1 Introduction

When you read this book, you, without doubt, already know what the **World Wide Web** is and have used it extensively. The World Wide Web (or the **Web** for short) has impacted on almost every aspect of our lives. It is the biggest and most widely known information source that is easily accessible and searchable. It consists of billions of interconnected documents (called **Web pages**) which are authored by millions of people. Since its inception, the Web has dramatically changed our information seeking behavior. Before the Web, finding information means asking a friend or an expert, or buying/borrowing a book to read. However, with the Web, everything is only a few clicks away from the comfort of our homes or offices. Not only can we find needed information on the Web, but we can also easily share our information and knowledge with others.

The Web has also become an important channel for conducting businesses. We can buy almost anything from online stores without needing to go to a physical shop. The Web also provides convenient means for us to communicate with each other, to express our views and opinions on anything, and to discuss with people from anywhere in the world. The Web is truly a **virtual society**. In this first chapter, we introduce the Web, its history, and the topics that we will study in the book.

1.1 What is the World Wide Web?

The World Wide Web is officially defined as a "wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents." In simpler terms, the Web is an Internet-based computer network that allows users of one computer to access information stored on another through the world-wide network called the **Internet**.

The Web's implementation follows a standard **client-server** model. In this model, a user relies on a program (called the **client**) to connect to a remote machine (called the **server**) where the data is stored. Navigating through the Web is done by means of a client program called the **browser**, e.g., Netscape, Internet Explorer, Firefox, etc. Web browsers work by sending requests to remote servers for information and then interpreting

the returned documents written in HTML and laying out the text and graphics on the user's computer screen on the client side.

The operation of the Web relies on the structure of its **hypertext** documents. Hypertext allows Web page authors to link their documents to other related documents residing on computers anywhere in the world. To view these documents, one simply follows the links (called **hyperlinks**).

The idea of hypertext was invented by Ted Nelson in 1965 [403], who also created the well known hypertext system Xanadu (http://xanadu.com/). Hypertext that also allows other media (e.g., image, audio and video files) is called **hypermedia**.

1.2 A Brief History of the Web and the Internet

Creation of the Web: The Web was invented in 1989 by **Tim Berners-Lee**, who, at that time, worked at CERN (Centre European pour la Recherche Nucleaire, or European Laboratory for Particle Physics) in Switzerland. He coined the term "World Wide Web," wrote the first World Wide Web server, httpd, and the first client program (a browser and editor), "**WorldWideWeb**".

It began in March 1989 when Tim Berners-Lee submitted a proposal titled "Information Management: A Proposal" to his superiors at CERN. In the proposal, he discussed the disadvantages of hierarchical information organization and outlined the advantages of a hypertext-based system. The proposal called for a simple protocol that could request information stored in remote systems through networks, and for a scheme by which information could be exchanged in a common format and documents of individuals could be linked by hyperlinks to other documents. It also proposed methods for reading text and graphics using the display technology at CERN at that time. The proposal essentially outlined a **distributed hypertext system**, which is the basic architecture of the Web.

Initially, the proposal did not receive the needed support. However, in 1990, Berners-Lee re-circulated the proposal and received the support to begin the work. With this project, Berners-Lee and his team at CERN laid the foundation for the future development of the Web as a distributed hypertext system. They introduced their server and browser, the protocol used for communication between clients and the server, the HyperText Transfer Protocol (**HTTP**), the HyperText Markup Language (**HTML**) used for authoring Web documents, and the Universal Resource Locator (**URL**). And so it began.

Mosaic and Netscape Browsers: The next significant event in the development of the Web was the arrival of **Mosaic**. In February of 1993, Marc Andreesen from the University of Illinois' NCSA (National Center for Supercomputing Applications) and his team released the first "Mosaic for X" graphical Web browser for UNIX. A few months later, different versions of Mosaic were released for Macintosh and Windows operating systems. This was an important event. For the first time, a Web client, with a consistent and simple point-and-click graphical user interface, was implemented for the three most popular operating systems available at the time. It soon made big splashes outside the academic circle where it had begun. In mid-1994, Silicon Graphics founder Jim Clark collaborated with Marc Andreessen, and they founded the company **Mosaic Communications** (later renamed as **Netscape Communications**). Within a few months, the **Netscape** browser was released to the public, which started the explosive growth of the Web. The **Internet Explorer** from Microsoft entered the market in August, 1995 and began to challenge Netscape.

The creation of the World Wide Web by Tim Berners-Lee followed by the release of the Mosaic browser are often regarded as the two most significant contributing factors to the success and popularity of the Web.

