Chapter 1

Introduction to Web Mining

1.1. Introduction

The World Wide Web (WWW) is a very rich and huge repository of information which is rapidly increasing with each passing day and continues to grow exponentially in its size and complexity. In this era, World Wide Web has become one of the most popular platforms for exchanging, sharing and retrieval of information. It is very cumbersome and time-consuming to retrieve relevant information from such huge amount of data because of its exponential growth and also due to the increasing complexity. So, it has become extremely essential to apply some techniques so that valuable information can be extracted and converted into knowledge. The Web undoubtedly is an interesting medium to propagate information. Before the Web in antiquity, people used to gather information from newspapers and magazines, communicate through postal services, purchase products from nearby markets and even do business manually through paper, but the Internet has really solved our millions of issues. It has changed the approach with which we interact with one another, do our business, collect information, accomplish our purchases, exchange and share different types of information with our colleagues. Due to globalization and modernization, all the above mentioned activities which we used to do manually earlier are now performed with the help of the Internet and hence results in the increase of data over the Web. Therefore, the presence of enormous and humongous information on the Web and its rapid growth paved the way for search engines so that the information on the Web can be retrieved quickly.
Search engines like Google, Bing, Yahoo and Baidu have become very vital tools for every computer user. However, search engines utmost help us by providing the relevant results and web users are compelled to go through these innumerable relevant results to find the desired information. It is because of the fact that the method followed by search engines generates relevant results instead of exact desired results.

Search engines currently are plagued by four problems. Firstly, search engines provide us hundreds of irrelevant results. Secondly, a limited coverage of Web i.e. search engines cannot index the entire Web. Thirdly, a limited query interface based on syntactic and fourthly, a limited customization to individual users. Since, it is apparent that for a web user to find the desired information on the Web using search engines is cumbersome and time consuming. Thus, it is very difficult to manage and handle this gigantic and massive data, 80% of which is unstructured, complex and disorganized that required to be analysed by traditional tools [Ferrara et al., 2014]. This data have resulted in a very complex structure of websites and therefore, on browsing and searching the intended users are facing a big challenge to find the appropriate and specific information. So, it may not be inappropriate to say, “Finding information on the Web is like finding a needle in a haystack”.

For instance, we want to find the list of Indian Universities offering doctoral programs in computer science. Or list of computer science professors in India and viewing their profiles one by one. Or State-wise population of India. To figure out this information, most likely we may Google it, but the result will turn out to be unsatisfactory. No doubt the Web arranges all this information, but it is hardly possible to do a structured query as this arranged information is all over the Web. To retrieve exactly what we want, we need to organize information about the Universities, faculties, departments and divisions etc., in a database and construct a simple query.
statement. For this we need to find techniques to map the information embedded in web pages to data records stored in a database. Such a database when combined with appropriate tools promises a number of compelling applications.

Web mining is a very broad research area emerged to solve the issues that arise due to the WWW phenomenon. Generally Web mining is termed as the process of applying Data Mining techniques and concepts to retrieve useful information from WWW. Consequently, Web mining does not only mean applying Data Mining techniques, but it also includes modification of existing approaches as well as new approaches that better suit the demands of the Web. Web mining also includes many techniques from Machine Learning like Natural Language Processing, Artificial Intelligence along with Information Retrieval disciplines. Web mining has been thus developed into an autonomous and fastest growing research area. It is an area which deals with the mining of specific information from the Web by using various techniques. In the field of Web mining research, usually following two problems are often pointed out viz. finding relevant information and finding needed information. Web users usually browse web pages directly or use search engines in order to retrieve particular information from the Web. When a user makes use of a search engine to retrieve information, he enters a query containing several keywords, the search engine in turn returns a list of ranked web pages that resembles the search query. But the two major concerns associated with web search are low precision and low recall [Kosala et al., 2000]. Low precision is caused due to the lot of irrelevant web pages returned by the search engine while as low recall is caused because of insufficiency of indexing of all web pages. A lot of web pages being un-indexed that are available on the Internet causes difficulty in retrieving the relevant information. Therefore, finding relevant web pages that best matches the query is becoming one of the popular research topics among
Web mining research scholars [Zhang et al., 2006]. Most of the search engines work only on the basis of key-word (s) entered by the user. Possibly such search engines don’t return those results that the user really needs. For example, if a computer science student wants to learn about python programming language, by just entering a python as a one-word query, he may be presented with information about the python, one kind of a poisonous snake and not with the programming language. So, in the context of web search, the semantics of web data [Ghani and Fano, 2002] is not often taken into account. The two problems mentioned above have put the web applications and existing search engines under significant trauma. A lot of efforts had been made to deal with such problems from various research domains such as Database, Data Mining and Machine Learning and have put forward a great challenge to Web mining researchers.

