Text Data Management and Analysis

ACM Books

Editor in Chief

M. Tamer Özsu, *University of Waterloo*

ACM Books is a new series of high-quality books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers. ACM Books publications are widely distributed in both print and digital formats through booksellers and to libraries (and library consortia) and individual ACM members via the ACM Digital Library platform.

Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining

ChengXiang Zhai, *University of Illinois at Urbana–Champaign*
Sean Massung, *University of Illinois at Urbana–Champaign*
2016

An Architecture for Fast and General Data Processing on Large Clusters

Matei Zaharia, *Massachusetts Institute of Technology*
2016

Reactive Internet Programming: State Chart XML in Action

Franck Barbier, *University of Pau, France*
2016

Verified Functional Programming in Agda

Aaron Stump, *The University of Iowa*
2016

The VR Book: Human-Centered Design for Virtual Reality

Jason Jerald, *NextGen Interactions*
2016

Ada's Legacy: Cultures of Computing from the Victorian to the Digital Age

Robin Hammerman, *Stevens Institute of Technology*
Andrew L. Russell, *Stevens Institute of Technology*
2016

Edmund Berkeley and the Social Responsibility of Computer Professionals

Bernadette Longo, *New Jersey Institute of Technology*
2015

Candidate Multilinear Maps

Sanjam Garg, *University of California, Berkeley*
2015

Smarter than Their Machines: Oral Histories of Pioneers in Interactive Computing
John Cullinane, *Northeastern University; Mossavar-Rahmani Center for Business and Government, John F. Kennedy School of Government, Harvard University*
2015

A Framework for Scientific Discovery through Video Games
Seth Cooper, *University of Washington*
2014

Trust Extension as a Mechanism for Secure Code Execution on Commodity Computers
Bryan Jeffrey Parno, *Microsoft Research*
2014

Embracing Interference in Wireless Systems
Shyamnath Gollakota, *University of Washington*
2014

Text Data Management and Analysis

A Practical Introduction to Information Retrieval and Text Mining

ChengXiang Zhai

University of Illinois at Urbana–Champaign

Sean Massung

University of Illinois at Urbana–Champaign

ACM Books #12

Copyright © 2016 by the Association for Computing Machinery
and Morgan & Claypool Publishers

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews—without the prior permission of the publisher.

Designations used by companies to distinguish their products are often claimed as trademarks or registered trademarks. In all instances in which Morgan & Claypool is aware of a claim, the product names appear in initial capital or all capital letters. Readers, however, should contact the appropriate companies for more complete information regarding trademarks and registration.

Text Data Management and Analysis

ChengXiang Zhai and Sean Massung

books.acm.org
www.morganclaypoolpublishers.com

ISBN: 978-1-97000-119-8 hardcover
ISBN: 978-1-97000-116-7 paperback
ISBN: 978-1-97000-117-4 ebook
ISBN: 978-1-97000-118-1 ePub

Series ISSN: 2374-6769 print 2374-6777 electronic

DOIs:

10.1145/2915031 Book
10.1145/2915031.2915032 Preface
10.1145/2915031.2915033 Chapter 1
10.1145/2915031.2915034 Chapter 2
10.1145/2915031.2915035 Chapter 3
10.1145/2915031.2915036 Chapter 4
10.1145/2915031.2915037 Chapter 5
10.1145/2915031.2915038 Chapter 6
10.1145/2915031.2915039 Chapter 7
10.1145/2915031.2915040 Chapter 8
10.1145/2915031.2915041 Chapter 9
10.1145/2915031.2915042 Chapter 10
10.1145/2915031.2915043 Chapter 11
10.1145/2915031.2915044 Chapter 12
10.1145/2915031.2915045 Chapter 13
10.1145/2915031.2915046 Chapter 14
10.1145/2915031.2915047 Chapter 15
10.1145/2915031.2915048 Chapter 16
10.1145/2915031.2915049 Chapter 17
10.1145/2915031.2915050 Chapter 18
10.1145/2915031.2915051 Chapter 19
10.1145/2915031.2915052 Chapter 20
10.1145/2915031.2915053 Appendices
10.1145/2915031.2915054 References
10.1145/2915031.2915055 Index

A publication in the ACM Books series, #12
Editor in Chief: M. Tamer Özsu, *University of Waterloo*
Area Editor: Edward A. Fox, *Virginia Tech*

First Edition

10 9 8 7 6 5 4 3 2 1

To Mei and Alex

To Kai

Contents

Preface

The growth of "big data" created unprecedented opportunities to leverage computational and statistical approaches to turn raw data into actionable knowledge that can support various application tasks. This is especially true for the optimization of decision making in virtually all application domains such as health and medicine, security and safety, learning and education, scientific discovery, and business intelligence. Just as a microscope enables us to see things in the "micro world" and a telescope allows us to see things far away, one can imagine a "big data scope" would enable us to extend our perception ability to "see" useful hidden information and knowledge buried in the data, which can help make predictions and improve the optimality of a chosen decision. This book covers general computational techniques for managing and analyzing large amounts of text data that can help users manage and make use of text data in all kinds of applications.

Text data include all data in the form of natural language text (e.g., English text or Chinese text): all the web pages, social media data such as tweets, news, scientific literature, emails, government documents, and many other kinds of enterprise data. Text data play an essential role in our lives. Since we communicate using natural languages, we produce and consume a large amount of text data every day on all kinds of topics. The explosive growth of text data makes it impossible, or at least very difficult, for people to consume all the relevant text data in a timely manner. Thus, there is an urgent need for developing intelligent information retrieval systems to help people manage the text data and get access to the needed relevant information quickly and accurately at any time. This need is a major reason behind the recent growth of the web search engine industry. Due to the fact that text data are produced by humans for communication purposes, they are generally rich in semantic content and often contain valuable knowledge, information, opinions, and preferences of people. Thus, as a special kind of "big data," text data offer a great opportunity to discover various kinds of knowledge useful for many applications, especially knowledge about human opinions and preferences, which is often

directly expressed in text data. For example, it is now the norm for people to tap into opinionated text data such as product reviews, forum discussions, and social media text to obtain opinions. Once again, due to the overwhelming amount of information, people need intelligent software tools to help discover relevant knowledge for optimizing decisions or helping them complete their tasks more efficiently. While the technology for supporting text mining is not yet as mature as search engines for supporting text access, significant progress has been made in this area in recent years, and specialized text mining tools have now been widely used in many application domains. The subtitle of this book suggests that we cover two major topics, *information retrieval* and *text mining*. These two topics roughly correspond to the techniques needed to build the two types of application systems discussed above (i.e., search engines and text analytics systems), although the separation of the two is mostly artificial and only meant to help provide a high-level structure for the book, and a sophisticated application system likely would use many techniques from both topic areas.