Internet: The Web would not be possible without the Internet, which provides the communication network for the Web to function. The **Internet** started with the computer network **ARPANET** in the Cold War era. It was produced as the result of a project in the United States aiming at maintaining control over its missiles and bombers after a nuclear attack. It was supported by Advanced Research Projects Agency (ARPA), which was part of the Department of Defense in the United States. The first ARPANET connections were made in 1969, and in 1972, it was demonstrated at the First International Conference on Computers and Communication, held in Washington D.C. At the conference, ARPA scientists linked computers together from 40 different locations.

In 1973, Vinton Cerf and Bob Kahn started to develop the protocol later to be called **TCP/IP** (**Transmission Control Protocol/Internet Protocol**). In the next year, they published the paper "Transmission Control Protocol", which marked the beginning of TCP/IP. This new protocol allowed diverse computer networks to interconnect and communicate with each other. In subsequent years, many networks were built, and many competing techniques and protocols were proposed and developed. However, ARPANET was still the backbone to the entire system. During the period, the network scene was chaotic. In 1982, the TCP/IP was finally adopted, and the **Internet**, which is a connected set of networks using the TCP/IP protocol, was born.

Search Engines: With information being shared worldwide, there was a need for individuals to find information in an orderly and efficient manner. Thus began the development of search engines. The search system **Excite** was introduced in 1993 by six Stanford University students. **EINet Galaxy** was established in 1994 as part of the MCC Research Consortium at the University of Texas. Jerry Yang and David Filo created **Yahoo!** in 1994, which started out as a listing of their favorite Web sites, and offered directory search. In subsequent years, many search systems emerged, e.g., **Lycos**, **Inforseek**, **AltaVista**, **Inktomi**, **Ask Jeeves**, **Northernlight**, etc.

Google was launched in 1998 by Sergey Brin and Larry Page based on their research project at Stanford University. Microsoft started to commit to search in 2003, and launched the **MSN** search engine in spring 2005. It used search engines from others before. **Yahoo!** provided a general search capability in 2004 after it purchased Inktomi in 2003.

W3C (**The World Wide Web Consortium**): W3C was formed in the December of 1994 by MIT and CERN as an international organization to lead the development of the Web. W3C's main objective was "to promote standards for the evolution of the Web and interoperability between WWW products by producing specifications and reference software." The first **International Conference on World Wide Web** (**WWW**) was also held in 1994, which has been a yearly event ever since.

From 1995 to 2001, the growth of the Web boomed. Investors saw commercial opportunities and became involved. Numerous businesses started on the Web, which led to irrational developments. Finally, the bubble burst in 2001. However, the development of the Web was not stopped, but has only become more rational since.

1.3 Web Data Mining

The rapid growth of the Web in the last decade makes it the largest publicly accessible data source in the world. The Web has many unique characteristics, which make mining useful information and knowledge a fascinating and challenging task. Let us review some of these characteristics.

1. The amount of data/information on the Web is huge and still growing. The coverage of the information is also very wide and diverse. One can find information on almost anything on the Web.
2. Data of all types exist on the Web, e.g., structured tables, semi-structured Web pages, unstructured texts, and multimedia files (images, audios, and videos).

3. Information on the Web is **heterogeneous**. Due to the diverse authorship of Web pages, multiple pages may present the same or similar information using completely different words and/or formats. This makes integration of information from multiple pages a challenging problem.
4. A significant amount of information on the Web is linked. Hyperlinks exist among Web pages within a site and across different sites. Within a site, hyperlinks serve as information organization mechanisms. Across different sites, hyperlinks represent implicit conveyance of authority to the target pages. That is, those pages that are linked (or pointed) to by many other pages are usually high quality pages or **authoritative pages** simply because many people trust them.
5. The information on the Web is noisy. The **noise** comes from two main sources. First, a typical Web page contains many pieces of information, e.g., the **main content** of the page, navigation links, advertisements, copyright notices, privacy policies, etc. For a particular application, only part of the information is useful. The rest is considered noise. To perform fine-grain Web information analysis and data mining, the noise should be removed. Second, due to the fact that the Web does not have quality control of information, i.e., one can write almost anything that one likes, a large amount of information on the Web is of low quality, erroneous, or even misleading.
6. The Web is also about services. Most commercial Web sites allow people to perform useful operations at their sites, e.g., to purchase products, to pay bills, and to fill in forms.
7. The Web is dynamic. Information on the Web changes constantly. Keeping up with the change and monitoring the change are important issues for many applications.
8. The Web is a virtual society. The Web is not only about data, information and services, but also about interactions among people, organizations and automated systems. One can communicate with people anywhere in the world easily and instantly, and also express one's views on anything in Internet forums, blogs and review sites.

