CHAPTER 1

INTRODUCTION

1.1. OVERVIEW OF THE RESEARCH

The success of digital revolution and the growth of the Internet have ensured that huge volumes of high-dimensional data are available to all users. The World Wide Web (WWW) has played an important role in making the data, even from geographically distant locations, easily accessible to users all over the world. The World Wide Web (WWW), in the current era of information explosion, has become the largest source of online data, which includes information in multiple mediums including text, graphics, videos and sound. WWW is considered as a global information medium where users read, write and communicate via computers connected to the Internet.

The impact of the Internet on everyday life is tremendous and it has changed the way of transacting business, providing and receiving education, organization management, etc. The manner of information collection and sharing has changed with the advancement of hardware and communication software. It has become the defacto technology for sharing new ideas and content exchange (Schall *et al.*, 2011).

It is a well-known fact that the Internet is a dynamic environment, whose evaluations are on par with the modern technologies like mobile and social networks. These changes have made possible a much greater reach of the Internet, increasing the number of websites and Internet users worldwide. The past few decades have envisaged an exponential growth related to both the number of websites available and number of users using these websites for various operations like information collection, entertainment and e-commerce (Yan *et al.*, 2014).

The information stored on the net is enormous, every second, millions of bytes are added all over the world. This indicates that the immense data and knowledge on the net is growing at a mind blowing rate and even estimates are proven wrong every second. According to the survey conducted by http://www.internetlivestats.com, the number of websites has recently exceeded 14.3 trillion during September 2014 and is being accessed by more than 2,756,198,420 users. According to http://www.factshunt.com, an average 103 million websites are added per annum in WWW and the total amount of data stored in these websites is over 1 Yotta-byte (1 000 000 000 000 000 000 000 000 bytes), out of which the accessible data is more than 672 Exabytes (672,000,000,000 Gigabytes).

One fraction of this growth is mainly contributed by e-commerce market, which is the only trillion-dollar industry growing at a double-digit percentage each year (http://www.internetretailer.com). According to the survey conducted by Moore (2014), there are more than 110,000 e-commerce websites in English using Alexa Global ranking method. This dynamic e-market is increasingly leaving antiquated marketing philosophies and strategies to the adoption of more customer-driven initiatives that seek to understand, attract, retain and build intimate long term relationship with profitable customers (Asiedu and Safo, 2013; Kotler and Kelvin, 2006). This paradigm shift has undauntedly led to the growing interest in Customer Relationship Management (CRM) initiatives that aim at ensuring customer identification and interactions, customization and personalization that unreservedly lead to customer satisfaction, retention and profitability, among other additional business benefits (Thompson, 2004; Ryals and Knox, 2001).

Thus, the 21st century e-commerce sites are increasingly becoming more customer-centric and are very much interested not just in acquiring new customers, but, more importantly, but also in retaining existing customers. On an

average, a business spends six times more to obtain new customers, using e-commerce websites than old customers (Lakshmi *et al.*, 2011). Thus, it is imperative to attract old customers to re-use the company services and to run a more profitable e-commerce company. For this purpose, these websites are currently using the details regarding the user's interactions and behaviour to identify their patterns and preferences, in order to improve their experience during a transaction and to serve them better.

These details, usually stored in web-logs, consist of "hidden asset" called "knowledge" that has become the most valuable resource to both web designers and owners. The knowledge discovery techniques used for these purposes, termed as Web Usage Mining (WUM), is a research field that focuses on analyzing and predicting user's experience / preferences using data mining techniques.

The result of such analysis can be used in several applications like personalizing user's web browsing experience (Pagar *et al.*, 2014) and predicting intuitive web pages that a user is likely to browse (Anitha, 2010). Such an analysis is widely popular in e-commerce and e-business websites also (Mohanty and Passi, 2010; Ya, 2012) In the field of e-commerce, the knowledge of users' intentions is used to provide only related and targeted marketing advertisements thus resulting in increased number of customers. Further, these details can also be used to improve the browsing experience of the user by predicting the next web page access of a user. These predictions aim to save browsing and searching time, while at the same time reducing the retrieval time and bandwidth load on network considerably.

