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

https://doi.org/10.1017/CBO9780511546914.001 Published online by Cambridge University Press

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!

https://doi.org/10.1017/CBO9780511546914.001 Published online by Cambridge University Press