Chapter 2
WEB MINING

Unprecedented popularity of Internet and the dependency of people to the World Wide Web for information gathering and sharing have made sophisticated Web Mining Techniques a primary concern. But the conventional Data Mining Techniques cannot be applied directly on Web Data due to its complex nature and size. This chapter presents a discussion of the special nature of web data, various components of Web Mining and its relevance. Special emphasis is given to Web Usage Mining and its applications.

2.1 Introduction

With the radical quick and explosive growth of information available over the Internet, World Wide Web has become a powerful platform to store, disseminate and retrieve information as well as mine useful knowledge. To be able to cope up with the abundance of available information, users of the Web need assistance of intelligent Web Mining software for finding, sorting, and filtering the available information [Etzioni, 1996]

Locating relevant information is the major problem with the web related search and applications. Users either browse web documents directly or use a Search Engine as a search assistant to locate required information on the web. Search Engines use keywords to retrieve the requested information and returns a list of ranked pages based on the relevance to the query [Xu, 2008]. However, query-based web search faces with two major problems. The first problem is low precision, which is caused by the large number of irrelevant pages returned by the search engine. Precision is the percentage of retrieved documents that are in fact relevant to the query. The second problem is low recall, which is due to the lack of capability of indexing all web pages available on the Internet. This causes the difficulty in locating the un-indexed information that is actually relevant. Recall tells what percentage of relevant documents on the net is retrieved by the search engine [Kosala and Blockeel, 2000]. Thus, the profusion of resources has prompted the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "Web Mining".

2.2 Data Mining

Data mining, also termed as Knowledge Discovery in Databases (KDD) is defined as the process of discovering useful patterns or knowledge from large 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. Major tasks in data mining are classification (supervised learning), clustering (unsupervised learning), association rule mining, and sequential pattern mining [Liu, 2007].

KDD process can be decomposed into three steps:

1. ***Pre-processing***: In real-world datasets, erroneous values can be recorded for a variety of reasons, including measurement errors, subjective judgments and malfunctioning or misuse of automatic recording equipment. Thus raw data is usually not suitable for mining. It may need to be cleaned in order to remove noises or abnormalities. When the volume of data is very huge, irrelevant attributes can be avoided through feature selection.

2. ***Data mining***: The processed data is then fed to a data mining algorithm which will produce patterns or knowledge.

3. ***Post-processing***: In many applications, all discovered patterns are not useful. Patterns that are useful for application are identified through the Post-processing step. Various evaluation and visualization techniques are applied to make the decision.

The data mining tasks are 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 advent of popularity of the Web and text documents, Web Mining and Text Mining are becoming increasingly important and popular [Han, Kamber, 2000; Liu, 2007].

2.3 Web Mining

The term Web mining was first proposed by Oren Etzioni in his paper [Etzioni, 1996]. In this work, he defines Web mining as the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. In the same paper, Etzioni came up with a question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has become the focus of quite a lot of research studies **[**Wang, 2000].

2.3.1 Why Web Mining?

From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. Web mining can be viewed as the collection of technologies to fulfil this potential. Interest in Web mining has grown rapidly in its short history, both in the research and practitioner communities. Web mining research has attracted many academicians and engineers from database management, information retrieval, artificial intelligence research, especially from data mining and knowledge discovery.

A panel organized at 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Nov. 1997 USA (ICTAI 1997) raised the question that whether there is any distinction between Web mining and general data mining. While no definitive conclusions were reached then, the tremendous attention by the researchers on various aspects of Web mining and the number of significant ideas that have been developed, have answered this question in the affirmative in a big way [Srivastava, Desikan and Kumar, 2004].

Web mining can be broadly defined as the discovery and analysis of useful information from the WWW. In web mining, data can be collected at the server side, client side, proxy servers, or obtained from an organization's database [Mobasher, Cooley and Srivastava, 2000]. Depending on the location of the source, the type of collected data differs. It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and meta information that might be available. This makes the techniques to be used for a particular task in web mining widely varying.

Some of the characteristics of web data are

• *Web page complexity far exceeds the complexity of any traditional text document collection.*

Although the Web functions as a huge digital library, the pages themselves lack a uniform structure and contain far more authoring style and content variations than any set of books or traditional text-based documents. Moreover, huge number of documents in this digital library have not been indexed, which makes searching the data it contains extremely difficult.