Such systems are termed as Web Access Recommendation System (WARS) and are generally used to locate relevant web pages or items in which the user is interested and are used to locate relevant pages, products, items or services in a website. According to Mulvenna *et al.* (2000), it is more efficient and user-

friendly to provide users with what they need automatically and without asking them explicitly for it. Several systems have been developed and implemented for this purpose (Niranjan *et al.*, 2010; Badhe and Shirsat, 2013; Suguna and Sharmila, 2013). But owing to the dynamic nature of the WWW and e-commerce industry, the area is still considered immature and requires more research to identify techniques for improving the efficiency of WARS in terms of accuracy and speed. This research work is another attempt towards improving the performance of WARS, using enhanced data mining techniques.

1.2. OVERVIEW OF WEB MINING

Web Mining is the process of discovering potentially useful and previously unknown information from the Web data (Kosala and Blockeel, 2000) using machine learning (data mining) techniques. The techniques in web mining focus on providing solutions to content provider, web designer and programmers to improve their website and also to the web users with navigation assistance tools. It is a part of data mining where knowledge is gained from WWW.

Web mining is defined as an application of data mining to extract knowledge from web data, where atleast one of structure (hyperlink) or usage (Web log) data is used in the mining process. Initially, it was used to observe the user behaviour from their viewing, book marking and browsing history. The term web mining was coined by Etzioni (1996) and was used to define task oriented method and later was defined to be data oriented method by Cooley *et al.* (1997). A general web mining scenario is given in Figure 1.1.

In general, web mining is the activity of identifying patterns ***p*** implied in large document collection C, which can be denoted by a mapping $\xi : C \rightarrow \boldsymbol{p}$. The general process of web mining consists of four different stages as shown in Figure 1.1.

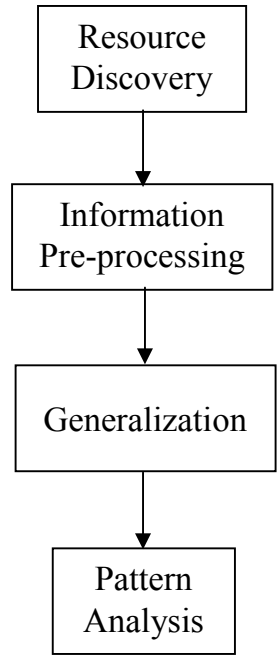

Figure 1.1 : General Framework of Web Mining

Resource Discovery is the task of retrieving information from web resources and documents. Web information retrieval is the process to find a subset S of appropriate number of documents relevant to a certain query q from large document collection C, which can also be denoted by a mapping $\xi : (c,q) \rightarrow S$. Information Pre-processing is the transform process of the result of resource discovery. Generalization is used to uncover general patterns at individual and across multiple sites. In this step, machine learning and traditional data mining techniques are typically used. Pattern Analysis is the validation of the mined patterns.

1.2.1. Categorization of Web Mining Approaches

Web mining approaches are broadly divided into two categories, namely, "Process-centric approach" and "Data-centric approach". Process centric approach views web mining as a sequence of task, while data centric approaches view web mining in terms of web data used during the process. Data centric approaches and solutions are more popular, and the present research work also adapts this

approach. According to the Data centric approach, web mining is defined as an application of data mining techniques to extract knowledge from web data, where atleast one of structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data). Web mining can be broadly divided into three distinct categories, namely, web content mining, web structure mining and web usage mining, according to the kinds of data to be mined (Figure 1.2).