1.2. Web Mining

The explosive growth of WWW makes it a single largest source of data in the World. As the Web is huge, diverse, dynamic and mostly unstructured, therefore mining the web is a very enthralling challenge [Zhang and Segall, 2008]. Oren Etzioni was the first to coin the term Web mining. According to him, “Web mining is the use of Data Mining techniques to automatically discover and extract information from Web documents and services” [Etzioni, 1996]. Two approaches were taken initially to define Web mining: Process-centric view and Data-centric view. Process-centric view of Web mining can be defined as a sequence of tasks while as Data-centric view [Cooley et al., 1997], defines Web mining in terms of the types of web data being used in the extraction process. As it is evident from the approach adopted in most research papers [Kosala and Blockeel, 2000; Borges and Levene, 1998], the second definition of Web mining has become more acceptable. According to the second definition, i.e. following the Data-centric
view of Web mining, Web mining may be defined as, “The application of Data Mining techniques to extract knowledge from web data”. According to [Markov and Larose, 2007], Web mining may be defined as follows: “For Web mining, we refer to the application of Data Mining methodologies, techniques, and models to the variety of data forms, structures and usage patterns that comprise the World Wide Web”. Web mining is basically the process of extracting valuable information from the Web in order to make it more useful and more profitable. Web mining uses two of the activated research areas, namely Data mining and World Wide Web. According to [Etzioni, 1996], the process of Web mining can be divided into various subtasks namely, Resource finding, Information Selection and Pre-processing, Generalization and Analysis. These are briefly described below and shown in figure 1.1:

I). **Resource Discovery**: It is the task of retrieving the required or needed information from the webpages. Data mining techniques such as classification and clustering are used to extract information.

II). **Information Selection and Pre-processing**: It deals with the task of representing the web pages. Binary, Term Frequency (tf), Inverse Document Frequency (idf), Term Frequency-Inverse Document Frequency (tf-idf), Weight Inverse Document Frequency (widf) and Latent Semantic Indexing (LSI) are different representations of a webpage. The brief explanation of different representations is given below [Manning et al., 2009]:

- Binary also known as “Set of Words”. The relevance or weight of feature is a binary value {0, 1} depending on whether the feature appears in the document or not.
- Term Frequency $t_{f,t,d}$ of term $t$ in document $d$ is defined as the number of times that $t$ occurs in $d$. 

• The Inverse Document Frequency is,

\[
idf_t = \log \frac{N}{df_t}
\]  

(1.1)

Where \( N \) is the total number of documents in a collection and document frequency \( df_t \) is defined to be the number of documents in the collection that contains the term \( t \).

• The Term Frequency-Inverse Document Frequency, \( tf-idf \) weighting scheme assigns to term \( t \), a weight in document \( d \) given by

\[
idf_{t,d} = tf_{t,d} \times idf_t
\]

(1.2)

III). **Generalization**: It is the task of automatically generating patterns. Machine learning or Data Mining techniques can be used to identify and generate patterns in a webpage or across multiple web pages.

IV). **Analysis**: In this phase, we analyse and validate mined patterns. To measure the accuracy of the retrieved pattern, accuracy measures are used.