• *The Web constitutes a highly dynamic information source.*

Not only does the Web continue to grow rapidly, the information it holds also receives constant updates. News, stock market, service centre, and corporate sites revise their Web pages regularly. Linkage information and access records also undergo frequent updates.

• *The Web serves a broad spectrum of user communities.*

The Internet's rapidly expanding user community connects millions of workstations. These users have markedly different backgrounds, interests, and usage purposes. Many lack good knowledge of the information network's structure and unaware of a particular search's heavy cost. They frequently get lost within the Web's ocean of information, and get disturbed at the many access hops and lengthy waits required to retrieve search results.

• *Only a small portion of the Web Pages contain truly relevant or useful information.*

A given user generally focuses on only a tiny portion of the Web, dismissing the rest as uninteresting data that serves only to swamp the desired search results [Sharma, 2011].

Web mining, though considered to be a particular application of data mining, warrants a separate field of research, mainly because of the aforesaid characteristics of the data and human related issues [Han, Kevin and Chang, 2002].

2.3.2 Components of Web Mining

According to Etzioni, Web mining can be viewed as consisting of four tasks as shown in Figure 2.1.

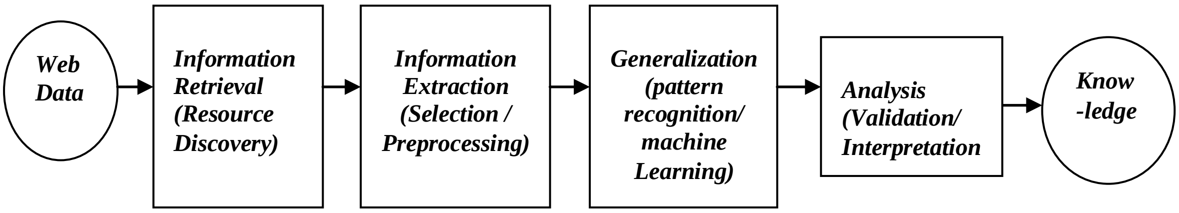
Web Data
Information Retrieval (Resource Discovery)
Information Extraction (Selection / Preprocessing)
Generalization (pattern recognition/ machine Learning)
Analysis (Validation/ Interpretation
Know -ledge

Figure 2.1: Web Mining Subtasks [Pal, Talwar and Mitra, 2002]

1. *Information Retrieval (IR) (Resource Discovery)*: Resource discovery or IR deals with automatic retrieval of all relevant documents, while at the same time ensuring that the non relevant ones are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.

2. *Information Selection/Extraction and Preprocessing*: Once the documents have been retrieved, the challenge is to automatically extract knowledge and other required information without human interaction. Information extraction (IE) is the task of identifying specific fragments of a single document that constitute its core semantic content.

3. *Generalization*: In this phase, pattern recognition and machine learning techniques are usually used on the extracted information. Most of the machine learning systems, deployed on the web, learn more about the user's interest than the web itself. A major obstacle when learning about the web is the labeling problem: data is abundant on the web, but it is unlabelled. Many data mining techniques require inputs labelled as positive (yes) or negative (no) examples with respect to some concept.

4. *Analysis*: Analysis is a data-driven problem which presumes that there is sufficient data available so that potentially useful information can be extracted and analyzed. Humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and or interpretation of the mined patterns which take place in this phase. Once the patterns have been discovered, analysts need appropriate tools to understand, visualize, and interpret these patterns.

Based on the aforesaid four phases [Figure 2.1], 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. Here, evaluation includes both generalization and analysis [Pal, Talwar and Mitra, 2002].

2.3.3 Category of Web Mining

Generally two different approaches were taken in defining Web mining. First was a ‘process-centric view’ that defined Web Mining as a sequence of tasks [Section 2.3.2] and the second, a ’data-centric view’ that defined Web Mining in terms of the type of Web data that is used in the mining process [Srivastava, Desikan and Kumar, 2004].

The data-centric view of Web mining is defined as: - Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Usage data. Thus, based on the primary kinds of data used in the mining process, web mining tasks can be categorized into three: Web Structure Mining, Web Content Mining and Web Usage Mining.