All these characteristics present both challenges and opportunities for mining and discovery of information and knowledge from the Web. In this book, we only focus on mining textual data. For mining of images, videos and audios, please refer to [143, 441].

To explore information mining on the Web, it is necessary to know data mining, which has been applied in many Web mining tasks. However, Web mining is not entirely an application of data mining. Due to the richness and diversity of information and other Web specific characteristics discussed above, Web mining has developed many of its own algorithms.

1.3.1 What is Data Mining?

Data mining is also called **knowledge discovery in databases** (**KDD**). It is commonly defined as the process of discovering useful **patterns** or knowledge from data sources, e.g., databases, texts, images, the Web, etc. The patterns must be valid, potentially useful, and understandable. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization.

There are many data mining tasks. Some of the common ones are **supervised learning** (or **classification**), **unsupervised learning** (or **clustering**), **association rule mining**, and **sequential pattern mining**. We will study all of them in this book.

A data mining application usually starts with an understanding of the application domain by **data analysts** (**data miners**), who then identify suitable data sources and the target data. With the data, data mining can be performed, which is usually carried out in three main steps:

- **Pre-processing**: The raw data is usually not suitable for mining due to various reasons. It may need to be cleaned in order to remove noises or abnormalities. The data may also be too large and/or involve many irrelevant attributes, which call for data reduction through sampling and attribute selection. Details about data pre-processing can be found in any standard data mining textbook.
- **Data mining**: The processed data is then fed to a data mining algorithm which will produce patterns or knowledge.
- **Post-processing**: In many applications, not all discovered patterns are useful. This step identifies those useful ones for applications. Various evaluation and visualization techniques are used to make the decision.

The whole process (also called the **data mining process**) is almost always iterative. It usually takes many rounds to achieve final satisfactory results, which are then incorporated into real-world operational tasks.

Traditional data mining uses structured data stored in relational tables, spread sheets, or flat files in the tabular form. With the growth of the Web and text documents, **Web mining** and **text mining** are becoming increasingly important and popular. Web mining is the focus of this book.

1.3.2 What is Web Mining?

Web mining aims to discover useful information or knowledge from the **Web hyperlink structure**, **page content**, and **usage data**. Although Web mining uses many data mining techniques, as mentioned above it is not

purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data. Many new mining tasks and algorithms were invented in the past decade. Based on the primary kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

- **Web structure mining**: Web structure mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links, we can discover important Web pages, which, incidentally, is a key technology used in search engines. We can also discover communities of users who share common interests. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.
- **Web content mining**: Web content mining extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics. These tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc, for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments. These are not traditional data mining tasks.
- **Web usage mining**: Web usage mining refers to the discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining applies many data mining algorithms. One of the key issues in Web usage mining is the pre-processing of clickstream data in usage logs in order to produce the right data for mining.

In this book, we will study all these three types of mining. However, due to the richness and diversity of information on the Web, there are a large number of Web mining tasks. We will not be able to cover them all. We will only focus on some important tasks and their algorithms.

The **Web mining process** is similar to the data mining process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a data warehouse. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages. We will devote a whole chapter on crawling.

Once the data is collected, we go through the same three-step process: data pre-processing, Web data mining and post-processing. However, the techniques used for each step can be quite different from those used in traditional data mining.

1.4 Summary of Chapters

This book consists of two main parts. The first part, which includes Chaps. 2–5, covers the major topics of data mining. The second part, which comprises the rest of the chapters, covers Web mining (including a chapter on Web search). In the Web mining part, Chaps. 7 and 8 are on Web structure mining, which are closely related to Web search (Chap. 6). Since it is difficult to draw a boundary between Web search and Web mining, Web search and mining are put together. Chaps 9–11 are on Web content mining, and Chap. 12 is on Web usage mining. Below we give a brief introduction to each chapter.

Chapter 2 – Association Rules and Sequential Patterns: This chapter studies two important data mining models that have been used in many Web mining tasks, especially in Web usage and content mining. Association rule mining finds sets of data items that occur together frequently. Sequential pattern mining finds sets of data items that occur together frequently in some sequences. Clearly, they can be used to find regularities in the Web data. For example, in Web usage mining, association rule mining can be used to find users' visit and purchase patterns, and sequential pattern mining can be used to find users' navigation patterns.

