 Old Main Hill
Logan, UT 84322-3900
Email: peg.howland@usu.edu
Homepage: http://www.math.usu.edu/~howland

Andreas G. K. Janecek
Research Lab for Computational Technologies and Applications
University of Vienna
Lenaugasse 2/8
A - 1080 Vienna
Austria
Email: andreas.janecek@univie.ac.at

Mostafa Keikha
Department of Electrical and Computer Engineering
University of Tehran
P.O. Box 14395-515, Tehran
Iran
Email: m.keikha@ece.ut.ac.ir

Mei Kobayashi
IBM Research, Tokyo Research Laboratory
1623-14 Shimotsuruma, Yamato-shi
Kanagawa-ken 242-8502
Japan
Email: mei@jp.ibm.com
Homepage: http://www.trl.ibm.com/people/meik

Jacob Kogan
Department of Mathematics and Statistics
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
Email: kogan@math.umbc.edu
Homepage: http://www.math.umbc.edu/~kogan

Christopher V. Kopek
Department of Computer Science
Wake Forest University
P.O. Box 7311
Winston-Salem, NC 27109
Email: kopecv5@wfu.edu

Brian T. Lamb
Department of Computer Science
Wake Forest University
P.O. Box 7311
Winston-Salem, NC 27109
Email: lambbt5@wfu.edu

Qi Li
Department of Computer Science
Western Kentucky University
1906 College Heights Boulevard #11076
Bowling Green, KY 42101-1076
Email: qi.li@wku.edu
Homepage: http://www.wku.edu/~qi.li

Robert Neumayer
Institute of Software Technology and Interactive Systems
Vienna University of Technology
Favoritenstraße 9-11/188/2
A - 1040 Vienna
Austria
Email: neumayer@ifs.tuwien.ac.at

Charles Nicholas
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
Email: nicholas@cs.umbc.edu
Homepage: http://www.cs.umbc.edu/~nicholas

Farhad Oroumchian
College of Information Technology
University of Wollongong in Dubai
P.O. Box 20183, Dubai
U.A.E.
Email: farhadoroumchian@uowdubai.ac.ae

Matthew E. Otey
Intelligent Systems Division
NASA Ames Research Center
Moffett Field, CA 94035
Email: otey@email.arc.nasa.gov

Haesun Park
Division of Computational Science and Engineering
College of Computing
Georgia Institute of Technology
266 Ferst Drive
Atlanta, GA 30332-0280
Email: hpark@cc.gatech.edu
Homepage: http://www.cc.gatech.edu/~hpark

Houduo Qi
Department of Mathematics
University of Southampton, Highfield
Southampton SO17 1BJ, UK
Email: hdqi@soton.ac.uk
Homepage: http://www.personal.soton.ac.uk/hdqi

Narjes Sharif Razavian
Department of Electrical and Computer Engineering
University of Tehran
P.O. Box 14395-515, Tehran
Iran
Email: n.razavian@ece.ut.ac.ir

Hassan Seyed Razi
Department of Electrical and Computer Engineering
University of Tehran
P.O. Box 14395-515, Tehran
Iran
Email: seyedraz@ece.ut.ac.ir

Pierre Senellart
INRIA Futurs & Université Paris-Sud
4 rue Jacques Monod
91893 Orsay Cedex
France
Email: pierre@senellart.com
Homepage: http://pierre.senellart.com/

Ashok N. Srivastava
Intelligent Systems Division
NASA Ames Research Center
Moffett Field, CA 94035
Email: ashok@email.arc.nasa.gov

Thomas S. Whaples
Department of Computer Science
Wake Forest University
P.O. Box 7311
Winston-Salem, NC 27109
Email: whapts3@wfu.edu

Mike Wiacek
Google, Inc.
1600 Amphitheatre Parkway
Mountain View, CA 94043
Email: mjwiacek@google.com

Zhonghang Xia
Department of Computer Science
Western Kentucky University
1906 College Heights Boulevard #11076
Bowling Green, KY 42101-1076
Email: zhonghang.xia@wku.edu
Homepage: http://www.wku.edu/~zhonghang.xia

Guangming Xing
Department of Computer Science
Western Kentucky University
1906 College Heights Boulevard #11076
Bowling Green, KY 42101-1076
Email: guangming.xing@wku.edu
Homepage: http://www.wku.edu/~guangming.xing

Dimitrios Zeimpekis
Department of Computer Engineering and Informatics
University of Patras
26500 Patras
Greece
Email: dsz@hpclab.ceid.upatras.gr