Web Content Mining**:** Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of Text Mining to Web content has been the most widely researched area. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research on this topic has drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, application of these techniques in Web Content Mining has been limited [Srivastava, Desikan and Kumar, 2004].

Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources. Multimedia data mining on the Web has gained many researchers’ attention recently. Working towards a unifying framework for representation, problem solving, and learning from multimedia is really a challenge, this research area is still in its infancy indeed, many works are waiting to be done [Wang, 2000, Petrushin, 2007].

Web Structure Mining**:** The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining categorizes the web pages and generates information, such as similarity and relationship between different Web sites.

Web structure mining can also have another direction – discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure of Web pages, which will in turn help in easy navigation through the pages, and also facilitates comparison and integration of web page schemes. This also facilitates the introduction of database techniques for accessing information in web pages by providing a reference schema.

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web Structure Mining is the process of discovering structure information from the Web. This can be further divided into two kinds based on the kind of structure information used.

- *Hyperlinks***:** A hyperlink is a structural unit that connects a location in a web page to different location, either within the same web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink. Desikan et al. provides a significant study on hyperlink analysis in [Desikan et al., 2002].
- *Document Structure*: The contents within a web page can be organized in a tree-structured format, based on the various HTML and XML tags within the page. Here, the mining efforts focus on automatically extracting document object model (DOM) structures out of documents [Srivastava, Desikan and Kumar, 2004].

Web Usage Mining**:** Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [Srivastava *et al,* 2000]. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

2.4. Web Usage Mining

Generally, web users are performing their interest-driven visits by clicking one or more functional web objects. They may exhibit different types of access interests associated with their navigational tasks during their surfing periods. Thus, employing data mining techniques on the observed usage data may lead to finding the underlying usage patterns. In addition, capturing of web user access pattern not only supports better understanding of user navigational behaviour, but also helps in efficiently improving Web site structure or design. This, furthermore, can be utilized to recommend or predict Web contents tailored and personalized to web users who can benefit from obtaining more preferred information at reduced waiting time [Madria et al., 1999].

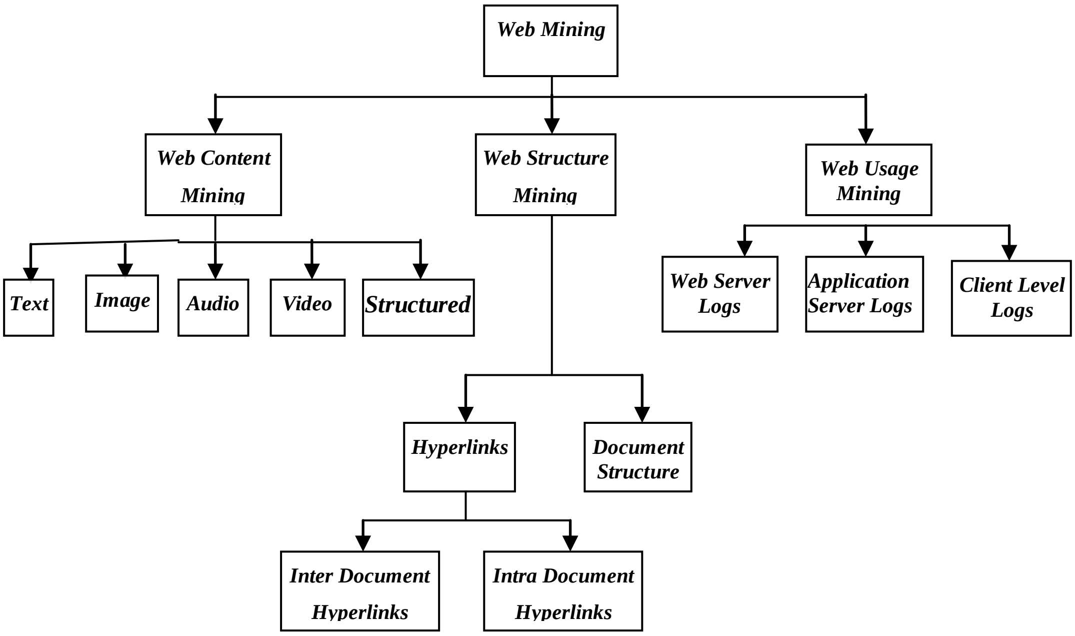

Figure 2.2: Web Mining Taxonomy [Srivastava, Desikan and Kumar, 2004]

2.4.1 Web Usage Data

Data can be collected at the server-level, client-level, proxy-level, or obtained from an organization’s database. Each type of data collection differs not only in terms of the location of the data source, but also in the kinds of data available, the segment of population from which the data was collected, and its method of implementation.