In contrast to structured data, which conform to well-defined schemas and are thus relatively easy for computers to handle, text has less explicit structure so the development of intelligent software tools discussed above requires computer processing to understand the content encoded in text. The current technology of natural language processing has not yet reached a point to enable a computer to precisely understand natural language text (a main reason why humans often should be involved in the loop), but a wide range of statistical and heuristic approaches to management and analysis of text data have been developed over the past few decades. They are usually very robust and can be applied to analyze and manage text data in any natural language, and about any topic. This book intends to provide a systematic introduction to many of these approaches, with an emphasis on covering the most useful knowledge and skills required to build a variety of practically useful text information systems.

This book is primarily based on the materials that the authors have used for teaching a course on the topic of text data management and analysis (i.e., CS410 Text Information Systems) at the University of Illinois at Urbana–Champaign, as well as the two Massive Open Online Courses (MOOCs) on "Text Retrieval and Search Engines" and "Text Mining and Analytics" taught by the first author on Coursera in 2015. Most of the materials in the book directly match those of these two MOOCs with also similar structures of topics. As such, the book can be used as a main reference book for any of these two MOOCs.

Information Retrieval (IR) is a relatively mature field and there are no shortage of good textbooks on IR; for example, the most recent ones include *Modern Information Retrieval: The Concepts and Technology behind Search* by Baeza-Yates

and Ribeiro-Neto [2011], *Information Retrieval: Implementing and Evaluating Search Engines* by Büttcher et al. [2010], *Search Engines: Information Retrieval in Practice* by Croft et al. [2009], and *Introduction to Information Retrieval* by Manning et al. [2008]. Compared with these existing books on information retrieval, our book has a broader coverage of topics as it attempts to cover topics in both information retrieval and text mining, and attempts to paint a general roadmap for building a text information system that can support both text information access and text analysis. For example, it includes a detailed introduction to word association mining, probabilistic topic modeling, and joint analysis of text and non-text data, which are not available in any existing information retrieval books. In contrast with IR, Text Mining (TM) is far from mature and is actually still in its infancy. Indeed, how to define TM precisely remains an open question. As such, it appears that there is not yet a textbook on TM. As a textbook on TM, our book provides a basic introduction to the major representative techniques for TM. By introducing TM and IR in a unified framework, we want to emphasize the importance of integration of IR and TM in any practical text information system since IR plays two important roles in any TM application. The first is to enable fast reduction of the data size by filtering out a large amount of non-relevant text data to obtain a small set of most relevant data to a particular application problem. The second is to support an analyst to verify and interpret any patterns discovered from text data where an analyst would need to use search and browsing functions to reach and examine the most relevant support data to the pattern.

Another feature that sets this book apart is the availability of a companion toolkit for information retrieval and text mining, i.e., the MeTA toolkit (available at https://meta-toolkit.org/), which contains implementations of many techniques discussed in the book. Many exercises in the book are also designed based on this toolkit to help readers acquire practical skills of experimenting with the learned techniques from the book and applying them to solve real-world application problems.

This book consists of four parts. Part I provides an overview of the content covered in the book and some background knowledge needed to understand the chapters later. Parts II and III contain the major content of the book and cover a wide range of techniques in IR (called Text Data Access techniques) and techniques in TM (called Text Data Analysis techniques), respectively. Part IV summarizes the book with a unified framework for text management and analysis where many techniques of IR and TM can be combined to provide more advanced support for text data access and analysis with humans in the loop to control the workflow.

The required background knowledge to understand the content in this book is minimal since the book is intended to be mostly self-contained. However, readers

are expected to have basic knowledge about computer science, particularly data structures and programming languages and be comfortable with some basic concepts in probability and statistics such as conditional probability and parameter estimation. Readers who do not have this background may still be able to follow the basic ideas of most of the algorithms discussed in the book; they can also acquire the needed background by carefully studying Chapter 2 of the book and, if necessary, reading some of the references mentioned in the Bibliographical Notes section of that chapter to have a solid understanding of all the major concepts mentioned therein. MeTA can be used by anyone to easily experiment with algorithms and build applications, but modifying it or extending it would require at least some basic knowledge of C++ programming.

The book can be used as a textbook for an upper-level undergraduate course on information retrieval and text mining or a reference book for a graduate course to cover practical aspects of information retrieval and text mining. It should also be useful to practitioners in industry to help them acquire a wide range of practical techniques for managing and analyzing text data that they can use immediately to build various interesting real-world applications.

Acknowledgments

This book is the result of many people's help. First and foremost, we want to express our sincere thanks to Edward A. Fox for his invitation to write this book for the ACM Book Series in the area of Information Retrieval and Digital Libraries, of which he is the Area Editor. We are also grateful to Tamer Ozsu, Editor-in-Chief of ACM Books, for his support and useful comments on the book proposal. Without their encouragement and support this book would have not been possible. Next, we are deeply indebted to Edward A. Fox, Donna Harman, Bing Liu, and Jimmy Lin for thoroughly reviewing the initial draft of the book and providing very useful feedback and constructive suggestions. While we were not able to fully implement all their suggestions, all their reviews were extremely helpful and led to significant improvement of the quality of the book in many ways; naturally, any remaining errors in the book are solely the responsibility of the authors.

Throughout the process of writing the book, we received strong support and great help from Diane Cerra, Executive Editor at Morgan & Claypool Publishers, whose regular reminders and always timely support are key factors that prevented us from having the risk of taking "forever" to finish the book; for this, we are truly grateful to her. In addition, we would like to thank Sara Kreisman for copyediting and Paul C. Anagnostopoulos and his production team at Windfall Software (Ted

Laux, Laurel Muller, MaryEllen Oliver, and Jacqui Scarlott) for their great help with indexing, illustrations, art proofreading, and composition, which ensured a fast and smooth production of the book.