Based on the four subtasks (Resource Discovery, Information Extraction and Pre-processing, Generalization and Analysis), Web mining can be viewed as the use of Data Mining techniques to automatically retrieve, extract and evaluate information for Knowledge Discovery from web documents and services [Sharma et al., 2011]. The process of Web mining is very similar to the Data Mining process. The main difference between the two is mainly in the data collection phase. In Web mining data is collected from various sources such as from web pages and web logs and is a substantial task which involves Crawling the large number of web pages, but in Data Mining, data are usually collected from an already stored Datawarehouse.
1.2.1. Types of Web Data

The Web consists of various forms of data. [Madria et al., 1999], asserts that the Web involves three types of data: data on the Web, Web log data and Web structure data. [Cooley et al., 2000], categorized the web data as content data, structure data, usage data, and user profile data. Based on the classification of what part of web data to mine the three most accepted categories of the Web mining are Web content mining, Web structure mining, and Web usage mining [Borges and Levene, 1998; Kosala and Blockeel, 2000; Madria et al., 1999]. Thus, the web data is generally categorized into three main types: web content data, web structure data and web usage data shown in figure 1.2.
Figure 1.2: Web Data Types

1.2.1.1. Web Content Data: Web content data are the data (texts, audios, videos and images) on web pages and other documents. Hypertext Markup Language (HTML) web pages are the common form of web content data. While viewing the HTML webpage in a browser, it may appear differently in different kinds of browsers, but the basic structure is same. Other forms of content data are Extensible Markup Language (XML) documents and dynamic server pages such as Pre-Hypertext Processor (PHP), Java Server Page (JSP) and Active Server Page (ASP).

1.2.1.2. Web Structure Data: Web structure data is the data from hyperlinks between web pages. It includes intra-document structure and inter-document structure. In a particular webpage, intra-page structure information is the information about arrangement of different HTML tags and inter-page structure information is the information about web pages connected to other web pages. The whole structure of a website can be represented using hyperlinks which form a graph known as Web graph. The Web graph depicts the overall structure of a website. In a graph, a webpage represents a node and a hyperlink connecting one page to another represents an edge. An edge in a graph can be an outgoing arc or an incoming arc. Web graph is mainly used in indexing and searching.
1.2.3. Web Usage Data: Web usage data is the data contained in http logs, app server logs, etc. Web log data is generated by a Web server when a user interacts with a website. This type of data is used to track different kinds of business events. Depending on the source of collection of data, the web usage data can be categorized into three main types: On the server side, the client side, and also the Proxy side. Generally, when a user requests for a webpage to web server, the web server processes the request and the webpage is sent to the user. This interaction between a user and a web server generates data on the server and all the information is recorded and stored in server log files. Web server on different kinds of usage logs are generated such as an access log (which stores information about IP Address, username, date and time request), agent log (that stores information about browser, its version and operating system of a user) and error log (which stores information about file not found, no data, aborted transmission). Log files of a Web Server can be analysed using software known as a traffic Analyser, which can help us to improve the structure of a website.

1.2.2. Characteristics of Web Data

Data on the Web is different and has its own features than the data stored in traditional database systems. Usually Web data show the following characteristics:

- The data on the Web is huge: It is very difficult to estimate the exact volume of data available on the Internet because the Web is growing at exponential growth with each passing day. In 1994, the World Wide Web Worm (W WWW), one of the first Web search engines, had an index of 110,000 Web pages and Web accessible documents [McBryan, 1994]. But as of now May, 2016, the top search engines can index at least 4.57 billion pages. This enormous and gigantic volume of data available on the Web makes traditional techniques difficult to handle it.
• The data on the Web is heterogeneous and distributed: As we know, Web is an interconnection of various computers over the Internet. Due to this property web data is usually scattered across a number of computers or servers, which are placed around the World at different locations. In addition to textual data, web pages also contain other forms of data such as images, audios and videos. To deal with such type of data, different web processing techniques are required.

• The data on the Web is noisy: A typical webpage in addition to the main content often contains large amount of noisy information such as banner advertisements, copyright notices, navigation bars, etc. This noisy data being irrelevant to user degrades and hampers the extraction process. So for web miners, it is important for them to differentiate the main content from noisy data.

• The data on the Web is dynamic: Because of the Web 2.0, as the number of web users increases with each day, the number of web pages also grows significantly. It is because of the fact that it is now very easy to upload the web pages and publish the content on the Web. As a result, the data on the Web is being updated and changed frequently. Monitoring and keeping the changes is one of the important issues.

All these features of the web data mentioned above clearly differentiate the web data from the data stored in traditional databases. These features present lot of challenges and also opportunities in order to extract accurately the useful information from the Web.