The usage data collected at the different sources such as Server level, Client Level and Proxy Level represent the navigation patterns of different segments of the overall web traffic [Cooley, 2000].

Server-Level Collection: Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. A web server log records the browsing behaviour of site visitors. The data recorded in server logs reflect the concurrent and interleaved access of a web site by multiple users. These log files can be stored in various formats such as Common Log Format (CLF) or Extended Common Log Format (ECLF). Most of the web servers follow common log format (CLF) as “ip address username password date/timestamp URL version status-code bytes-sent“. ECLF contains referrer and agent information in addition. Referrer is the referring link URL and user agent is the string describing the type and version of browser software used. Web cache and the IP address misinterpretation are the two drawbacks in the server log. Web cache keeps track of web pages that requests and saves a copy of these pages for a certain period. If there is a request for the same page, the cached page is used instead of making new request to the server. Therefore, these requests are not recorded into the log files.

Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. In order to handle this problem, web servers also store other kind of usage information such as cookies in separate logs, or appended to the CLF or ECLF logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Packet sniffing technology (also referred to as “network monitors”) is an alternative method for collecting usage data through server logs. Packet sniffers monitor network traffic coming to a web server and extract usage data directly from TCP/IP packets. Besides usage data, the server side log also provides access to the “site files”, e.g. content data, structure information, local databases, and web page meta-information such as the size of a file and its last modified time [Srivastava et al., 2000].

Client Level Collection: Client-side collection can be implemented by using a remote agent (such as Java scripts or Java applets) or by modifying the source code of an existing browser (such as Mozilla) to enhance its data collection capabilities [Srivastava et al., 2000]. Client level data collection need cooperation of users, but has the advantage of eliminating the caching and session identification problem.

Proxy Level Collection: A web proxy acts as an intermediary between client browsers and Web servers. Proxy-level caching can be used to reduce the loading time of a web page experienced by users as well as the network traffic load at the server and client sides. Proxy level data reveals the details of HTTP requests from multiple users to multiple servers. This information is used for identifying the browsing behaviour of a group of users [Srivastava et al., 2000].

2.4.2 Data Preprocessing

The information available in the web is heterogeneous and unstructured. Therefore, data preprocessing has a fundamental and crucial role in discovering patterns. The goal of preprocessing is to transform the raw click stream data into a set of user profiles. Data preprocessing presents a number of unique challenges which makes it a complex and time demanding phase. A variety of algorithms and heuristic techniques have been introduced for preprocessing tasks such as merging and cleaning, user and session identification etc [Mobasher, Cooley and Srivastava, 2000; Eirinaki, Vazirgiannis, 2003].