The content of the book and our understanding of the topics covered in the book have benefited from many discussions and interactions with a large number of people in both the research community and industry. Due to space limitations, we can only mention some of them here (and have to apologize to many whose names are not mentioned): James Allan, Charu Aggarwal, Ricardo Baeza-Yates, Nicholas J. Belkin, Andrei Broder, Jamie Callan, Jaime Carbonell, Kevin C. Chang, Yi Chang, Charlie Clarke, Fabio Crestani, W. Bruce Croft, Maarten de Rijke, Arjen de Vries, Daniel Diermeier, AnHai Doan, Susan Dumais, David A. Evans, Edward A. Fox, Ophir Frieder, Norbert Fuhr, Evgeniy Gabrilovich, C. Lee Giles, David Grossman, Jiawei Han, Donna Harman, Marti Hearst, Jimmy Huang, Rong Jin, Thorsten Joachims, Paul Kantor, David Karger, Diane Kelly, Ravi Kumar, Oren Kurland, John Lafferty, Victor Lavrenko, Lillian Lee, David Lewis, Jimmy Lin, Bing Liu, Wei-Ying Ma, Christopher Manning, Gary Marchionini, Andrew McCallum, Alistair Moffat, Jian-Yun Nie, Douglas Oard, Dragomir R. Radev, Prabhakar Raghavan, Stephen Robertson, Roni Rosenfeld, Dan Roth, Mark Sanderson, Bruce Schatz, Fabrizio Sebastiani, Amit Singhal, Keith van Rijsbergen, Luo Si, Noah Smith, Padhraic Smyth, Andrew Tomkins, Ellen Voorhees, and Yiming Yang, Yi Zhang, Justin Zobel. We want to thank all of them for their indirect contributions to this book. Some materials in the book, especially those in Chapter 19, are based on the research work done by many Ph.D. graduates of the Text Information Management and Analysis (TIMAN) group at the University of Illinois at Urbana–Champaign, under the supervision by the first author. We are grateful to all of them, including Tao Tao, Hui Fang, Xuehua Shen, Azadeh Shakery, Jing Jiang, Qiaozhu Mei, Xuanhui Wang, Bin Tan, Xu Ling, Younhee Ko, Alexander Kotov, Yue Lu, Maryam Karimzadehgan, Yuanhua Lv, Duo Zhang, V.G.Vinod Vydiswaran, Hyun Duk Kim, Kavita Ganesan, Parikshit Sondhi, Huizhong Duan, Yanen Li, Hongning Wang, Mingjie Qian, and Dae Hoon Park. The authors' own work included in the book has been supported by multiple funding sources, including NSF, NIH, NASA, IARPA, Air Force, ONR, DHS, Alfred P. Sloan Foundation, and many companies including Microsoft, Google, IBM, Yahoo!, LinkedIn, Intel, HP, and TCL. We are thankful to all of them.

The two Massive Open Online Courses (MOOCs) offered by the first author for the University of Illinois at Urbana–Champaign (UIUC) in 2015 on Coursera (i.e., *Text Retrieval and Search Engines* and *Text Mining and Analytics*) provided a direct basis for this book in the sense that many parts of the book are based primarily on the transcribed notes of the lectures in these two MOOCs. We thus would like

to thank all the people who have helped with these two MOOCs, especially TAs Hussein Hazimeh and Alex Morales, and UIUC instruction support staff Jason Mock, Shannon Bicknell, Katie Woodruff, and Edward Noel Dignan, and the Head of Computer Science Department, Rob Rutenbar, whose encouragement, support, and help are all essential for these two MOOCs to happen. The first author also wants to thank UIUC for allowing him to use the sabbatical leave in Fall 2015 to work on this book. Special thanks are due to Chase Geigle, co-founder of MeTA. In addition to all the above, the second author would like to thank Chase Geigle, Jason Cho, and Urvashi Khandelwal (among many others) for insightful discussion and encouragement.

Finally, we would like to thank all our family members, particularly our wives, Mei and Kai, for their love and support. The first author wants to further thank his brother Chengxing for the constant intellectual stimulation in their regular research discussions and his parents for cultivating his passion for learning and sharing knowledge with others.

ChengXiang Zhai
Sean Massung
June 2016

PART I

OVERVIEW AND BACKGROUND

1 Introduction

In the last two decades, we have experienced an explosive growth of online information. According to a study done at University of California Berkeley back in 2003: ". . . the world produces between 1 and 2 exabytes (1018 petabytes) of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. Printed documents of all kinds comprise only .03% of the total." [Lyman et al. 2003]

A large amount of online information is textual information (i.e., in natural language text). For example, according to the Berkeley study cited above: "Newspapers represent 25 terabytes annually, magazines represent 10 terabytes . . . office documents represent 195 terabytes. It is estimated that 610 billion emails are sent each year representing 11,000 terabytes." Of course, there are also blog articles, forum posts, tweets, scientific literature, government documents, etc. Roe [2012] updates the email count from 610 billion emails in 2003 to 107 *trillion* emails sent in 2010. According to a recent IDC report report [Gantz & Reinsel 2012], from 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes.

While, in general, all kinds of online information are useful, textual information plays an especially important role and is arguably the most useful kind of information for the following reasons.

Text (natural language) is the most natural way of encoding human knowledge. As a result, most human knowledge is encoded in the form of text data. For example, scientific knowledge almost exclusively exists in scientific literature, while technical manuals contain detailed explanations of how to operate devices.

Text is by far the most common type of information encountered by people. Indeed, most of the information a person produces and consumes daily is in text form.

Text is the most expressive form of information in the sense that it can be used to describe other media such as video or images. Indeed, image search engines such as those supported by Google and Bing often rely on matching companion text of images to retrieve "matching" images to a user's keyword query.

The explosive growth of online text information has created a strong demand for intelligent software tools to provide the following two related services to help people manage and exploit big text data.

Text Retrieval. The growth of text data makes it impossible for people to consume the data in a timely manner. Since text data encode much of our accumulated knowledge, they generally cannot be discarded, leading to, e.g., the accumulation of a large amount of literature data which is now beyond any individual's capacity to even skim over. The rapid growth of online text information also means that no one can possibly digest all the new information created on a daily basis. Thus, there is an urgent need for developing intelligent text retrieval systems to help people get access to the needed relevant information quickly and accurately, leading to the recent growth of the web search industry. Indeed, web search engines like Google and Bing are now an essential part of our daily life, serving millions of queries daily. In general, search engines are useful anywhere there is a relatively large amount of text data (e.g., desktop search, enterprise search or literature search in a specific domain such as PubMed).