1.2.3. Taxonomy of Web Mining
As Web involves three types of data viz. Web Structure Data, Web Log Data and the Actual Data on the WWW, and as such Web Mining is divided into three types, namely Web Structure Mining, Web Usage Mining and Web
Content Mining respectively. [Kosala and Blockeel, 2000] also suggested three categories of Web mining based on the type of data being mined i.e. mining for content known as Web content mining, mining for link structure known as Web structure mining or mining for navigational patterns known as Web usage mining. According to [Chakrabarti, 2003], Web mining is an umbrella term used to describe three different types of Data Mining, i.e. content mining, usage mining and structure mining. The classification of Web mining is shown below in Figure 1.3.

![Web Mining Taxonomy](image)

**Figure 1.3: Web Mining Taxonomy**

1.2.3.1. **Web Content Mining:** This category of Web mining mines or extracts valuable information from contents of web pages. Web content may consist of texts, images, audios or videos. Web content mining does not provide any information about the structure of content, but deals with the extraction of information from web pages. It targets knowledge discovery in which it collects information from text documents, multimedia documents such as images and videos, which are linked in a Web. Web content mining techniques mine structured, semi-structured and multimedia data. There are two views of web content mining, Information Retrieval (IR) view and
Database view [Kosala and Blockeel, 2000]. The main aim of IR view of web content mining is to assist or improve the information finding or filtering of information to users based on user profiles. The database view of web content mining is to model data on the Web and to integrate them for more sophisticated queries. Clustering, Classification and Association Rules are some of the techniques used by web content mining. The real world example of web content mining is Clustering of profiles to get a prediction of the users behavior.

1.2.3.2. **Web Usage Mining**: This category of Web mining refers to the process of identifying user access patterns by analyzing users usage data on the Web. By analysing users usage data one can record his behaviour while browsing or making transactions and thus can better understand and serve the user depending on his needs. Web usage data can be easily captured by web servers, web proxies and client applications. It can also help us to identify those users who have accessed similar webpages and thus it can be very useful in collaborative web searching and filtering. Such as Amazon.com uses collaborative filtering to suggest books to those customers based on the liking of other customers having similar interest or purchasing histories. In 2002, Haung and his co-workers in Taiwan used Hopfield Net to model user interests and product profiles in an online bookstore [Chen,2002]. Also, there are web log analysis programs some are free such as Webalizer and Google Analytics and some are commercial such as Web Trends and Web Side Story. Google Analytic is a service offered by Google, generates detailed statistics about the visits to a website. A typical web log record looks:

```
"145.208.78.51-[26/Feb/2012:10:16:30-0500] "HTTP/.........html HTTP/1.0 2002781"
```
Where 145.208.78.51 is a client IP address, - is the user log name for anonymous user, 26/ Feb/ 2010 : 10: 16: 30- 0500 is the date and time zone that shows the date and time when a user sends a request, http:// .... html/ HTTP/ 1.0 is the URL of a webpage that the user has requested, 200 is a status code returned to the user and 2781 is the number of bytes transferred.

1.2.3.3. Web Structure Mining: This category of Web mining is defined as the process of discovering structural information from the Web. According to Kosala et al., Web structure mining is defined as a tool that defines the structural relationship between web pages [Kosala and Blockeel, 2000]. This type of mining can be performed at either intra-page (document) level or at inter-page (hyperlink level). The research at an inter - page level is known also as hyperlink analysis. It generates structural summary about web pages by analysing the link. The Web is treated as a graph, whereas web pages are nodes and hyperlinks are the edges connecting the nodes as shown in figure 1.4. It focuses on the inter-document structure within the Web to discover the link structure of hyperlinks. It can be used to discover the structure of web documents itself and the nature of the network on the website. The model can be used to categorize the web pages and integrate different web pages. Thus, it helps us in discovering similarities between the websites. Hyperlink Induced Topic Search (HITS) and PageRank (PR) are most popular algorithms that focus on the link structure of the Web to find the importance of the webpage. The real world example of PageRank Algorithm is PageRank Checker, which is an algorithm used by Google Search to Rank websites in their search engine results.
1.3. Web Content Mining

