Book Reviews

The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data

Ronen Feldman and James Sanger
(Bar-Ilan University and ABS Ventures)

Cambridge, England: Cambridge University Press, 2007, xii+410 pp; hardbound,
ISBN 0-521-83657-3, $70.00

Reviewed by
Rada Mihalcea
University of North Texas

Text mining is the process of discovering information in large text collections, and automatically identifying interesting patterns and relationships in textual data. It is a relatively new research area, which has recently raised much interest among the research and industry communities, mainly due to the continuously increasing amount of information available on the Web and elsewhere. Text mining is a highly interdisciplinary research area, bringing together research insights from the fields of data mining, natural language processing, machine learning, and information retrieval. In particular, text mining is closely related to the older area of data mining, which targets the extraction of interesting information from data records, although text mining is allegedly more difficult, as the source data consists of unstructured collections of documents rather than structured databases.

The book by Feldman and Sanger is a thorough introduction to text mining, covering the general architecture of text mining systems, along with the main techniques used by such systems. It addresses both the theory and practice of text mining, and it illustrates the different techniques with real-world scenarios and practical applications. It is particularly relevant for students and professional practitioners, being structured as a self-contained handbook that does not require previous experience in any of the research fields involved.

The book is structured into twelve chapters, which gradually introduce the area of text mining and related topics, starting with an introduction to the task of text mining, and ending with examples of practical applications from three different domains.

The first chapter can be regarded as an overview of the book. It starts by defining the problem of text mining and the key elements in text mining: the document collections, the document features (words, terms, and concepts), and the role of background knowledge in text mining. It then briefly touches upon the possible applications of text mining, such as pattern discovery and trend analysis, and shortly discusses the interface layer of text mining systems. The second part of the chapter lays down the general architecture of a text mining system, which also serves as a rough guide for the rest of the book, as it describes the main components of a text mining system that are described in detail in subsequent chapters.

Chapter 2 is one of the longest chapters in the book, and also one of the most dense in terms of newly introduced concepts. Despite being a more difficult read compared to the other chapters in the book, I found it to be the most informative with respect to operations specific to text mining. The chapter starts by defining the core text mining operations, including distributions, sets, and associations, and introduces

the main techniques for isolating interesting patterns and analyzing trends over time. The second part of the chapter overviews the role of background knowledge in text mining. The authors describe several ontologies and lexicons, and show with evidence from a real-world example (the FACT system they developed in the late 1990s) how background knowledge can be effectively integrated into text mining systems. A shortcoming of this section is the interchangeable use of "domain ontology" and "background knowledge," which can be confusing for computational linguists, who typically make a distinction between these terms. Finally, the third part of the chapter briefly describes query languages, which are later addressed in detail in Chapter 9.

Chapter 3 is meant as a short introduction to text preprocessing techniques, including tokenization, tagging, and parsing. This chapter, as well as several of the following chapters addressing text classification and information extraction, were most likely included in the book because of the authors' intention to make the book self-contained and appealing even for those with no background in computational linguistics.

Chapters 4 and 5 describe techniques for text classification and clustering, which are relevant to the selection of documents addressed by a text mining system. Chapter 4 describes the representation of documents for the purpose of text categorization, and introduces several machine-learning algorithms, including decision trees, naive Bayes, and SVMs, as well as committees of classifiers through bagging and boosting. Chapter 5 introduces several clustering algorithms, including agglomerative clustering, expectation maximization, and *K*-means, as well as techniques specific to clustering textual data such as latent semantic analysis.

The next three chapters, 6, 7, and 8, address the task of information extraction, which is a key element in any text mining system. Chapter 6 begins by defining the problem of information extraction, the architecture of a typical information extraction system, and the main knowledge-based and structural approaches to information extraction. Chapter 7 is dedicated to probabilistic models, including maximum entropy, hidden Markov models, and conditional random fields. Chapter 8 then describes text preprocessing using these probabilistic models, including part-of-speech tagging, shallow parsing, and named-entity tagging. The chapter also addresses bootstrapping techniques for information extraction. One drawback of this section of the book is the fact that the division of the material among these three chapters is not always very clear, and some parts could have been organized differently. For instance, the probabilistic models and their applications are split among Chapters 7 and 8, although they could have been combined under one chapter. Chapter 8 includes material on bootstrapping approaches for information extraction, which seems unrelated to the rest of this chapter and instead could have been included in Chapter 6.

Chapter 9 addresses techniques for interfacing with text mining systems, including browsing and displaying of distributions, associations, and hierarchies, as well as query languages and query refinement. Visualization techniques are then addressed in Chapter 10, which describes visual interfaces to text mining systems, such as association graphs, histograms, and self-organizing maps. The chapter is rich in illustrations, which contribute to a better understanding of the differences between the various visualization methods.

The relationships discovered by a text mining system can be further explored by using techniques for link analysis, which are described in Chapter 11. The chapter begins with a brief introduction to graph theory, followed by a description of several graph centrality methods and algorithms for network partitioning. The chapter also provides pointers to software packages for link analysis.

Finally, Chapter 12 illustrates the application of the text mining concepts introduced in the book to three practical text mining problems: industry literature mining, patent analysis, and protein interactions. By showing examples of concrete applications, the authors demonstrate the applicability of the text mining theory introduced in the book to practical real-world scenarios.

Overall, *The Text Mining Handbook* is a good introduction to text mining, written by leading experts in the field. The book is well written and addresses both the theory and practice of text mining, which makes it appealing for researchers and practitioners alike.

By being self-contained, with several chapters addressing in detail the main topics relevant to text mining, the book is highly recommended to those who would like to start delving into the area of text mining without having any previous background in computational linguistics. Although experts in computational linguistics will most likely find that they can safely skip over several of the text processing chapters (e.g., the introductory chapters on text preprocessing, or the chapters on text classification and information extraction), they will certainly find a lot of value in the chapters addressing the specific task of text mining (mainly Chapters 2, 10, 11, and 12).

Rada Mihalcea is an Assistant Professor of Computer Science at the University of North Texas. Her research interests are in lexical semantics, multilingual natural language processing, minimally supervised natural language learning, and graph-based algorithms for natural language processing. She served as the president of the ACL Special Interest Group on the Lexicon (SIGLEX) from 2004 to 2007, and she is currently a board member of the ACL Special Interest Group on Natural Language Learning (SIGNLL). Mihalcea's address is: Department of Computer Science, University of North Texas, P.O. Box 311366, Denton, TX, 76203; e-mail: rada@cs.unt.edu.