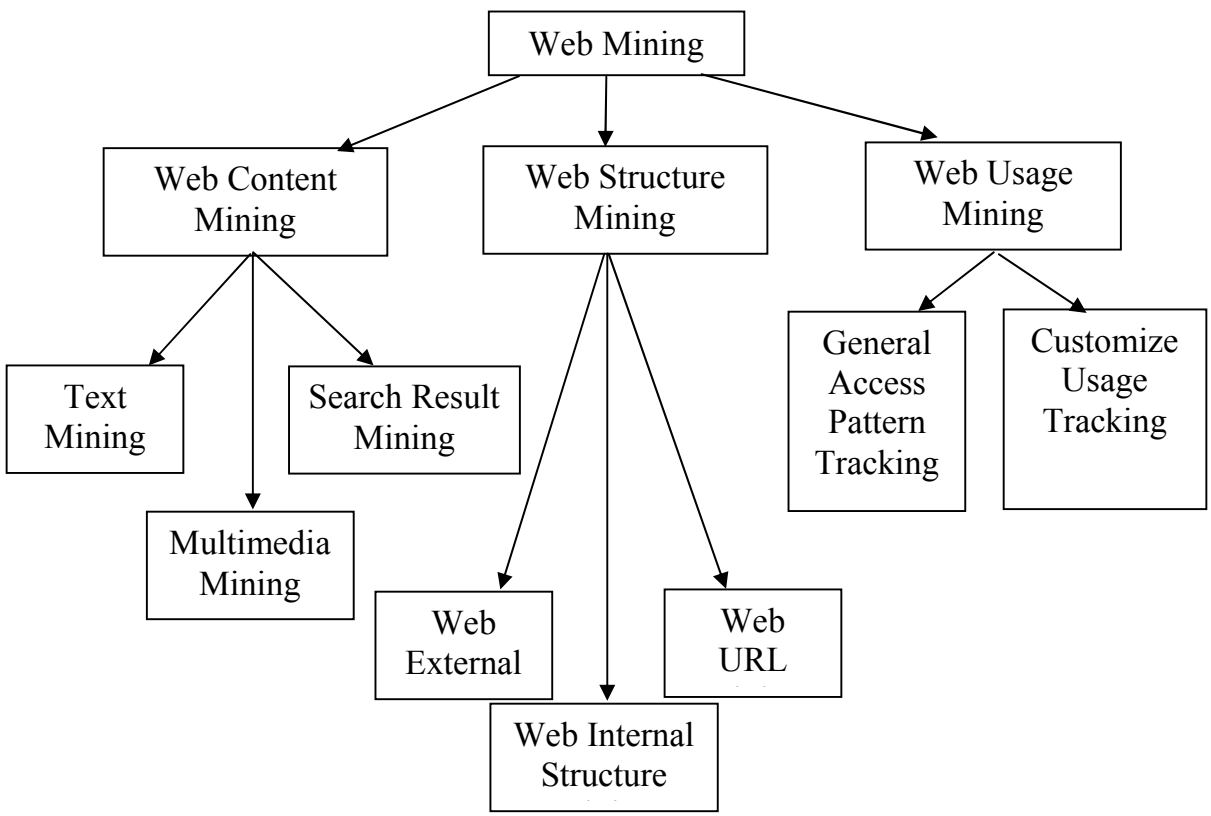

Figure 1.2 : Categorization of Web Mining

Web content mining is the application of data mining techniques to contents published on the Internet, usually as HTML (semi-structured), plaintext (unstructured) or XML (structured) documents. Web structure mining helps the users to retrieve the relevant documents by analyzing the link structure of the Web (Kumar and Singh, 2010).

Web structure mining is the art of discovering the link structures of the Web pages, so that tasks like cataloging and generating information such as the similarity and relationship between them can be performed by taking advantage of their hyperlink topology. Web usage mining on the other hand, involves the analysis and discovery of user access patterns from Web servers in order to serve the users' needs better.

Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behaviour and the web structure, thereby improving the design of this colossal collection of resources (Srivastava *et al.*, 2000). There are two main trends in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools exist, but they are limited and usually unsatisfactory, and this research work aims to provide a tool to analyze the usage navigation pattern. Customized usage tracking analyzes individual trends. Its purpose is to customize web sites to users. The information displays the depth of the site structure, and the format of the resources can be dynamically customized for each user, over time based on their access patterns.