Chapter 3 – Supervised Learning: Supervised learning is perhaps the most frequently used mining/learning technique in both practical data mining and Web mining. It is also called **classification**, which aims to learn a classification function (called a **classifier**) from data that are labeled with pre-defined classes or categories. The resulting classifier is then applied to classify future data instances into these classes. Due to the fact that the data instances used for learning (called the **training data**) are labeled with pre-defined classes, the method is called supervised learning.

Chapter 4 – Unsupervised Learning: In unsupervised learning, the data used for learning has no pre-defined classes. The learning algorithm has to find the hidden structures or regularities in the data. One of the key unsupervised learning techniques is **clustering**, which organizes data instances into **groups** or **clusters** according to their similarities (or differences). Clustering is widely used in Web mining. For example, we can cluster Web pages into groups, where each group may represent a particular topic. We can also cluster documents into a hierarchy of clusters, which may represent a topic hierarchy.

Chapter 5 – Partially Supervised Learning: Supervised learning requires a large number of labeled data instances to learn an accurate classifier. Labeling, which is often done manually, is labor intensive and time

consuming. To reduce the manual labeling effort, **learning from labeled and unlabeled examples** (or **LU learning**) was proposed to use a small set of labeled examples (data instances) and a large set of unlabeled examples for learning. This model is also called **semi-supervised learning**.

Another learning model that we will study is called **learning from positive and unlabeled examples** (or **PU learning**), which is for two-class classification. However, there are no labeled negative examples for learning. This model is useful in many situations. For example, we have a set of Web mining papers and we want to identify other Web mining papers in a research paper repository which contains all kinds of papers. The set of Web mining papers can be treated as the positive data, and the papers in the research repository can be treated as the unlabeled data.

Chapter 6 – Information Retrieval and Web Search: Search is probably the largest application on the Web. It has its root in **information retrieval** (or IR for short), which is a field of study that helps the user find needed information from a large collection of text documents. Given a query (e.g., a set of **keywords**), which expresses the user's information need, an IR system finds a set of documents that is relevant to the query from its underlying collection. This is also how a Web search engine works.

Web search brings IR to a new height. It applies some IR techniques, but also presents a host of interesting problems due to special characteristics of the Web data. First of all, Web pages are not the same as plain text documents because they are semi-structured and contain hyperlinks. Thus, new methods have been designed to produce better Web IR (or search) systems. Another major issue is efficiency. Document collections used in traditional IR systems are not large, but the number of pages on the Web is huge. For example, Google claimed that it indexed more than 8 billion pages when this book was written. Web users demand very fast responses. No matter how effective a retrieval algorithm is, if the retrieval cannot be done extremely efficiently, few people will use it. In the chapter, several other search related issues will also be discussed.

Chapter 7 – Link Analysis: Hyperlinks are a special feature of the Web, which have been exploited for many purposes, especially for Web search. Google's success is largely attributed to its hyperlink-based ranking algorithm called **PageRank**, which is originated from **social network analysis**. In this chapter, we will first introduce some main concepts of social network analysis and then describe two most well known Web link analysis algorithms, PageRank and HITS. In addition, we will also study several community finding algorithms. When Web pages link to one another, they form Web communities, which are groups of content creators that share

some common interests. Communities not only manifest in hyperlinks, but also in other contexts such as emails and Web page contents.

Chapter 8 – Web Crawling: A Web **crawler** is a program that automatically traverses the Web's hyperlink structure and downloads each linked page to a local storage. Crawling is often the first step of Web mining or in building a Web search engine. Although conceptually easy, building a practical crawler is by no means simple. Due to efficiency and many other concerns, it involves a great deal of engineering. There are two types of crawlers: **universal crawlers** and **topic crawlers**. A universal crawler downloads all pages irrespective of their contents, while a topic crawler downloads only pages of certain topics. The difficulty in topic crawling is how to recognize such pages. We will study several techniques for this purpose.

Chapter 9 – Structured Data Extraction: Wrapper Generation: A large number of pages on the Web contain structured data, which are usually data records retrieved from underlying databases and displayed in Web pages following some fixed templates. Structured data often represent their host pages' essential information, e.g., lists of products and services. Extracting such data allows one to provide value added services, e.g., comparative shopping, and meta-search. There are two main approaches to extraction. One is the supervised approach, which uses supervised learning to learn data extraction rules. The other is the unsupervised pattern discovery approach, which finds repeated patterns (hidden templates) in Web pages for data extraction.

