Chapter 1

# AN INTRODUCTION TO TEXT MINING

Charu C. Aggarwal
*IBM T. J. Watson Research Center*
*Yorktown Heights, NY*
charu@us.ibm.com

ChengXiang Zhai
*University of Illinois at Urbana-Champaign*
*Urbana, IL*
czhai@cs.uiuc.edu

**Abstract**

The problem of text mining has gained increasing attention in recent years because of the large amounts of text data, which are created in a variety of social network, web, and other information-centric applications. Unstructured data is the easiest form of data which can be created in any application scenario. As a result, there has been a tremendous need to design methods and algorithms which can effectively process a wide variety of text applications. This book will provide an overview of the different methods and algorithms which are common in the text domain, with a particular focus on mining methods.

## 1. Introduction

Data mining is a field which has seen rapid advances in recent years [8] because of the immense advances in hardware and software technology which has lead to the availability of different kinds of data. This is particularly true for the case of text data, where the development of hardware and software platforms for the web and social networks has enabled the rapid creation of large repositories of different kinds of data. In particular, the web is a technological enabler which encourages the

C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, DOI 10.1007/978-1-4614-3223-4_1,

creation of a large amount of text content by different users in a form which is easy to store and process. The increasing amounts of text data available from different applications has created a need for advances in algorithmic design which can learn interesting patterns from the data in a dynamic and scalable way.

While structured data is generally managed with a database system, text data is typically managed via a search engine due to the lack of structures [5]. A search engine enables a user to find useful information from a collection conveniently with a keyword query, and how to improve the effectiveness and efficiency of a search engine has been a central research topic in the field of information retrieval [13, 3], where many related topics to search such as text clustering, text categorization, summarization, and recommender systems are also studied [12, 9, 7].

However, research in information retrieval has traditionally focused more on facilitating information access [13] rather than analyzing information to discover patterns, which is the primary goal of text mining. The goal of information access is to connect the right information with the right users at the right time with less emphasis on processing or transformation of text information. Text mining can be regarded as going beyond information access to further help users analyze and digest information and facilitate decision making.There are also many applications of text mining where the primary goal is to analyze and discover any interesting pattterns, including trends and outliers, in text data, and the notion of a query is not essential or even relevant.

Technically, mining techniques focus on the primary models, algorithms and applications about what one can learn from different kinds of text data. Some examples of such questions are as follows:

- What are the primary supervised and unsupervised models for learning from text data? How are traditional clustering and classification problems different for text data, as compared to the traditional database literature?

- What are the useful tools and techniques used for mining text data? Which are the useful mathematical techniques which one should know, and which are repeatedly used in the context of different kinds of text data?

- What are the key application domains in which such mining techniques are used, and how are they effectively applied?

A number of key characteristics distinguish text data from other forms of data such as relational or quantitative data. This naturally affects the

mining techniques which can be used for such data. The most important characteristic of text data is that it is *sparse* and *high dimensional*. For example, a given corpus may be drawn from a lexicon of about 100,000 words, but a given text document may contain only a few hundred words. Thus, a corpus of text documents can be represented as a *sparse term-document matrix* of size $n \times d$, when $n$ is the number of documents, and $d$ is the size of the lexicon vocabulary. The $(i, j)$th entry of this matrix is the (normalized) frequency of the $j$th word in the lexicon in document $i$. The large size and the sparsity of the matrix has immediate implications for a number of data analytical techniques such as dimensionality reduction. In such cases, the methods for reduction should be specifically designed while taking this characteristic of text data into account. The variation in word frequencies and document lengths also lead to a number of issues involving document representation and normalization, which are critical for text mining.

Furthermore, text data can be analyzed at different levels of representation. For example, text data can easily be treated as a bag-of-words, or it can be treated as a string of words. However, in most applications, it would be desirable to represent text information *semantically* so that more meaningful analysis and mining can be done. For example, representing text data at the level of named entities such as people, organizations, and locations, and their relations may enable discovery of more interesting patterns than representing text as a bag of words. Unfortunately, the state of the art methods in natural language processing are still not robust enough to work well in unrestricted text domains to generate accurate semantic representation of text. Thus most text mining approaches currently still rely on the more shallow word-based representations, especially the bag-of-wrods approach, which, while losing the positioning information in the words, is generally much simpler to deal with from an algorithmic point of view than the string-based approach. In special domains (e.g., biomedical domain) and for special mining tasks (e.g., extraction of knowledge from the Web), natural language processing techniques, especially information extraction, are also playing an important role in obtaining a semantically more meaningful representation of text.