1.3. WEB USAGE MINING

Web usage mining, popularly also known as web log mining, works on secondary data like web log file and click streams to extract knowledge with regard to web usage. It is the process which uses data mining techniques abundantly, the result of which can be used for various purposes like personalization, system improvement and site modification. Web Usage Mining (WUM) prediction process is structured according to two components performed

training and testing with respect to the Web server activity (Frias-Martinez and Karamcheti, 2003; Baraglia and Silvestri, 2007; Jalali *et al.*, 2008). The training component is aimed at building the knowledge base by analyzing historical data, such as server access log files, that is then used in the testing component. It works on details collected from web servers, site contents, visitor information and external channels. The challenge in mining of web data is twofold:

(i) The quality of web data varies considerably, and

(ii) Their integration with data from other sources is difficult.

The solutions and results of web mining are, in most of the cases, custom made for each application, and some conventional presentation tools also exist. There are three generic types of Web applications:

- Revolutionary applications: They have emerged from the Web and have no counterpart in the pre-Web era,

- Innovative applications: They have emerged from Information Technology. The capabilities and particularities of the Web have a major impact on them. Applications like e-learning belong to this category, and

- Web-empowered conventional applications: They were transferred in the Web context; the Web revolutionized the way of doing them. Examples include marketing of products, literature search and imaging and public relations.

Figure 1.3 shows the general architecture of web usage mining systems.

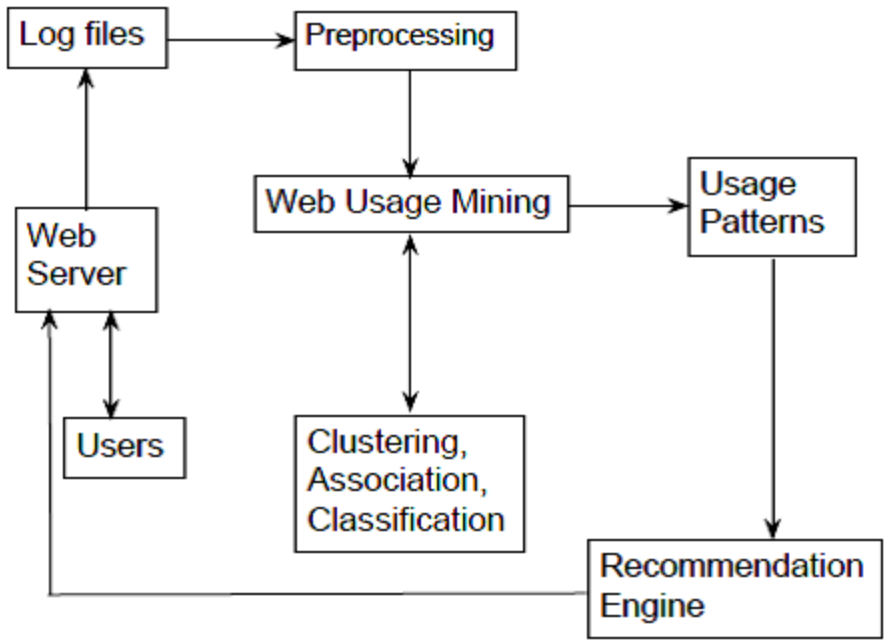

Figure 1.3 : Web Usage Mining Architecture

1.3.1. Application Areas

Web data have been applied to a wide range of applications to discover knowledge. Some of them are listed below (Figure 1.4):-

- Personalization,
- System Improvement,
- Site Modification,
- Business Intelligence, and
- Usage characterization.

Personalizing the Web experience for a user is highly advantageous for many Web-based applications, like individualized marketing for e-commerce. Making dynamic recommendations to a Web user, based on her/his profile in addition to usage behavior is very attractive to many applications, like cross-sales and up-sales in e-commerce. Web usage mining is an excellent approach for achieving this goal. The WebWatcher (Joachims *et al.*, 1997), SiteHelper (Ngu

and Wu, 1997), Letizia (Lieberman, 1995) and clustering work by Mobasher *et al.* (1999) and Yan *et al.* (1996) have all concentrated on providing Web Site personalization based on usage information.