The process of extracting useful and valuable information from web pages is known as Web content mining. [Linoff and Berry, 2001] define web content mining as “the process of extracting useful information from the text, images and other forms of content that make up the page”. According to [Kuhlins and Tredwell, 2002], 80% of the websites available on the Web get their data from an underlying databases upon a request by a user and content mining can be thought of as the reverse of this process. Web content mining is related to Data Mining and Text Mining as a lot of Data Mining techniques are useful in web content mining and a large amount of the web contents are texts. On the other hand, it is different from Data Mining and Text Mining because web data are structured, semi-structured and/or unstructured, while as Data Mining deals with structured data and Text Mining with unstructured data. Web content mining requires inventive applications of Data Mining and/or techniques of text mining and also its own distinctive approaches. In the area of web content mining, there has been a rapid development and expansion of activities in the past few years because of the exceptional growth of the Web contents and most importantly because of the
various benefits of such mining. Multimedia Data Mining is the part of the web content mining, which can be defined as the process of mining online multimedia sources. This research area has recently gained a lot of popularity, but it is still in its infancy and lot of work is required to be done. [Oh et al., 2003] propose a technique which could be used to mine video data in order to find and extract interesting patterns from the motions of videos. However, due to the lack of structure and heterogeneity of Web data, automatic extraction of targeted information from Web still presents many challenges. The various research problems in the field of web content mining are briefly discussed in following sections:

- **Data/information extraction**: Extraction of data from web pages allows one to provide services. Our focus of research is on the extraction of structured data from Web pages, such as extraction of products from e-commerce web pages.

- **Web information integration and schema matching**: No doubt the Web contains an enormous amount of data. But the same kind of information on different websites or even on a single webpage has been represented in a different fashion. Matching or identifying semantically similar data from different web pages is a very important research problem with many practical applications.

- **Opinion extraction from online sources**: Forums, Blogs and Customer reviews on products are the main sources of online opinions. Opinion mining is the study of opinions, sentiments and emotions expressed in the form of text. Mining and analyzing opinions are of great importance for business decision makers in marketing intelligence.

- **Segmenting Web pages and detecting noise**: In many applications, sometimes a user is interested in only the main content of the webpage. A
webpage in addition to main content, also contains banner advertisements, navigation hyperlinks and copyright notices. Segmenting the webpage automatically in order to extract the main content of the pages is another interesting research problem. All these research problems mentioned above present the major research challenges, and of course have many real-life applications as well.

1.4. Applications of Web Mining

1.4.1. Predicting the future using web data: We have terabytes and terabytes of data available on the Web and fortunately now we have abled technology to process such amount of data. Data scientists in every field such as health and business take this data in its raw form both unstructured and unorganized form, get aggregate information out of it, thus finally get knowledge to solve various real world problems. This application can be best explained with the prediction of cholera outbreak in many countries. When cholera re-emerged in Angola on April 7, 2006 with 240,000 cases, data scientists started investigating the cause and found that the outbreak of cholera was based on droughts and storms in Anglo, droughts which had occurred two years before the storm in Anglo. But they could not infer too much from this information, and then they started to find and identify if such similar events happened in any part of the world to find their correlation. They found that in 1960, Bangladesh also had significant cholera cases, most of them were after droughts. So, they start to analyse what was common in both Anglo and Bangladesh. They found that both countries had a low Gross Domestic Product and low concentration of water and in such countries if we have a drought and years later a storm, then the probability of cholera is much higher. Based on the above statistics, a prediction system was developed that would give the probability of cholera. During April 2011, the system flashes the outbreak of cholera in Cuba. As there was no
outbreak of cholera earlier in Cuba for almost ten decades, they thought there is a bug. But in January 2013, Cuba fights to contain the most severe cholera. The predictive model built cannot prevent droughts and storms, but can prevent people from dying because of cholera by taking preventive measures in time. No doubt not only web data can predict the future, but it can also improve it.

1.4.2. User activity identification and tracking in e-commerce: Let us take a scenario of customer behaviour, both in a general store and in an online store. In a general store, data about the customer can be collected only at one point known as checkout counter or ‘point-of-sale’. No other information about the customer behaviour can be predicted throughout his shopping process. But in case of an online store, the complete behaviour of customer can be recorded with the help of web log files (web logs are maintained by web servers and contain information about the user accessing the site). User activity identification and tracking can tell us who are our best customers, what makes our customer leave, what makes our customer stay, what mix of items they buy and what action should be taken to prevent them from leaving. Hence, nearly every Organization in this competitive market will eventually need such type of information in order to remain ahead of the curve.

