Excerpt from: ChengXiang Zhai and Sean Massung: Text Data Management and Analysis. A Practical Introduction to Information Retrieval and Text Mining. Morgan & Claypool, 2016. Pages 3–13.

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.

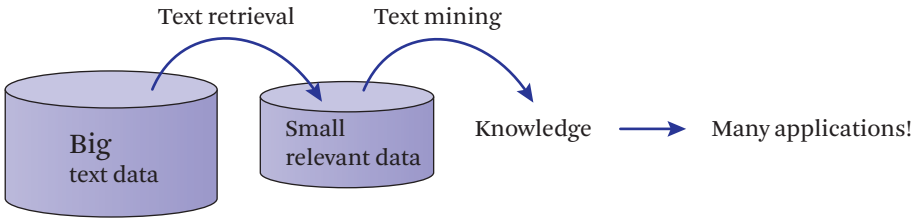

Figure 1.1 **Text retrieval and text mining are two main techniques for analyzing big text data.**

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

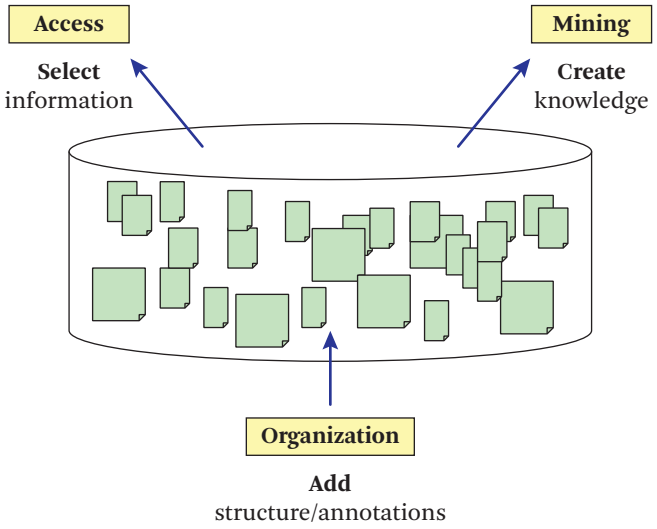

Figure 1.2 **Information access, knowledge acquisition, and text organization are three major capabilities of a text information system with text organization playing a supporting role for information access and knowledge acquisition. Knowledge acquisition is also often referred to as text mining.**

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

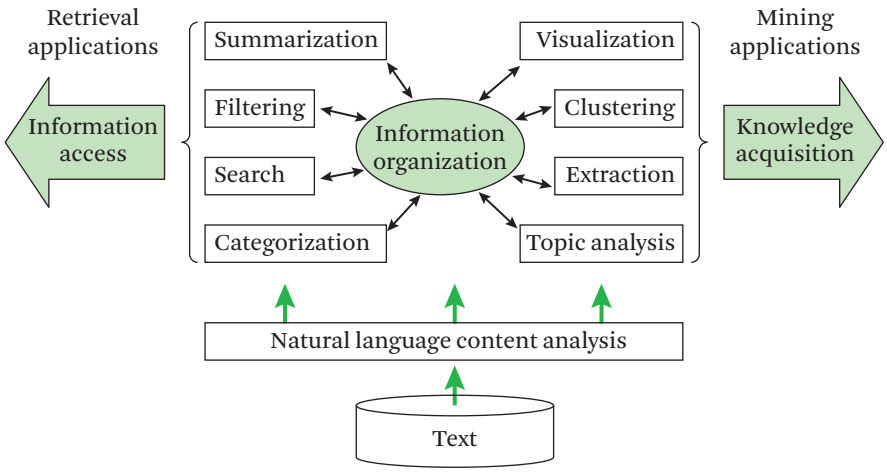

Figure 1.3 **Conceptual framework of text information systems.**

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