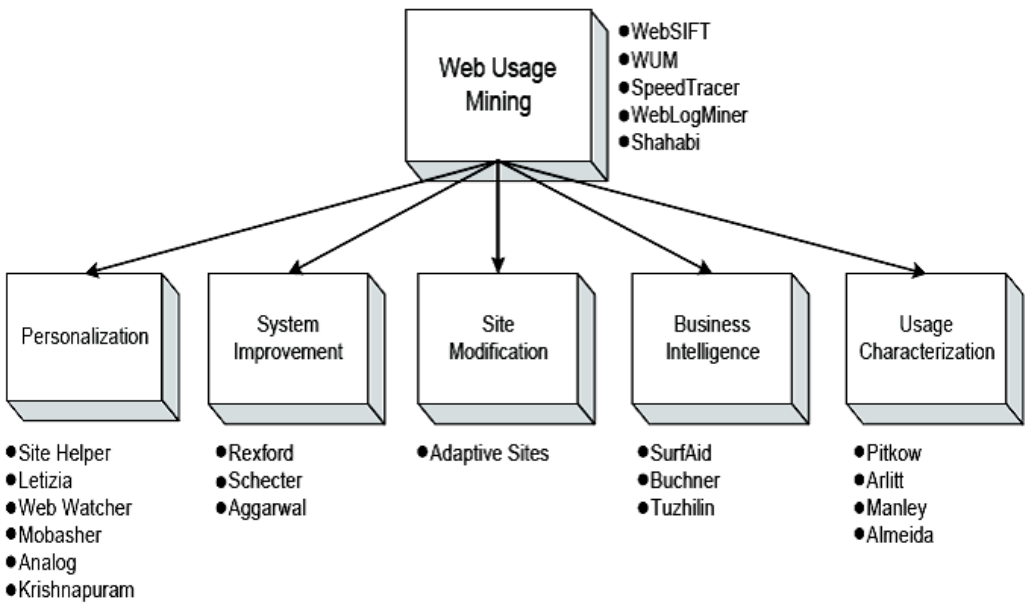

Figure. 1.4 : Applications Areas

Performance and other service quality attributes are crucial to user satisfaction from services such as databases, networks, etc. Similar qualities are expected from the users of Web services. Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission (Cohen *et al.*, 1998), load balancing or data distribution. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate (Fawcett and Provost, 1999). Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins, etc.

The attractiveness of a web site, in terms of both content and structure, is crucial to many applications, like a product catalog for e-commerce. In a similar

fashion, information on how customers are using a web site is critical information for marketers of e-tailing business.

The client side activity log provides detailed information about the user's interaction with the browser interface as well as the navigational strategy used to browse a particular site. This information can be analyzed to extract several usage characteristics of the web site.

1.3.2. Requirements of Web Usage Mining

It is necessary to examine what kind of features a WUM system is expected to possess in order to conduct effective and efficient Web usage mining and what kind of challenges may be faced in the process of developing new Web usage mining techniques. A Web usage mining system should be able to

- Gather useful usage data thoroughly,
- Filter out irrelevant usage data,
- Establish the actual usage data,
- Discover interesting navigation patterns,
- Display the navigation patterns clearly,
- Analyze and interpret the navigation patterns correctly, and
- Apply the mining results effectively.

The result of web usage mining depends on the web usage data and is explained in the next section.

1.4. WEB USAGE DATA

Web usage data captures the identity or origin of Web users along with their browsing behavior at a web site. Three types of usage data are used in web usage mining. They are web server data, application server data and application level data. Web Server Data correspond to the user logs that are collected at Web server. Some of the typical data collected at a Web server include IP addresses,

page references and access time of the users. Application Server Data are produced by commercial application servers (e.g., Weblogic [BEA], BroadVision [BV], StoryServer [VIGN], etc.) and have significant features in the framework to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs. Finally, in application level data, new kinds of events can always be defined in an application and logging can be turned on for them – generating histories of these specially defined events. The usage data can also be split into three different kinds, on the basis of the source of its collection: on the server side, the client side and the proxy side. The key issue is that, on the server side, there is an aggregate picture of the usage of a service by all users, while on the client side there is complete picture of usage of all services, by a particular client, with the proxy side being somewhere in the middle.