Text Mining. Due to the fact that text data are produced by humans for communication purposes, they are generally rich in semantic content and often contain valuable knowledge, information, opinions, and preferences of people. As such, they offer great opportunity for discovering various kinds of knowledge useful for many applications, especially knowledge about human opinions and preferences, which is often directly expressed in text data. For example, it is now the norm for people to tap into opinionated text data such as product reviews, forum discussions, and social media text to obtain opinions about topics interesting to them and optimize various decision-making tasks such as purchasing a product or choosing a service. Once again, due to the overwhelming amount of information, people need intelligent software tools to help discover relevant knowledge to optimize decisions or help them complete their tasks more efficiently. While the technology for supporting text mining is not yet as mature as search engines for supporting text access, sig-

nificant progress has been made in this area in recent years, and specialized text mining tools have now been widely used in many application domains.

In contrast to structured data, which conform to well-defined schemas and are thus relatively easy for computers to handle, text has less explicit structure, so the development of intelligent software tools discussed above requires computer processing to understand the content encoded in text. The current technology of natural language processing has not yet reached a point to enable a computer to precisely understand natural language text (a main reason why humans often should be involved in the loop), but a wide range of statistical and heuristic approaches to management and analysis of text data have been developed over the past few decades. They are usually very robust and can be applied to analyze and manage text data in any natural language, and about any topic. This book intends to provide a systematic introduction to many of these approaches, with an emphasis on covering the most useful knowledge and skills required to build a variety of practically useful text information systems.

The two services discussed above (i.e., text retrieval and text mining) conceptually correspond to the two natural steps in the process of analyzing any "big text data" as shown in Figure 1.1. While the raw text data may be large, a specific application often requires only a small amount of most relevant text data, thus conceptually, the very first step in any application should be to identify the *relevant text data* to a particular application or decision-making problem and avoid the unnecessary processing of large amounts of non-relevant text data. This first step of converting the raw big text data into much smaller, but highly relevant text data is often accomplished by techniques of text retrieval with help from users (e.g., users may use multiple queries to collect all the relevant text data for a decision problem). In this first step, the main goal is to connect users (or applications) with the most relevant text data.

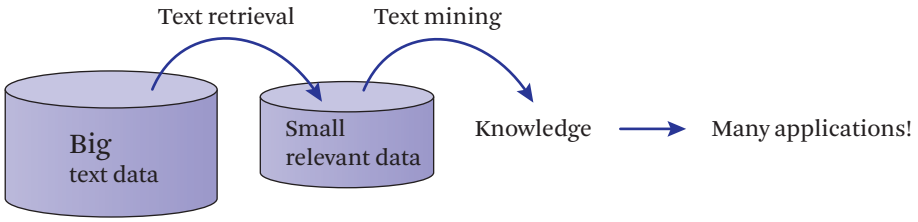

Figure 1.1 **Text retrieval and text mining are two main techniques for analyzing big text data.**

Once we obtain a small set of most relevant text data, we would need to further analyze the text data to help users digest the content and knowledge in the text data. This is the text mining step where the goal is to further discover knowledge and patterns from text data so as to support a user's task. Furthermore, due to the need for assessing trustworthiness of any discovered knowledge, users generally have a need to go back to the original raw text data to obtain appropriate context for interpreting the discovered knowledge and verify the trustworthiness of the knowledge, hence a search engine system, which is primarily useful for text access, also has to be available in any text-based decision-support system for supporting knowledge provenance. The two steps are thus conceptually interleaved, and a full-fledged intelligent text information system must integrate both in a unified framework.

It is worth pointing out that put in the context of "big data," text data is very different from other kinds of data because it is generally produced directly by humans and often also meant to be consumed by humans as well. In contrast, other data tend to be machine-generated data (e.g., data collected by using all kinds of physical sensors). Since humans can understand text data far better than computers can, involvement of humans in the process of mining and analyzing text data is absolutely crucial (much more necessary than in other big data applications), and how to optimally divide the work between humans and machines so as to optimize the collaboration between humans and machines and maximize their "combined intelligence" with minimum human effort is a general challenge in all applications of text data management and analysis. The two steps discussed above can be regarded as two different ways for a text information system to assist humans: information retrieval systems assist users in finding from a large collection of text data the most relevant text data that are actually needed for solving a specific application problem, thus effectively turning big raw text data into much smaller relevant text data that can be more easily processed by humans, while text mining application systems can assist users in analyzing patterns in text data to extract and discover useful actionable knowledge directly useful for task completion or decision making, thus providing more direct task support for users.

With this view, we partition the techniques covered in the book into two parts to match the two steps shown in Figure 1.1, which are then followed by one chapter to discuss how all the techniques may be integrated in a unified text information system. The book attempts to provide a complete coverage of all the major concepts, techniques, and ideas in information retrieval and text data mining from a practical viewpoint. It includes many hands-on exercises designed with a companion software toolkit MeTA to help readers learn how to apply techniques of information

retrieval and text mining to real-world text data and learn how to experiment with and improve some of the algorithms for interesting application tasks. This book can be used as a textbook for computer science undergraduates and graduates, library and information scientists, or as a reference book for practitioners working on relevant application problems in analyzing and managing text data.

1.1 Functions of Text Information Systems

From a user's perspective, a text information system (TIS) can offer three distinct, but related capabilities, as illustrated in Figure 1.2.