Chapter 10 – Information Integration: Due to diverse authorships of the Web, different Web sites typically use different words or terms to express the same or similar information. In order to make use of the data or information extracted from multiple sites to provide value added services, we need to semantically integrate the data/information from these sites in order to produce consistent and coherent databases. Intuitively, integration means to match columns in different data tables that contain the same type of information (e.g., product names) and to match data values that are semantically the same but expressed differently in different sites.

Chapter 11 – Opinion Mining: Apart from structured data, the Web also contains a huge amount of unstructured text. Analyzing such text is also of great importance. It is perhaps even more important than extracting structured data because of the sheer volume of valuable information of almost any imaginable types contained in it. This chapter will only focus on mining people's **opinions** or **sentiments** expressed in **product reviews**, **forum discussions** and **blogs**. The task is not only technically challenging,

but also very useful in practice because businesses and organizations always want to know consumer opinions on their products and services.

Chapter 12 – Web Usage Mining: Web usage mining aims to study user clicks and their applications to e-commerce and business intelligence. The objective is to capture and model **behavioral patterns** and **profiles** of users who interact with a Web site. Such patterns can be used to better understand the behaviors of different user segments, to improve the organization and structure of the site, and to create **personalized experiences** for users by providing dynamic **recommendations** of products and services.

1.5 How to Read this Book

This book is a textbook although two chapters are contributed by two other researchers. The contents of the two chapters have been carefully edited and integrated into the common framework of the whole book. The book is suitable for both graduate students and senior undergraduate students in the fields of computer science, information science, engineering, statistics, and social science. It can also be used as a reference by researchers and practitioners who are interested in or are working in the field of Web mining, data mining or text mining.

As mentioned earlier, the book is divided into two parts. Part I (Chaps. 2–5) covers the major topics of data mining. Text classification and clustering are included in this part as well. Part II, which includes the rest of the chapters, covers Web mining (and search). In general, all chapters in Part II require some techniques in Part I. Within each part, the dependency is minimal except Chap. 5, which needs several techniques from Chap. 4.

To Instructors: This book can be used as a class text for a one-semester course on Web data mining. In this case, there are two possibilities. If the students already have data mining or machine learning background, the chapters in Part I can be skipped. If the students do not have any data mining background, I recommend covering some selected sections from each chapter of Part I before going to Part II. The chapters in Part II can be covered in any sequence. You can also select a subset of the chapters according to your needs.

The book may also be used as a class text for an introductory data mining course where Web mining concepts and techniques are introduced. In this case, I recommend first covering all the chapters in Part I and then selectively covering some chapters or sections from each chapter in Part II depending on needs. It is usually a good idea to cover some sections of

Chaps. 6 and 7 as search engines fascinate most students. I also recommend including one or two lectures on data pre-processing for data mining since the topic is important for practical data mining applications but is not covered in this book. You can find teaching materials on data pre-processing from most introductory data mining books.

Supporting Materials: Updates to chapters and teaching materials, including lecture slides, data sets, implemented algorithms, and other resources, are available at http://www.springer.com/3-540-37881-2.

Bibliographic Notes

The W3C Web site (http://www.w3.org/) is the most authoritative resource site for information on Web developments, standards and guidelines. The history of the Web and hypertext, and Tim Berners-Lee's original proposal can all be found there. Many other sites also contain information about the history of the Web, the Internet and search engines, e.g., http://www.elsop.com/wrc/h_web.htm, http://www.zeltser.com/web-history/, http://www.isoc.org/internet/history/, http://www.livinginternet.com, http://www.w3c.rl.ac.uk/primers/history/origins.htm and http://searchenginewatch.com/.

There are some earlier introductory texts on Web mining, e.g., those by Baldi et al. [33] and Chakrabarti [85]. There are also several application oriented books, e.g., those by Linoff and Berry [338], and Thuraisingham [515], and edited volumes by Djeraba et al. [143], Scime [480], and Zhong et al. [617].

On data mining, there are many textbooks, e.g., those by Duda et al. [155], Dunham [156], Han and Kamber [218], Hand et al. [221], Larose [305], Langley [302], Mitchell [385], Roiger and Geatz [467], Tan et al. [512], and Witten and Frank [549]. Application oriented books include those by Berry and Linoff [49], Pyle [450], Parr Rud [468], and Tang and MacLennan [514]. Several edited volumes exist as well, e.g., those by Fayyad et al. [174], Grossman et al. [208], and Wang et al. [533].

Latest research results on Web mining can be found in a large number of conferences and journals (too many to list) due to the interdisciplinary nature of the field. All the journals and conferences related to the Web technology, information retrieval, data mining, databases, artificial intelligence, and machine learning may contain Web mining related papers.