1.4.1. Data Sources

The web usage data can be collected either from server side, client side, proxy servers or from an organization's database (Srivastava *et al.*, 2000). All these data collected represent the navigation patterns of different segments of the overall Web traffic, ranging from single-user, single-site browsing behavior to multi-user and multi-site access patterns. The potential data sources are explained in this section and are illustrated in Figure 1.5.

- **Server Level Collection**

A Web server log is an important source for performing Web Usage Mining, because it explicitly records the browsing behavior of site visitors. The data recorded in server logs reflects the (possibly concurrent) access of a web site by multiple users. These log files can be stored in various formats such as Common log or Extended log formats. Besides usage data, the server side also provides content data, structure information and Web page meta-information (such

as the size of a file and its last modified time). The Web server also relies on other utilities such as CGI scripts to handle data sent back from client browsers.

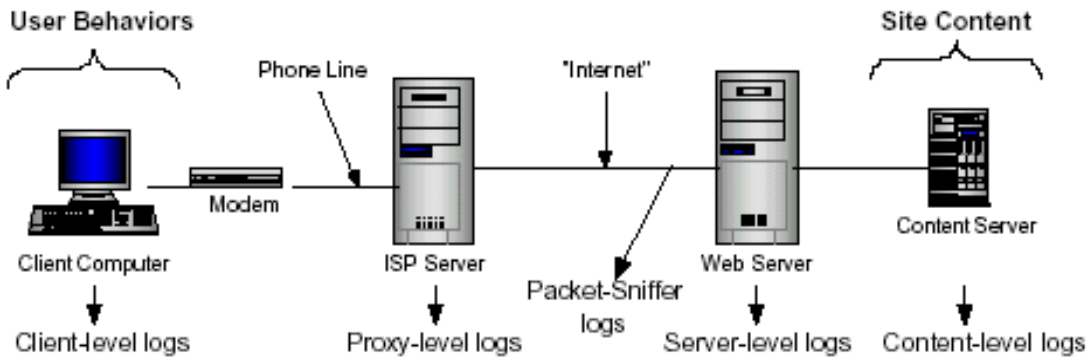

Figure 1.5 : Potential Data Sources

- **Client Level Collection**

Client-side data collection can be implemented by using a remote agent (such as Javascripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities. The implementation of client-side data collection methods requires user co-operation, either in enabling the functionality of the Javascripts and Java applets, or to voluntarily use the modified browser.

Client-side collection has an advantage over server-side collection, because it ameliorates both the caching and session identification problems. However, Java applets perform no better than server logs in terms of determining the actual view time of a page. In fact, it may incur some additional overhead, especially when the Java applet is loaded for the first time. Javascripts, on the other hand, consume little interpretation time but cannot capture all user clicks (such as reload or back buttons). These methods will collect only single-user, single-site browsing behavior.

A modified browser is much more versatile and will allow data collection about a single user over multiple web sites. The most difficult part of using this method is convincing the users to use the browser for their daily browsing activities. This can be done by offering incentives to users who are willing to use the browser, similar to the incentive programs offered by companies such as NetZero and AllAdvantage that reward users for clicking on banner advertisements while surfing the Web.

- **Proxy Level Collection**

A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy 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 (Cohen *et al.*, 1998). The performance of proxy caches depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

1.4.2. Types of Web Data

World Wide Web contains various information sources in different formats and involves three types of data (Figure 1.6).

Web content data is the data, which web pages are designed for presenting to the users. Web content data consists of free text, semi-structured data like HTML pages and more structured data like automatically generated HTML pages, XML files or data in tables related to web content. Textual, image, audio and video data types falls into this category.