Information Access. This capability gives a user access to the useful information when the user needs it. With this capability, a TIS can connect the right information with the right user at the right time. For example, a search engine enables a user to access text information through querying, whereas a recommender system can push relevant information to a user as new information items become available. Since the main purpose of Information Access is to connect a user with relevant information, a TIS offering this capability

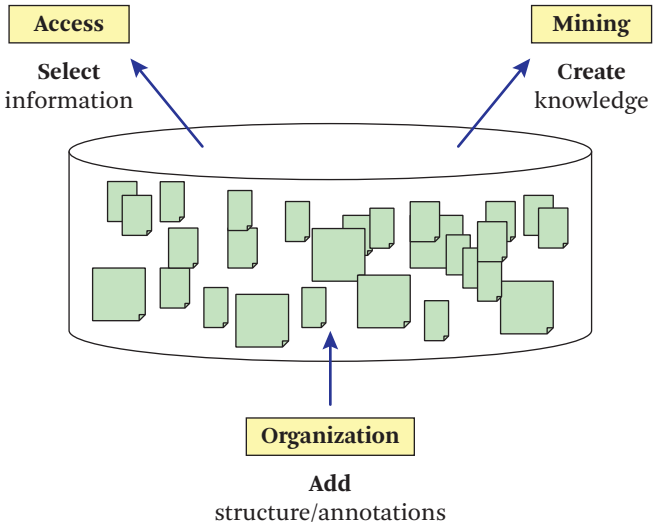

Figure 1.2 **Information access, knowledge acquisition, and text organization are three major capabilities of a text information system with text organization playing a supporting role for information access and knowledge acquisition. Knowledge acquisition is also often referred to as text mining.**

generally only does minimum analysis of text data sufficient for matching relevant information with a user's information need, and the original information items (e.g., web pages) are often delivered to the user in their original form, though summaries of the delivered items are often provided. From the perspective of text analysis, a user would generally need to read the information items to further digest and exploit the delivered information.

Knowledge Acquisition (Text Analysis). This capability enables a user to acquire useful knowledge encoded in the text data that is not easy for a user to obtain without synthesizing and analyzing a relatively large portion of the data. In this case, a TIS can analyze a large amount of text data to discover interesting patterns buried in text. A TIS with the capability of knowledge acquisition can be referred to as an analysis engine. For example, while a search engine can return relevant reviews of a product to a user, an analysis engine would enable a user to obtain directly the major positive or negative opinions about the product and to compare opinions about multiple similar products. A TIS offering the capability of knowledge acquisition generally would have to analyze text data in more detail and synthesize information from multiple text documents, discover interesting patterns, and create new information or knowledge.

Text Organization. This capability enables a TIS to annotate a collection of text documents with meaningful (topical) structures so that scattered information can be connected and a user can navigate in the information space by following the structures. While such structures may be regarded as "knowledge" acquired from the text data, and thus can be directly useful to users, in general, they are often only useful for facilitating either information access or knowledge acquisition, or both. In this sense, the capability of text organization plays a supporting role in a TIS to make information access and knowledge acquisition more effective. For example, the added structures can allow a user to search with constraints on structures or browse by following structures. The structures can also be leveraged to perform detailed analysis with consideration of constraints on structures.

Information access can be further classified into two modes: *pull* and *push*. In the pull mode, the user takes initiative to "pull" the useful information out from the system; in this case, the system plays a passive role and waits for a user to make a request, to which the system would then respond with relevant information. This mode of information access is often very useful when a user has an *ad hoc*

information need, i.e., a temporary information need (e.g., an immediate need for opinions about a product). For example, a search engine like Google generally serves a user in pull mode. In the push mode, the system takes initiative to "push" (recommend) to the user an information item that the system believes is useful to the user. The push mode often works well when the user has a relatively stable information need (e.g., hobby of a person); in such a case, a system can know "in advance" a user's preferences and interests, making it feasible to recommend information to a user without having the user to take the initiative. We cover both modes of information access in this book.

The pull mode further consists of two complementary ways for a user to obtain relevant information: *querying* and *browsing*. In the case of querying, the user specifies the information need with a (keyword) query, and the system would take the query as input and return documents that are estimated to be relevant to the query. In the case of browsing, the user simply navigates along structures that link information items together and progressively reaches relevant information. Since querying can also be regarded as a way to navigate, in one step, into a set of relevant documents, it's clear that browsing and querying can be interleaved naturally. Indeed, a user of a web search engine often interleaves querying and browsing.

Knowledge acquisition from text data is often achieved through the process of text mining, which can be defined as mining text data to discover useful knowledge. Both the data mining community and the natural language processing (NLP) community have developed methods for text mining, although the two communities tend to adopt slightly different perspective on the problem. From a data mining perspective, we may view text mining as mining a special kind of data, i.e., text. Following the general goals of data mining, the goal of text mining would naturally be regarded as to discover and extract interesting patterns in text data, which can include latent topics, topical trends, or outliers. From an NLP perspective, text mining can be regarded as to partially understand natural language text, convert text into some form of knowledge representation and make limited inferences based on the extracted knowledge. Thus a key task is to perform *information extraction*, which often aims to identify and extract mentions of various entities (e.g., people, organization, and location) and their relations (e.g., who met with whom). In practice, of course, any text mining applications would likely involve both pattern discovery (i.e., data mining view) and information extraction (i.e., NLP view), with information extraction serving as enriching the semantic representation of text, which enables pattern

finding algorithms to generate semantically more meaningful patterns than directly working on word or string-level representations of text. Due to our emphasis on covering general and robust techniques that can work for all kinds of text data without much manual effort, we mostly adopt the data mining view in this book since information extraction techniques tend to be more language-specific and generally require much manual effort. However, it is important to stress that information extraction is an essential component in any text information system that attempts to support deeper knowledge discovery or semantic analysis.

Applications of text mining can be classified as either direct applications, where the discovered knowledge would be directly consumed by users, or indirect applications, where the discovered knowledge isn't necessarily directly useful to a user, but can indirectly help a user through better support of information access. Knowledge acquisition can also be further classified based on what knowledge is to be discovered. However, due to the wide range of variations of the "knowledge," it is impossible to use a small number of categories to cover all the variations. Nevertheless, we can still identify a few common categories which we cover in this book. For example, one type of knowledge that a TIS can discover is a set of topics or subtopics buried in text data, which can serve as a concise summary of the major content in the text data. Another type of knowledge that can be acquired from opinionated text is the overall sentiment polarity of opinions about a topic.

1.2 Conceptual Framework for Text Information Systems

Conceptually, a text information system may consist of several modules, as illustrated in Figure 1.3.