1.4.3. Google Search: Google is World’s one of the most famous and popular search engines so far. On an average Google processes over 40,000 search queries every second. As it spends less time on searching and produces a quantitative and quick search result, that is why, it is different from other search engines. It is the first search engine to use web structure mining in order to mine the information from the Web. In order to rank and calculate the importance of web pages, Google uses the Link structure of the Web. [Page and Brin, 1998], proposed the PageRank Algorithm and according to
them PageRank is defined as: “Assume page A has \( T_1 \ldots T_n \) pages which point to A. The PageRank of a page A, \( PR(A) \) can be calculated as follows:

\[
PR(A) = (1 - d) + d \left( \frac{PR(T_1)}{C(T_1)} + \ldots + \frac{PR(T_n)}{C(T_n)} \right)
\]  

(1.3)

Where \( d \) is a damping factor, which can be set between 0 and 1 and we usually set \( d \) to 0.85 and \( C(A) \) is the number of links going out of page A”.

1.4.4. Detection of fake reviews and reviewers: Everybody goes online before getting inline to buy anything or go anywhere in order to satisfy themselves with someone else’s feedback about the product, hotel or restaurant. If a customer wants to buy a product, he will read the reviews of that product and if most of the reviews are positive, then the chances of buying are very high otherwise certainly the customer shall not buy that product. People who want to damage the reputation write fake reviews of products, hotels and restaurants, which is an illegal activity and is termed as opinion spamming. Detecting fake reviews and reviewers is very important for every organization and with the help of Web mining techniques we can solve such problems.

1.4.5. Fraud detection: With the advancement in technology, fraud is also rising dramatically, resulting in loss of billions of dollars each year worldwide. Detecting and preventing fraud has become very important to the growing popularity of e-commerce and telecommunication systems. For each user, e-commerce sites maintain a signature based on one’s buying patterns (e.g. types of items bought). If this buying pattern changes significantly, that means the signal is a fraud. HNC software has developed software known as Falcon, which uses domain knowledge and neural networks for credit card fraud detection [Bolton and Hand, 2002].
1.5. Identification of the Research Problem

The most common and widely used way of disseminating and retrieving information now-a-days is a WWW without any doubt. The previous decade experienced a remarkable growth in computer technology, such that with the press of a fingertip the information about a particular topic appeared in monitors within seconds. As time passed by the complexity of the Web increased due to the enormously large amount of data. Therefore, it is very difficult to guess and estimate the exact volume of data available on the Internet, as the Web is growing exponentially with each passing day. It has been estimated that the data on the Web is doubling after every 1.2 years resulting in information overload. [Toffler, 1970], a Futurologist was the first to coin the term “Information Overload”. Indeed, the Internet is considered as an ocean of information, but the growth of Internet exceeded all expectations. So extraction of data according to users need became a tedious and cumbersome task and as a result, mining became an essential technique to extract the valuable information from the Web. Data available on the Web is structured, unstructured and semi-structured according to the database researchers. Generally, data contained in relational databases are referred to as structured data, XML data are referred to as semi-structured data and text documents, presentations and spreadsheets are referred to as unstructured data. To be more precise, structured data is intended for machines while as unstructured data is intended for people, and semi-structured data, somewhere in between. Extraction of data is extremely valuable for extracting specific information from a large number of web pages. On the Internet, same type of information is usually scattered across different websites. If we can collect the information together, save in structured form, it would be beneficial. Extracting data from the web using already existing tools such as search engines have some limitations. Search engines help us to browse and visit a huge amount of websites in order to retrieve the relevant
information from the Web but manually visiting these sites and extracting data is time-consuming. Also data on the Web is diverse and scattered so it is necessary to integrate and extract data from web pages. Therefore, researchers have started extracting and integrating the information from the Web and thus one can easily access and retrieve the relevant data from the Web. Consider a webpage as shown in figure 1.5. Such a webpage, not only contain just unstructured data, but also contains a large amount of structured data. As can be seen from the figure, the middle part of the webpage contains data with similar tuples and the rest of the webpage contains advertisements, logos, etc. If we want to extract structured data using search engines, it is not possible because search engines treat a webpage as a single unit. i.e. either it will retrieve a full webpage or will not retrieve.