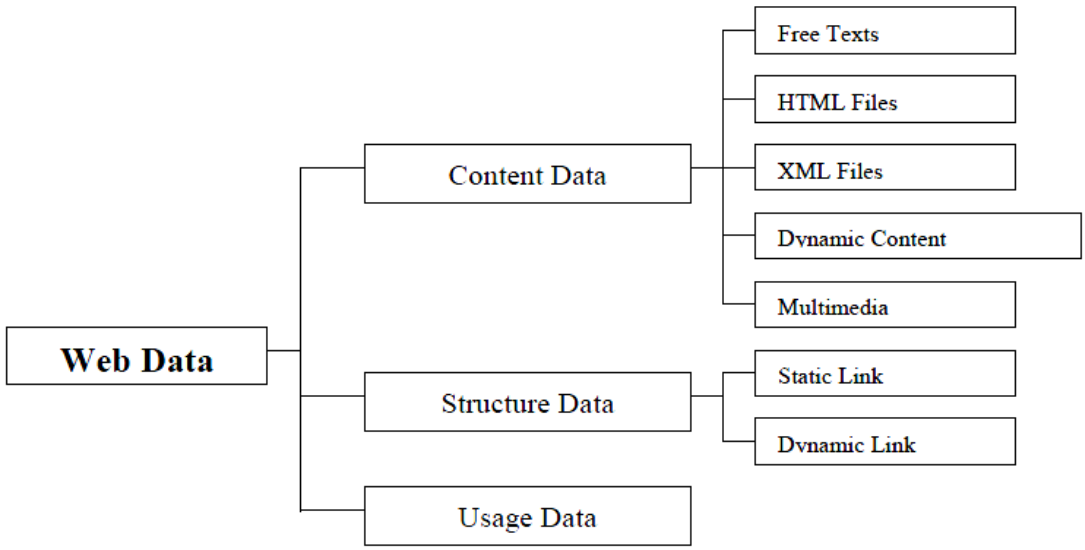

Figure 1.6 : Web Data Categorization

Web structure data describes the organization of the content. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. Inter-page structure information comprises of hyper-links connecting one page to another. Web graph is constructed by hyperlinks information from web pages. The web graph has been widely adopted as the core describing the web structure. It is most widely accepted way of representing web structure related to web page connectivity (dynamic and static links).

Web usage data (focus of this research) includes web log data from web server access logs, proxy server logs, browser logs, registration data, cookies and any other data generated, as the results of web user interactions with web servers. Web log data is created on web server. Every Web server has a unique IP address and a domain name. When any user enters (a URL) any browser, this request is sent to the web server. Subsequent to the operation, web server fetches the page and sends it to the user's browser. Web server data are created from the relationship between web user's interaction with a web site and the web server. A web server log, containing Web server data, is created as a result of the httpd

process that is run on Web servers (Buchner and Mulvenna, 1998). All types of server activities such as success, errors and lack of response are logged into a server log file (Bertot *et al.*, 1997). Web servers dynamically produce and update four types of "usage" log files: access log, agent log, error log and referrer log.

Web Access Logs has fields containing web server data, including the date, time, user's IP address, user action, request method and requested data. Error Logs includes data about specific events such as "file not found", "document contains no data", or configuration errors; providing server administrator information on "problematic and erroneous" links on the server. Other type of data recorded to the error log is aborted transmissions. Agent logs provides data on the browser, browser version and operating system of the requesting user.

Generally, Web server logs are stored in Common Logfile Format (CLF) or Extended Logfile Format (ELF). CLF includes date, time, client IP, remote log name of a user, bytes transferred, server name, requested URL and http status code returned. ELF includes bytes sent and received, server name, IP address, port, request query, requested service name, time elapsed for transaction to complete, version of transfer protocol used, user agent which is the browser program making the request, cookie ID and referrer. Web server logging tools, also known as Web traffic analyzers, analyze the log files of a Web server and produce reports from this information from this data source. These data can be used in the planning and optimizing web site structure.

1.5. GENERAL WEB LOG MINING SYSTEM

A general web log mining system (Figure 1.7) consists of three steps, namely, preprocessing, pattern discovery and pattern analysis.