First, there is a need for a module of *content analysis* based on natural language processing techniques. This module allows a TIS to transform raw text data into more meaningful representations that can be more effectively matched with a user's query in the case of a search engine, and more effectively processed in general in text analysis. Current NLP techniques mostly rely on *statistical machine learning* enhanced with limited linguistic knowledge with variable depth of understanding of text data; shallow techniques are robust, but deeper semantic analysis is only feasible for very limited domains. Some TIS capabilities (e.g., summarization) tend to require deeper NLP than others (e.g., search). Most text information systems use very shallow NLP, where text would simply be represented as a "*bag of words*," where words are basic units for representation and the order of words is ignored (although the counts of words are retained). However, a more sophisticated representation is

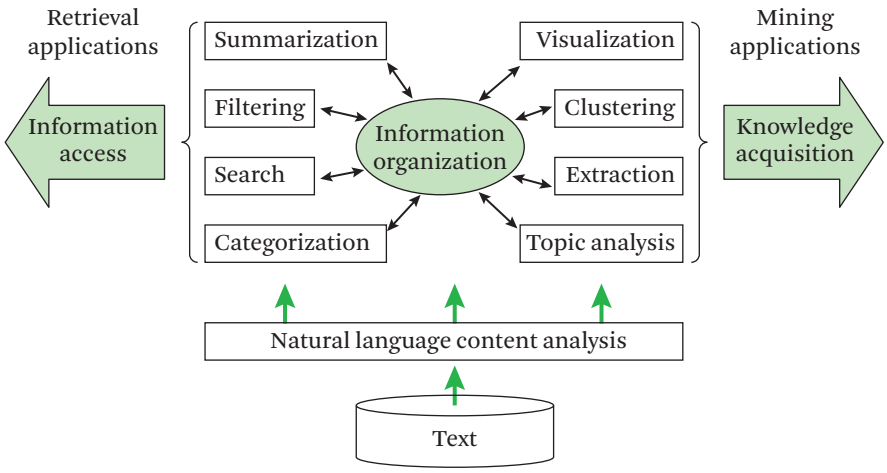

Figure 1.3 **Conceptual framework of text information systems.**

also possible, which may be based on recognized entities and relations or other techniques for more in-depth understanding of text.

With content analysis as the basis, there are multiple components in a TIS that are useful for users in different ways. The following are some commonly seen functions for managing and analyzing text information.

Search. Take a user's query and return relevant documents. The search component in a TIS is generally called a search engine. Web search engines are among the most useful search engines that enable users to effectively and efficiently deal with a huge amount of text data.

Filtering/Recommendation. Monitor an incoming stream, decide which items are relevant (or non-relevant) to a user's interest, and then recommend relevant items to the user (or filter out non-relevant items). Depending on whether the system focuses on recognizing relevant items or non-relevant items, this component in a TIS may be called a recommender system (whose goal is to recommend relevant items to users) or a filtering system (whose goal is to filter out non-relevant items to allow a user to keep only the relevant items). Literature recommender and spam email filter are examples of a recommender system and a filtering system, respectively.

Categorization. Classify a text object into one or several of the predefined categories where the categories can vary depending on applications. The categorization component in a TIS can annotate text objects with all kinds of meaningful categories, thus enriching the representation text data, which further enables more effective and deeper text analysis. The categories can also be used for organizing text data and facilitating text access. Subject categorizers that classify a text article into one or multiple subject categories and sentiment taggers that classify a sentence into positive, negative, or neutral in sentiment polarity are both specific examples of a text categorization system.

Summarization. Take one or multiple text documents, and generate a concise summary of the essential content. A summary reduces human effort in digesting text information and may also improve the efficiency in text mining. The summarization component of a TIS is called a summarizer. News summarizer and opinion summarizer are both examples of a summarizer.

Topic Analysis. Take a set of documents and extract and analyze topics in them. Topics directly facilitate digestion of text data by users and support browsing of text data. When combined with the companion non-textual data such as time, location, authors, and other meta data, topic analysis can generate many interesting patterns such as temporal trends of topics, spatiotemporal distributions of topics, and topic profiles of authors.

Information Extraction. Extract entities, relations of entities or other "knowledge nuggets" from text. The information extraction component of a TIS enables construction of entity-relation graphs. Such a knowledge graph is useful in multiple ways, including support of navigation (along edges and paths of the graph) and further application of graph mining algorithms to discover interesting entity-relation patterns.

Clustering. Discover groups of similar text objects (e.g., terms, sentences, documents, . . .). The clustering component of a TIS plays an important role in helping users explore an information space. It uses empirical data to create meaningful structures that can be useful for browsing text objects and obtaining a quick understanding of a large text data set. It is also useful for discovering outliers by identifying the items that do not form natural clusters with other items.

Visualization. Visually display patterns in text data. The visualization component is important for engaging humans in the process of discovering interesting patterns. Since humans are very good at recognizing visual patterns,

visualization of the results generated from various text mining algorithms is generally desirable.

This list also serves as an outline of the major topics to be covered later in this book. Specifically, search and filtering are covered first in Part II about text data access, whereas categorization, clustering, topic analysis, and summarization are covered later in Part III about text data analysis. Information extraction is not covered in this book since we want to focus on general approaches that can be readily applied to text data in *any* natural language, but information extraction often requires language-specific techniques. Visualization is also not covered due to the intended focus on algorithms in this book. However, it must be stressed that both information extraction and visualization are very important topics relevant to text data analysis and management. Readers interested in these techniques can find some useful references in the Bibliographic Notes at the end of this chapter.

1.3 Organization of the Book

The book is organized into four parts, as shown in Figure 1.4.

Part I. Overview and Background. This part consists of the first four chapters and provides an overview of the book and background knowledge, including basic concepts needed for understanding the content of the book that some readers may not be familiar with, and an introduction to the MeTA toolkit used for exercises in the book. This part also gives a brief overview of natural language processing techniques needed for understanding text data and obtaining informative representation of text needed in all text data analysis applications.