Recently, there has been rapid growth of text data in the context of different web-based applications such as social media, which often occur in the context of multimedia or other heterogeneous data domains. Therefore, a number of techniques have recently been designed for the *joint mining* of text data in the context of these different kinds of data domains. For example, the Web contains text and image data which are often intimately connected to each other and these links can be used

to improve the learning process from one domain to another. Similarly, cross-lingual linkages between documents of different languages can also be used in order to transfer knowledge from one language domain to another. This is closely related to the problem of transfer learning [11].

The rest of this chapter is organized as follows. The next section will discuss the different kinds of algorithms and applications for text mining. We will also point out the specific chapters in which they are discussed in the book. Section 3 will discuss some interesting future research directions.

## 2. Algorithms for Text Mining

In this section, we will explore the key problems arising in the context of text mining. We will also present the organization of the different chapters of this book in the context of these different problems. We intentionally leave the definition of the concept "text mining" vague to broadly cover a large set of related topics and algorithms for text analysis, spanning many different communities, including natural language processing, information retrieval, data mining, machine learning, and many application domains such as the World Wide Web and Biomedical Science. We have also intentionally allowed (sometimes significant) overlaps between chapters to allow each chapter to be relatively self contained, thus useful as a standing-alone chapter for learning about a specific topic.

**Information Extraction from Text Data:** Information Extraction is one of the key problems of text mining, which serves as a starting point for many text mining algorithms. For example, extraction of entities and their relations from text can reveal more meaningful semantic information in text data than a simple bag-of-words representation, and is generally needed to support inferences about knowledge buried in text data. Chapter 2 provides an survey of key problems in Information Extraction and the major algorithms for extracting entities and relations from text data.

**Text Summarization:** Another common function needed in many text mining applications is to summarize the text documents in order to obtain a brief overview of a large text document or a set of documents on a topic. Summarization techniques generally fall into two categories. In extractive summarization, a summary consists of information units extracted from the original text; in contrast, in abstractive summarization, a summary may contain "synthesized" information units that may not necessarily occur in the text documents. Most existing summarization methods are extractive, and in Chapter 3, we give a brief survey of these

commonly used summarization methods.

**Unsupervised Learning Methods from Text Data:** Unsupervised learning methods do not require any training data, thus can be applied to any text data without requiring any manual effort. The two main unsupervised learning methods commonly used in the context of text data are *clustering* and *topic modeling*. The problem of clustering is that of segmenting a corpus of documents into partitions, each corresponding to a topical cluster. The problems of clustering and topic modeling are closely related. In topic modeling we use a probabilistic model in order to determine a *soft* clustering, in which each document has a membership probability of the cluster, as opposed to a hard segmentation of the documents. Topic models can be considered as the process of clustering with a generative probabilistic model. Each *topic* can be considered a probability distribution over words, with the representative words having the highest probability. Each document can be expressed as a probabilistic combination of these different topics. Thus, a topic can be considered to be analogous to a cluster, and the membership of a document to a cluster is probabilistic in nature. This also leads to a more elegant cluster membership representation in cases in which the document is known to contain distinct topics. In the case of hard clustering, it is sometimes challenging to assign a document to a single cluster in such cases. Furthermore, topic modeling relates elegantly to the dimension reduction problem, where each topic provides a conceptual dimension, and the documents may be represented as a linear probabilistic combination of these different topics. Thus, topic-modeling provides an extremely general framework, which relates to both the clustering and dimension reduction problems. In chapter 4, we study the problem of clustering, while topic modeling is covered in two chapters (Chapters 5 and 8). In Chapter 5, we discuss topic modeling from the perspective of dimension reduction since the discovered topics can serve as a low-dimensional space representation of text data, where semantically related words can "match" each other, which is hard to achieve with bag-of-words representation. In chapter 8, topic modeling is discussed as a general probabilistic model for text mining.

**LSI and Dimensionality Reduction for Text Mining:** The problem of dimensionality reduction is widely studied in the database literature as a method for representing the underlying data in compressed format for indexing and retrieval [10]. A variation of dimensionality reduction which is commonly used for text data is known as *latent semantic indexing* [6]. One of the interesting characteristics of latent semantic indexing is that it brings our the key semantic aspects of the text data, which makes it more suitable for a variety of mining applications. For ex-

ample, the noise effects of synonymy and polysemy are reduced because of the use of such dimensionality reduction techniques. Another family of dimension reduction techniques are probabilistic topic models,notably PLSA, LDA, and their variants; they perform dimension reduction in a probabilistic way with potentially more meaningful topic representations based on word distributions. In chapter 5, we will discuss a variety of LSI and dimensionality reduction techniques for text data, and their use in a variety of mining applications.