![Image](image.png)

**Figure 1.5:** A Structured Webpage containing four Data Records

The extraction of structured data from webpages got the attention in the mid-1990’s and researchers started to work on this extraction problem. Manual, Semi-automatic and Fully-automatic are the three main approaches of web data extraction [Liu,2012]. Structured web data extraction is the problem of extracting data records from the webpages. Structured data
extraction can be thought of as a reverse engineering task, i.e. given a webpage, underlying data model can be retrieved where from data can be extracted. Such data on the Web is usually embedded in template generated webpages [Cafarella et al., 2011]. Upon a web page request, structured data are retrieved from an underlying database and inserted into web pages using some fixed template. According to [Liu, 2012; Cafarella et al., 2011] structured data on the Web are data records that are being retrieved from the underlying database. The web pages may contain structured data in the form of tables, horizontal lists, vertical lists etc. As there are billions of web pages on the Web, that differ in their visual style and the underlying structure. So, automatically extracting structured data from such pages is not a trivial task and much of the efforts are laid by various researchers to tackle the problem [Bohannon et al., 2012; Dalvi et al., 2011; Furche et al., 2012a]. Structured data extraction can basically alter the approach with which we see and browse the Web today. Structured web data extraction approaches are important for organizations. Also, the success of an organization in today’s era hinges on finding and reacting to competitive pressures [Connote, 2012]. For example, if a business company wants to extract data from hundreds of sources, it is sensibly infeasible to try manual approaches, practically a fully-automatic approach should be developed by the company to achieve better and cost effective results. This thesis discusses and examines how to automatically extract structured data from various web pages. Web can be thought as “Database of Everything”, taking such web pages into consideration. Thus, extracting and integrating data from multiple webpages promises a number of compelling applications.

1.6. Objectives of the Research

The main Objectives of our research are listed below:

I. To review state-of-the art data extraction approaches.
II. To propose a method for automatic extraction of structured data (ESD) from a single page.

III. To propose a method for automatic extraction of structured data from multiple e-commerce websites

IV. To Propose a method for analysis of customers buying behavior.

1.7. Organization of the Thesis

Our Thesis is organized into six chapters. Chapter 1 introduces the research problem and its background. The concept of information overload and evolution of Web mining has been elaborated in detail. Pertinent concepts of Web mining such as characteristics of web data, types of web data and taxonomy of Web mining have also been described in an efficacious manner. We have also presented few research problems in the field of web content mining, which is the field under study. A few real life applications of Web mining have also been discussed. The problem of extraction of structured data from web pages has also been elucidated clearly.

In Chapter 2 related areas of Web mining, namely Data Mining, Information Retrieval and Information Extraction have been presented with diagrammatic representation which unfolds their boundaries. The detailed discussion about the extraction problem and classification of retrieval and mining techniques have also been illustrated. The three main approaches used for data extraction viz. Manual approaches, Semi-automatic approaches and Fully-automatic approaches have been presented in a concise manner. The other associated classification of structured data extraction approaches which is based on the techniques used by each approach to generate wrapper has also been explained in detail. Various state-of-art approaches both theoretical and practical, to structured data extraction has been outlined and explained in detail in a historical perspective.
Chapter 3 discusses the two main categories of automatic data extraction approaches, pattern search based methods and visual signal aided methods. A Fully-automatic method known as Extraction of Structured Data (ESD) has been proposed for automatic extraction of structured data from a single list page using segmentation and block similarity calculation. Each stage of ESD has been discussed in detail in this chapter. We have also analysed various string similarity measures. The method has been tested, which shows the efficacy of the proposed method.

Chapter 4 briefly discusses the various extraction techniques of product information from multiple e-commerce websites. This chapter proposes a method for automatic extraction of product information from multiple e-commerce web pages. We have also discussed the usefulness of extraction method so that web shoppers will be convinced and benefited. The said method generates wrapper, which can be re-used to extract data from web pages of the same template. The method uses structural similarity of DOM trees for efficient clustering. At the end of this chapter, the method is tested and compared to other approaches.

Chapter 5 discusses a model-based on modern Data Mining techniques. The model can predict the buying behaviour of a customer within an e-commerce system.

Chapter 6 concludes the thesis highlighting some of the major contributions of this work. The contributions of the problem at hand have been presented point wise. Lastly a brief discussion on future work has been presented.