Part II. Text Data Access. This part consists of Chapters 5–11, covering the major techniques for supporting text data access. This part provides a systematic discussion of the basic information retrieval techniques, including the formulation of retrieval tasks as a problem of ranking documents for a query (Chapter 5), retrieval models that form the foundation of the design of ranking functions in a search engine (Chapter 6), feedback techniques (Chapter 7), implementation of retrieval systems (Chapter 8), and evaluation of retrieval systems (Chapter 9). It then covers web search engines, the most important application of information retrieval so far (Chapter 10), where techniques for analyzing links in text data for improving ranking of text objects are introduced and application of supervised machine learning to combine multiple

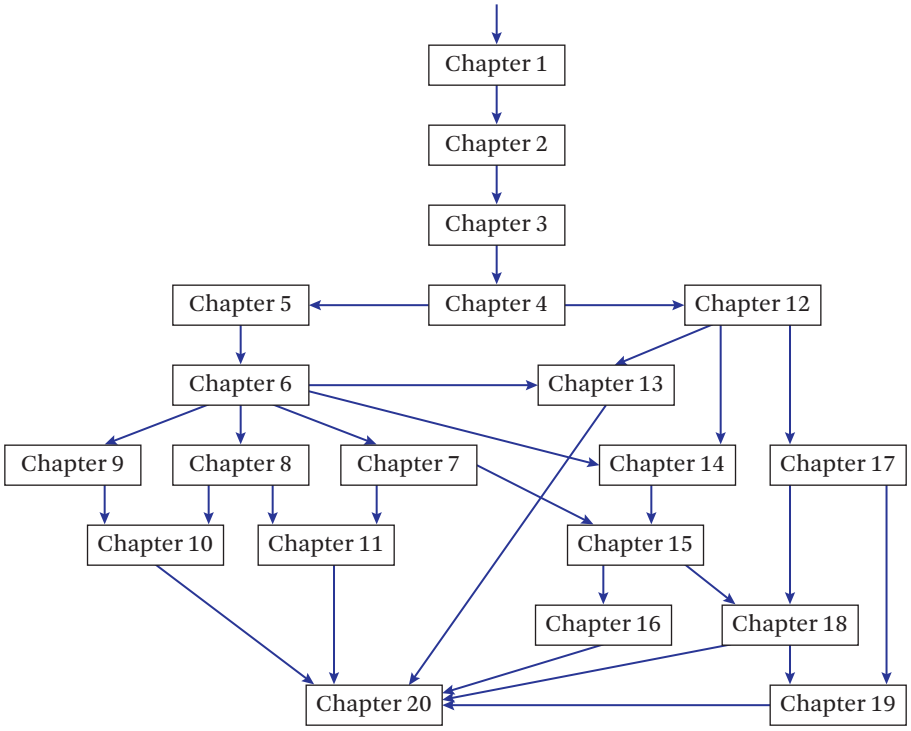

Figure 1.4 **Dependency relations among the chapters.**

features for ranking is briefly discussed. The last chapter in this part (Chapter 11) covers recommender systems which provide a "push" mode of information access, as opposed to the "pull" mode of information access supported by a typical search engine (i.e., querying by users).

Part III. Text Data Analysis. This part consists of Chapters 12–19, covering a variety of techniques for analyzing text data to facilitate user digestion of text data and discover useful topical or other semantic patterns in text data. Chapter 12 gives an overview of text analysis from the perspective of data mining, where we may view text data as data generated by humans as "subjective sensors" of the world; this view allows us to look at the text analysis problem in the more general context of data analysis and mining in general, and facilitates the discussion of joint analysis of text and non-text data. This is followed by multiple chapters covering a number of the most useful general techniques for analyzing text data without or with only minimum human effort. Specifically, Chapter 13 discusses techniques for discovering two fundamental se-

mantic relations between lexical units in text data, i.e., paradigmatic relations and syntagmatic relations, which can be regarded as an example of discovering knowledge about the natural language used to generate the text data (i.e., linguistic knowledge). Chapter 14 and Chapter 15 cover, respectively, two closely related techniques to generate and associate meaningful structures or annotations with otherwise unorganized text data, i.e., text clustering and text categorization. Chapter 16 discusses text summarization useful for facilitating human digestion of text information. Chapter 17 provides a detailed discussion of an important family of probabilistic approaches to discovery and analysis of topical patterns in text data (i.e., topic models). Chapter 18 discusses techniques for analyzing sentiment and opinions expressed in text data, which are key to discovery of knowledge about preferences, opinions, and behavior of people based on analyzing the text data produced by them. Finally, Chapter 19 discusses joint analysis of text and non-text data, which is often needed in many applications since it is in general beneficial to use as much data as possible for gaining knowledge and intelligence through (big) data analysis.

Part IV. Unified Text Management and Analysis System. This last part consists of Chapter 20 where we attempt to discuss how all the techniques discussed in this book can be conceptually integrated in an operator-based unified framework, and thus potentially implemented in a general unified system for text management and analysis that can be useful for supporting a wide range of different applications. This part also serves as a roadmap for further extension of MeTA to provide effective and general high-level support for various applications and provides guidance on how MeTA may be integrated with many other related existing toolkits, including particularly search engine systems, database systems, natural language processing toolkits, machine learning toolkits, and data mining toolkits.

Due to our attempt to treat all the topics from a practical perspective, most of the discussions of the concepts and techniques in the book are informal and intuitive. To satisfy the needs of some readers that might be interested in deeper understanding of some topics, the book also includes an appendix with notes to provide a more detailed and rigorous explanation of a few important topics.

1.4 How to Use this Book

Due to the extremely broad scope of the topics that we would like to cover, we have to make many tradeoffs between breadth and depth in coverage. When making

such a tradeoff, we have chosen to emphasize the coverage of the basic concepts and practical techniques of text data mining at the cost of not being able to cover many advanced techniques in detail, and provide some references at the end of many chapters to help readers learn more about those advanced techniques if they wish to. Our hope is that with the foundation received from reading this book, you will be able to learn about more advanced techniques by yourself or via another resource. We have also chosen to cover more general techniques for text management and analysis and favor techniques that can be applicable to any text in any natural language. Most techniques we discuss can be implemented without any human effort or only requiring minimal human effort; this is in contrast to some more detailed analysis of text data, particularly using natural language processing techniques. Such "deep analysis" techniques are obviously very important and are indeed necessary for some applications where we would like to go in-depth to understand text in detail. However, at this point, these techniques are often not scalable and they tend to require a large amount of human effort. In practice, it would be beneficial to combine both kinds of techniques.

We envision three main (and potentially overlapping) categories of readers.