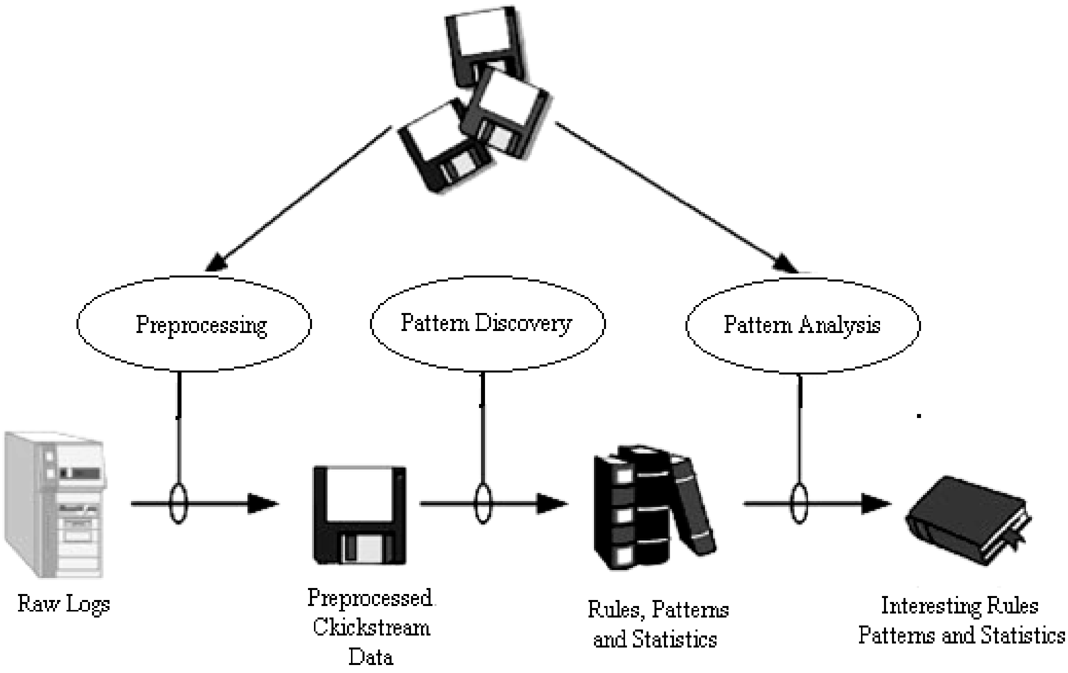
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Ckickstream Data
Rules, Patterns and Statistics
Interesting Rules Patterns and Statistics

Figure 2.3: Web Usage Mining Tasks [Mobasher, Cooley and Srivastava, 2000]

2.4.2.1 Data Cleaning

Data cleaning consists of removing all the data tracked in Web logs that are useless for mining purposes [Tan and Kumar, 2002; Anderson, 2002] e.g. requests for graphical page content (e.g., jpg and gif images), requests for any other file which might be included into a web page, navigation sessions performed by robots and Web Spiders. While requests for graphical contents and files are easy to eliminate, robots and Web Spider’s navigation patterns must be explicitly identified. This is usually done, for instance, by referring to the remote hostname, by referring to the user agent, or by checking the access to the robots.txt file. However, some robots actually send a false user agent in HTTP request. In these cases, a heuristic based on navigational behaviour is used to separate robot sessions from actual users-sessions [Berendt et al., 2002]. The search engine navigational paths are characterized by breadth first navigation in the tree representing the web site structure and by unassigned referrer. The heuristic is based on this assumption and a classification of navigations. Well known robots-navigational paths are used to train the classifier, and the model obtained is used to classify further navigational sessions even if there is no prior knowledge about the user agent that generated them.

2.4.2.2 User Identification

Identification of unique users is necessary for analyzing access behavior of users. In majority of web servers, users are treated as anonymous. The simplest method to identify user is to assign different user id to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. Referrer information and a user agent in the Extended Log Format may be used to overcome this problem. If the IP address of a user is same as previous entry and user agent is different, then the user is assumed as a new user. If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the user, then the user is identified as a new user in the same address. Caching problem can be rectified by assigning a short expiration time to HTML pages enforcing the browser to retrieve every page from the server [Eirinaki, Vazirgiannis, 2003].

2.4.2.3 Session Identification

The goal of session identification is to divide web logs of each user into individual access sessions. A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period. Once a user was identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction [Scime, 2005].

2.4.2.4 Other Preprocessing Tasks

For some usage mining purposes, further preprocessing tasks like Path Completion and Formatting are also applied. Path completion is used to find the actual access path among web pages by adding the important accesses that are not recorded in the web log. The referrer field in the web logs can be checked to find out from which page the request has come. If the referrer is unavailable, the link structure of the website can also help to estimate the access path of users. The goal of transaction identification is to create meaningful clusters of requested web pages for each user. Therefore, the task of identifying transactions is to divide a larger transaction into multiple smaller ones or merge smaller transactions into larger ones. Once the appropriate preprocessing steps have been applied to the server log, a final preparation module is used to properly format the sessions or transactions for the type of data mining to be accomplished [Cooley, Srivastava and Mobasher, 1999].