**Supervised Learning Methods for Text Data:** Supervised learning methods are general machine learning methods that can exploit training data (i.e., pairs of input data points and the corresponding desired output) to learn a classifier or regression function that can be used to compute predictions on unseen new data. Since a wide range of application problems can be cast as a classification problem (that can be solved using supervised learning), the problem of supervised learning is sometimes also referred to as classification. Most of the traditional methods for text mining in the machine learning literature have been extended to solve problems of text mining. These include methods such as rule-based classifier, decision trees, nearest neighbor classifiers, maximum-margin classifiers, and probabilistic classifiers. In Chapter 6, we will study machine learning methods for automated text categorization, a major application area of supervised learning in text mining. A more general discussion of supervised learning methods is given in Chapter 8. A special class of techniques in supervised learning to address the issue of lack of training data, called *transfer learning*, are covered in Chapter 7.

**Transfer Learning with Text Data:** The afore-mentioned example of cross-lingual mining provides a case where the attributes of the text collection may be heterogeneous. Clearly, the feature representations in the different languages are heterogeneous, and it can often provide useful to transfer knowledge from one domain to another, especially when their is paucity of data in one domain. For example, labeled English documents are copious and easy to find. On the other hand, it is much harder to obtain labeled Chinese documents. The problem of transfer learning attempts to *transfer* the learned knowledge from one domain to another. Some other scenarios in which this arises is the case where we have a mixture of text and multimedia data. This is often the case in many web-based and social media applications such as *Flickr*, *Youtube* or other multimedia sharing sites. In such cases, it may be desirable to transfer the learned knowledge from one domain to another with the use of cross-media transfer. Chapter 7 provides a detailed survey of such learning techniques.

**Probabilistic Techniques for Text Mining:** A variety of probabilistic methods, particularly unsupervised topic models such as PLSA and LDA and supervised learning methods such as conditional random fields are used frequently in the context of text mining algorithms. Since such methods are used frequently in a wide variety of contexts, it is useful to create an organized survey which describes the different tools and techniques that are used in this context. In Chapter 8, we introduce the basics of the common probabilistic models and methods which are often used in the context of text mining. The material in this chapter is also relevant to many of the clustering, dimensionality reduction, topic modeling and classification techniques discussed in Chapters 4, 5, 6 and 7.

**Mining Text Streams:** Many recent applications on the web create massive streams of text data. In particular web applications such as social networks which allow the simultaneous input of text from a wide variety of users can result in a continuous stream of large volumes of text data. Similarly, news streams such as *Reuters* or aggregators such as *Google news* create large volumes of streams which can be mined continuously. Such text data are more challenging to mine, because they need to be processed in the context of a one-pass constraint [1]. The one-pass constraint essentially means that it may sometimes be difficult to store the data offline for processing, and it is necessary to perform the mining tasks continuously, as the data comes in. This makes algorithmic design a much more challenging task. In chapter 9, we study the common techniques which are often used in the context of a variety of text mining tasks.

**Cross-Lingual Mining of Text Data:** With the proliferation of web-based and other information retrieval applications to other applications, it has become particularly useful to apply mining tasks in different languages, or use the knowledge or corpora in one language to another. For example, in cross-language mining, it may be desirable to cluster a group of documents in different languages, so that documents from different languages but similar semantic topics may be placed in the same cluster. Such cross-lingual applications are extremely rich, because they can often be used to leverage knowledge from one data domain into another. In chapter 10, we will study methods for cross-lingual mining of text data, covering techniques such as machine translation, cross-lingual information retrieval, and analysis of comparable and parallel corpora.

**Text Mining in Multimedia Networks:** Text often occurs in the context of many multimedia sharing sites such as *Flickr* or *Youtube*. A natural question arises as to whether we can enrich the underlying mining process by simultaneously using the data from other domains

together with the text collection. This is also related to the problem of transfer learning, which was discussed earlier. In chapter 11, a detailed survey will be provided on mining other multimedia data together with text collections.

**Text Mining in Social Media:** One of the most common sources of text on the web is the presence of social media, which allows human actors to express themselves quickly and freely in the context of a wide range of subjects [2]. Social media is now exploited widely by commercial sites for influencing users and targeted marketing. The process of mining text in social media requires the special ability to mine dynamic data which often contains poor and non-standard vocabulary. Furthermore, the text may occur in the context of linked social networks. Such links can be used in order to improve the quality of the underlying mining process. For example, methods that use both link and content [4] are widely known to provide much more effective results which use only content or links. Chapter 12 provides a detailed survey of text mining methods in social media.