Students. This book is specifically designed to give you hands-on experience in working with real text mining tools and applications. If used individually, we suggest first reading through Chapters 1–4 in order to get a good understanding of the prerequisite knowledge in this book. Chapters 1, 2, and 3 will familiarize you with the concepts and vocabulary necessary to understand the future chapters. Chapter 4 introduces you to the companion toolkit MeTA, which is used in exercises in each chapter. We hope the exercises and chapter descriptions provide inspiration to work on your own text mining project. The provided code in MeTA should give a large head start and allow you to focus more on your contribution.

If used in class, there are several logical flows that an instructor may choose to take. As prerequisite knowledge, we assume some basic knowledge in probability and statistics as well as programming in a language such as C++ or Java. MeTA is written in modern C++, although some exercises may be accomplished only by modifying config files.

Instructors. We have gathered a logical and cohesive collection of topics that may be combined together for various course curricula. For example, Part 1 and Part 2 of the book may be used as an undergraduate introduction to *Information Retrieval* with a focus on how search engines work. Exercises assume basic programming experience and a little mathematical background in probability and statistics. A different undergraduate course may choose to survey

the entire book as an *Introduction to Text Data Mining*, while skipping some chapters in Part 2 that are more specific to search engine implementation and applications specific to the Web. Another choice would be using all parts as a supplemental graduate textbook, where there is still some emphasis on practical programming knowledge that can be combined with reading referenced papers in each chapter. Exercises for graduate students could be implementing some methods they read in the references into MeTA.

The exercises at the end of each chapter give students experience working with a powerful—yet easily understandable—text retrieval and mining toolkit in addition to written questions. In a programming-focused class, using the MeTA exercises is strongly encouraged. Programming assignments can be created from selecting a subset of exercises in each chapter. Due to the modular nature of the toolkit, additional programming experiments may be created by extending the existing system or implementing other well-known algorithms that do not come with MeTA by default. Finally, students may use components of MeTA they learned through the exercises to complete a larger final programming project. Using different corpora with the toolkit can yield different project challenges, e.g., review summary vs. sentiment analysis.

Practitioners. Most readers in industry would most likely use this book as a reference, although we also hope that it may serve as some inspiration in your own work. As with the student user suggestion, we think you would get the most of this book by first reading the initial three chapters. Then, you may choose a chapter relevant to your current interests and delve deeper or refresh your knowledge.

Since many applications in MeTA can be used simply via config files, we anticipate it as a quick way to get a handle on your dataset and provide some baseline results without any programming required.

The exercises at the end of each chapter can be thought of as default implementations for a particular task at hand. You may choose to include MeTA in your work since it uses a permissive free software license. In fact, it is dual-licensed under MIT and University of Illinois/NCSA licenses. Of course, we still encourage and invite you to share any modifications, extensions, and improvements with MeTA that are not proprietary for the benefit of all the readers.

No matter what your goal, we hope that you find this book useful and educational. We also appreciate your comments and suggestions for improvement of the book. Thanks for reading!

Bibliographic Notes and Further Reading

There are already multiple excellent text books in information retrieval (IR). Due to the long history of research in information retrieval and the fact that much foudational work has been done in 1960s, even some very old books such as van Rijsbergen [1979] and Salton and McGill [1983] and Salton [1989] remain very useful today. Another useful early book is Frakes and Baeza-Yates [1992]. More recent ones include Grossman and Frieder [2004], Witten et al. [1999], and Belew [2008]. The most recent ones are Manning et al. [2008], Croft et al. [2009], Büttcher et al. [2010], and Baeza-Yates and Ribeiro-Neto [2011]. Compared with these books, this book has a broader view of the topic of information retrieval and attempts to cover both text retrieval and text mining. While some existing books on IR have also touched some topics such as text categorization and text clustering, which we classify as text mining topics, no previous book has included an in-depth discussion of topic mining and analysis, an important family of techniques very useful for text mining. Recommender systems also seem to be missing in the existing books on IR, which we include as an alternative way to support users for text access complementary with search engines. More importantly, this book treats all these topics in a more systematic way than existing books by framing them in a unified coherent conceptual framework for managing and analyzing big text data; the book also attempts to minimize the gap between abstract explanation of algorithms and practical applications by providing a companion toolkit for many exercises. Readers who want to know more about the history of IR research and the major early milestones should take a look at the collection of readings in Sparck Jones and Willett [1997].

The topic of text mining has also been covered in multiple books (e.g., Feldman and Sanger [2007]). A major difference between this book and those is our emphasis on the integration of text mining and information retrieval with a belief that any text data application system must involve humans in the loop and search engines are essential components of any text mining systems to support two essential functions: (1) help convert a large raw text data set into a much smaller, but more relevant text data set which can be efficiently anlayzed by using a text mining algorithm (i.e., data reduction) and (2) help users verify the source text articles from which knowledge is discovered by a text mining algorithm (i.e., knowledge provenance). As a result, this book provides a more complete coverage of techniques required for developing big text data applications.

The focus of this book is on covering algorithms that are general and robust, which can be readily applied to any text data in any natural language, often with no or minimum human effort. An evitable cost of this focus is its lack of coverage

of some key techniques important for text mining, notably the information extraction (IE) techniques which are essential for text mining. We decided not to cover IE because the IE techniques tend to be language-specific and require non-trivial manual work by humans. Another reason is that many IE techniques rely on supervised machine learning approaches, which are well covered in many existing machine learning books (see, e.g., Bishop 2006, Mitchell 1997). Readers who are interested in knowing more about IE can start with the survey book [Sarawagi 2008] and review articles [Jiang 2012].

From an application perspective, another important topic missing in this book is information visualization, which is due to our focus on the coverage of models and algorithms. However, since every application system must have a user-friendly interface to allow users to optimally interact with a system, those readers who are interested in developing text data application systems will surely find it useful to learn more about user interface design. An excellent reference to start with is Hearst [2009], which also has a detailed coverage of information visualization.

Finally, due to our emphasis on breadth, the book does not cover any component algorithm in depth. To know more about some of the topics, readers can further read books in natural language processing (e.g., Jurafsky and Martin 2009, Manning and Schütze 1999), advanced books on IR (e.g., Baeza-Yates and Ribeiro-Neto [2011]), and books on machine learning (e.g., Bishop [2006]). You may find more specific recommendations of readings relevant to a particular topic in the Bibliographic Notes at the end of each chapter that covers the corresponding topic.