2.4.3 Pattern Discovery Techniques in Web Usage Mining

The pattern discovery phase consists of different techniques derived from various fields such as statistics, machine learning, data mining, pattern recognition, etc. applied to the Web domain and to the available data [Srivastava et al., 2000].

The most commonly used techniques applied to Web Usage Data are:-

2.4.3.1 Statistical Analysis

Statistical analysis is performed by many tools to give a description of the traffic on a web site, like most visited pages, average daily hits, etc. Statistical analysis is performed on variables such as page views, viewing time and length of access path. This type of knowledge is potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task and providing support for marketing decisions.

2.4.3.2 Classification

Classification is the process of building a model to classify a class of objects so as to predict the class label of a future object whose class is not known. Since the class label of each training sample is provided, this process is also known as supervised learning. For Web Usage Mining, classification is usually used to construct profiles of users belonging to a particular class or category. Classification can be done by the use of learning algorithm such as Decision Tree classifier, Naïve Basic algorithm etc. [Srivastava et al., 2000].

2.4.3.3 Clustering

Clustering is a technique for grouping a set of physical or abstract objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A cluster of data objects can be treated collectively as one group in practical applications. There exist a large number of clustering algorithms. The choice of a clustering algorithm depends both on the type of data available, and on its purpose and application [Berkhin, 2002].

2.4.3.4 Association Rule Mining

Association rule mining finds interesting association or correlation relationships among a large set of data items. A typical example of association rule mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”. For web usage mining, association rules can be used to find correlations between web pages (or products in an e-commerce website) accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Apart from being exploited for business applications, the associations can also be used for web recommendation [Lin, Alvarez and Ruiz, 2000], personalization [Forsat, Meybodii and Neiat, 2009; Shinde, Kulkarni, 2011; Ansari et al., 2000] or improving the system’s performance through predicting and pre-fetching of web data [Chen and Zhang, 2003; Domènechv et al., 2007].

2.4.3.5 Sequential Pattern Mining

Web logs can be treated as a collection of sequences of access events from one user or session in timestamp ascending order. A web access pattern [Agrawal and Srikant, 1995; Pei et al., 2000] is a sequential pattern in a large set of pieces of web logs, which is pursued frequently by users. Such knowledge can be used for discovering useful user access trends and predicting future visit patterns, which is helpful for pre-fetching documents, recommending web pages, or placing advertisements aimed at certain user groups. Sequential pattern mining techniques [Agrawal and Srikant, 1995] are commonly used for discovering web access patterns from web logs.

2.5 Applications of Web Usage Mining

- Web Usage Mining offers the ability to analyze massive volumes of click stream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for web personalization, e-CRM and other interactive marketing programs.
- Web Usage Mining is used to determine common behaviors or traits of users.
- By determining frequent access behavior of users, needed links can be identified to improve the overall performance of future accesses. Information concerning frequently accessed pages can be used for caching.
- In addition to modifications to the linkage structure, identifying common access behaviors are used to improve the actual design of Web pages and to make other modifications to the site.
- Web usage patterns are used to gather business intelligence to improve Customer attraction, Customer retention, sales, marketing and advertisement, cross sales.
- Mining of web usage patterns help in the study of how browsers are used and the user’s interaction with a browser interface.
- Usage characterization is used to look into navigational strategy when browsing a particular site.
- Web usage mining of patterns provides a key to understanding Web traffic behaviour, which can be used to deal with policies on web caching, network transmission, load balancing, or data distribution.
- Web usage and data mining is also useful for detecting intrusion and attempted break-ins to the system.
- Web usage mining supports Counter Terrorism and Fraud Detection by detecting unusual accesses to secure data.

- Web usage mining is used in e-Learning, e-Business, e-Commerce, e-Services, e-Education, e-Newspapers, e-Government and Digital Libraries.

2.6 Summary

This chapter introduced basic concepts on Web Mining and its difference from conventional Data Mining. Special focus is given to Web Usage Mining and discussed in detail. Finally, applications of Web usage mining are highlighted. Searching, comprehending and using the semi structured information stored on the Web poses a significant challenge because this data is more sophisticated and dynamic than the information stored in commercial database systems.