**Opinion Mining from Text Data:** A considerable amount of text on web sites occurs in the context of product reviews or opinions of different users. Mining such opinionated text data to reveal and summarize the opinions about a topic has widespread applications, such as in supporting consumers for optimizing decisions and business intelligence. spam opinions which are not useful and simply add noise to the mining process. Chapter 13 provides a detailed survey of models and methods for opinion mining and sentiment analysis.

**Text Mining from Biomedical Data:** Text mining techniques play an important role in both enabling biomedical researchers to effectively and efficiently access the knowledge buried in large amounts of literature and supplementing the mining of other biomedical data such as genome sequences, gene expression data, and protein structures to facilitate and speed up biomedical discovery. As a result, a great deal of research work has been done in adapting and extending standard text mining methods to the biomedical domain, such as recognition of various biomedical entities and their relations, text summarization, and question answering. Chapter 14 provides a detailed survey of the models and methods used for biomedical text mining.

## 3. Future Directions

The rapid growth of online textual data creates an urgent need for powerful text mining techniques. As an interdisciplinary field, text data mining spans multiple research communities, especially data mining,

natural language processing, information retrieval, and machine learning with applications in many different areas, and has attracted much attention recently. Many models and algorithms have been developed for various text mining tasks have been developed as we discussed above and will be surveyed in the rest of this book.

Looking forward, we see the following general future directions that are promising:

- **Scalable and robust methods for natural language understanding:** Understanding text information is fundamental to text mining. While the current approaches mostly rely on bag of words representation, it is clearly desirable to go beyond such a simple representation. Information extraction techniques provide one step forward toward semantic representation, but the current information extraction methods mostly rely on supervised learning and generally only work well when sufficient training data are available, restricting its applications. It is thus important to develop effective and robust information extraction and other natural language processing methods that can scale to multiple domains.

- **Domain adaptation and transfer learning:** Many text mining tasks rely on supervised learning, whose effectiveness highly depends on the amount of training data available. Unfortunately, it is generally labor-intensive to create large amounts of training data. Domain adaptation and transfer learning methods can alleviate this problem by attempting to exploit training data that might be available in a related domain or for a related task. However, the current approaches still have many limitations and are generally inadequate when there is no or little training data in the target domain. Further development of more effective domain adaptation and transfer learning methods is necessary for more effective text mining.

- **Contextual analysis of text data:** Text data is generally associated with a lot of context information such as authors, sources, and time, or more complicated information networks associated with text data. In many applications, it is important to consider the context as well as user preferences in text mining. It is thus important to further extend existing text mining approaches to further incorporate context and information networks for more powerful text analysis.

- **Parallel text mining:** In many applications of text mining, the amount of text data is huge and is likely increasing over time,

thus it is infeasible to store the data in one machine, making it necessary to develop parallel text mining algorithms that can run on a cluster of computers to perform text mining tasks in parallel. In particular, how to parallelize all kinds of text mining algorithms, including both unsupervised and supervised learning methods is a major future challenge. This direction is clearly related to cloud computing and data-intensive computing, which are growing fields themselves.

## References

[1] C. Aggarwal. *Data Streams: Models and Algorithms*, Springer, 2007.

[2] C. Aggarwal. *Social Network Data Analytics*, Springer, 2011.

[3] R. A. Baeza-Yates, B. A. Ribeiro-Neto, *Modern Information Retrieval - the concepts and technology behind search, Second edition*, Pearson Education Ltd., Harlow, England, 2011.

[4] S. Chakrabarti, B. Dom, P. Indyk. Enhanced Hypertext Categorization using Hyperlinks, *ACM SIGMOD Conference*, 1998.

[5] W. B. Croft, D. Metzler, T. Strohma, *Search Engines - Information Retrieval in Practice*, Pearson Education, 2009.

[6] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman. Indexing by Latent Semantic Analysis. *JASIS*, 41(6), pp. 391–407, 1990.

[7] D. A. Grossman, O. Frieder, *Information Retrieval: Algorithms and Heuristics (The Kluwer International Series on Information Retrieval)*, Springer-Verlag New York, Inc, 2004.

[8] J. Han, M. Kamber. *Data Mining: Concepts and Techniques*, 2nd Edition, Morgan Kaufmann, 2005.

[9] C. Manning, P. Raghavan, H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[10] I. T. Jolliffee. Principal Component Analysis. *Springer*, 2002.

[11] S. J. Pan, Q. Yang. A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering*, 22(10): pp 1345–1359, Oct. 2010.

[12] G. Salton. *An Introduction to Modern Information Retrieval*, Mc Graw Hill, 1983.

[13] K. Sparck Jones P. Willett (ed.). *Readings in Information Retrieval*, Morgan Kaufmann Publishers Inc, 1997.