The Text Mining Handbook

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.

Dr. Ronen Feldman is a Senior Lecturer in the Mathematics and Computer Science Department of Bar-Ilan University and Director of the Data and Text Mining Laboratory. Dr. Feldman is cofounder, Chief Scientist, and President of ClearForest, Ltd., a leader in developing next-generation text mining applications for corporate and government clients. He also recently served as an Adjunct Professor at New York University's Stern School of Business. A pioneer in the areas of machine learning, data mining, and unstructured data management, he has authored or coauthored more than 70 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.

THE TEXT MINING HANDBOOK

Advanced Approaches in Analyzing Unstructured Data

Ronen Feldman
Bar-Ilan University, Israel

James Sanger
ABS Ventures, Waltham, Massachusetts

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA

www.cambridge.org
Information on this title: www.cambridge.org/9780521836579

First published 2007

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data

Feldman, Ronen, 1962–
The text mining handbook : advanced approaches in analyzing unstructured data /
Ronen Feldman, James Sanger.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-83657-3 (hardback)
1. Data mining – Handbooks, manuals, etc. I. Sanger, James, 1965– II. Title.
QA76.9.D343F45 2006
005.74 – dc22 2005029330

ISBN-13 978-0-521-83657-9 hardback
ISBN-10 0-521-83657-3 hardback

Cambridge University Press has no responsibility for
the persistence or accuracy of URLs for external or
third-party Internet Web sites referred to in this publication
and does not guarantee that any content on such
Web sites is, or will remain, accurate or appropriate.

In loving memory of my father, Issac Feldman

Contents

Preface

The information age has made it easy to store large amounts of data. The proliferation of documents available on the Web, on corporate intranets, on news wires, and elsewhere is overwhelming. However, although the amount of data available to us is constantly increasing, our ability to absorb and process this information remains constant. Search engines only exacerbate the problem by making more and more documents available in a matter of a few key strokes.

Text mining is a new and exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, natural language processing (NLP), information retrieval (IR), and knowledge management. Text mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results.

This book presents a general theory of text mining along with the main techniques behind it. We offer a generalized architecture for text mining and outline the algorithms and data structures typically used by text mining systems.

The book is aimed at the advanced undergraduate students, graduate students, academic researchers, and professional practitioners interested in complete coverage of the text mining field. We have included all the topics critical to people who plan to develop text mining systems or to use them. In particular, we have covered preprocessing techniques such as text categorization, text clustering, and information extraction and analysis techniques such as association rules and link analysis.

The book tries to blend together theory and practice; we have attempted to provide many real-life scenarios that show how the different techniques are used in practice. When writing the book we tried to make it as self-contained as possible and have compiled a comprehensive bibliography for each topic so that the reader can expand his or her knowledge accordingly.

BOOK OVERVIEW

The book starts with a gentle introduction to text mining that presents the basic definitions and prepares the reader for the next chapters. In the second chapter we describe the core text mining operations in detail while providing examples for each operation. The third chapter serves as an introduction to text mining preprocessing techniques. We provide a taxonomy of the operations and set the ground for Chapters IV through VII. Chapter IV offers a comprehensive description of the text categorization problem and outlines the major algorithms for performing text categorization.

Chapter V introduces another important text preprocessing task called text clustering, and we again provide a concrete definition of the problem and outline the major algorithms for performing text clustering. Chapter VI addresses what is probably the most important text preprocessing technique for text mining – namely, information extraction. We describe the general problem of information extraction and supply the relevant definitions. Several examples of the output of information extraction in several domains are also presented.

In Chapter VII, we discuss several state-of-the-art probabilistic models for information extraction, and Chapter VIII describes several preprocessing applications that either use the probabilistic models of Chapter VII or are based on hybrid approaches incorporating several models. The presentation layer of a typical text mining system is considered in Chapter IX. We focus mainly on aspects related to browsing large document collections and on issues related to query refinement. Chapter X surveys the common visualization techniques used either to visualize the document collection or the results obtained from the text mining operations. Chapter XI introduces the fascinating area of link analysis. We present link analysis as an analytical step based on the foundation of the text preprocessing techniques discussed in the previous chapters, most specifically information extraction. The chapter begins with basic definitions from graph theory and moves to common techniques for analyzing large networks of entities.

Finally, in Chapter XII, three real-world applications of text mining are considered. We begin by describing an application for articles posted in *BioWorld* magazine. This application identifies major biological entities such as genes and proteins and enables visualization of relationships between those entities. We then proceed to the GeneWays application, which is based on analysis of *PubMed* articles. The next application is based on analysis of U.S. patents and enables monitoring trends and visualizing relationships between inventors, assignees, and technology terms.

The appendix explains the DIAL language, which is a dedicated information extraction language. We outline the structure of the language and describe its exact syntax. We also offer several code examples that show how DIAL can be used to extract a variety of entities and relationships. A detailed bibliography concludes the book.

ACKNOWLEDGMENTS

This book would not have been possible without the help of many individuals. In addition to acknowledgments made throughout the book, we feel it important to

take the time to offer special thanks to an important few. Among these we would like to mention especially Benjamin Rosenfeld, who devoted many hours to revising the categorization and clustering chapters. The people at ClearForest Corporation also provided help in obtaining screen shots of applications using ClearForest technologies – most notably in Chapter XII. In particular, we would like to mention the assistance we received from Rafi Vesserman, Yonatan Aumann, Jonathan Schler, Yair Liberzon, Felix Harmatz, and Yizhar Regev. Their support meant a great deal to us in the completion of this project.

Adding to this list, we would also like to thank Ian Bonner and Kathy Bentaieb of Inxight Software for the screen shots used in Chapter X. Also, we would like to extend our appreciation to Andrey Rzhetsky for his personal screen shots of the GeneWays application.

A book written on a subject such as text mining is inevitably a culmination of many years of work. As such, our gratitude is extended to both Haym Hirsh and Oren Etzioni, early collaborators in the field.

In addition, we would like to thank Lauren Cowles of Cambridge University Press for reading our drafts and patiently making numerous comments on how to improve the structure of the book and its readability. Appreciation is also owed to Jessica Farris for help in keeping two very busy coauthors on track.

Finally it brings us great pleasure to thank those dearest to us – our children Yael, Hadar, Yair, Neta and Frithjof – for leaving us undisturbed in our rooms while we were writing. We hope that, now that the book is finished, we will have more time to devote to you and to enjoy your growth. We are also greatly indebted to our dear wives Hedva and Lauren for bearing with our long hours on the computer, doing research, and writing the endless drafts. Without your help, confidence, and support we would never have completed this book. Thank you for everything. We love you!