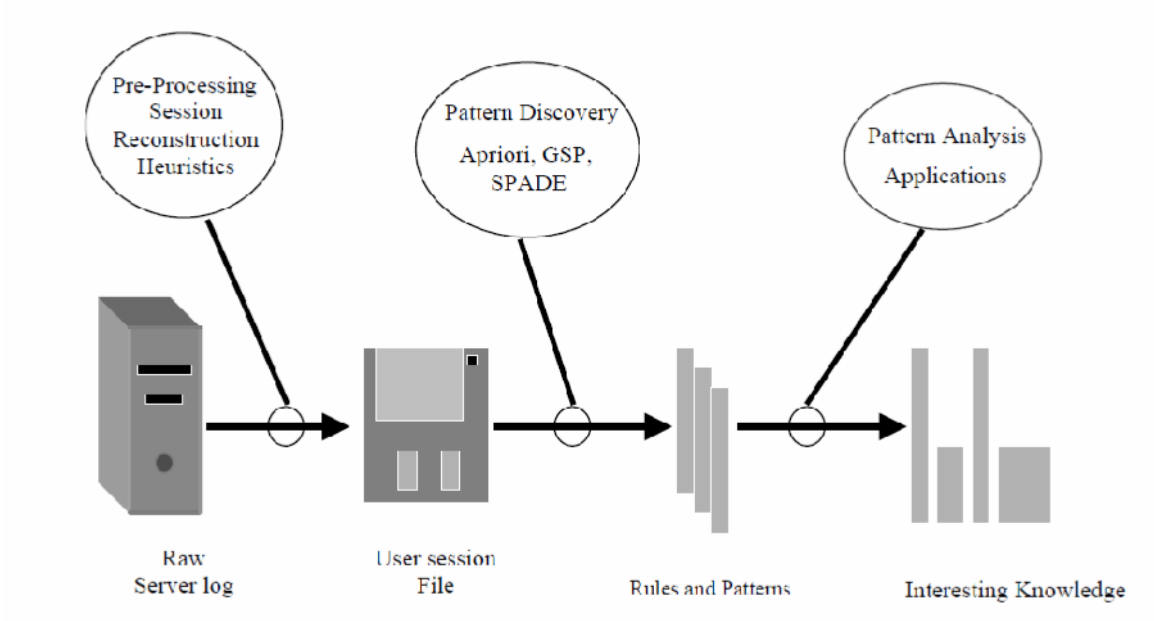

Figure. 1.7 : Phases of Web Usage Mining System

1.5.1. Raw Server Log

Raw server log or Web log is a log file that is created and maintained automatically by a web server. Every "hit" to the web site, including each view of a HTML document, image or other object, is logged. The raw web log file format is essentially one line of text for each hit to the web site. This contains information on who was visiting the site, where they came from and exactly what they were doing on the web site. An example of typical web log format along with a partial sample file is shown in Figures 1.8 and 1.9.

```
<ip_addr><base_url> - <date><method><file><protocol><code><bytes><referrer><user_agent>
```

Figure 1.8 : Format of Web Log File

```
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:21 -0600] "GET /Calls/CWOM.html
HTTP/1.0" 200 3942 "http://www.lycos.com/cgi-
bin/pursuit?query=advertising+psychology&maxhits=20&cat=dir" "Mozilla/4.5 [en] (Win98; I)"
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:23 -0600] "GET
/Calls/Images/earthani.gif HTTP/1.0" 200 10689 "http://www.acr-news.org/Calls/CWOM.html"
"Mozilla/4.5 [en] (Win98; I)"
```

Figure 1.9 : Sample Web Log File

Typically a web log file stores information on

1. IP address of the computer making the request,
2. user ID, (this field is not used in most cases),
3. date and time of the request,
4. a status field indicating if the request was successful,
5. size of the file transferred, and
6. referring URL, that is, the URL of the page which contains the link that generated the request; name and version of the browser being used.

Each user has an entry in the log file with a unique IP address whenever an access is made to a web page of a website or portal. In general, the web log file has the following characteristics.

- The log file is text file. Its records are identical in format,
- Each record in the log file represents a single HTTP request,
- A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL and the browser information,
- A browser may fire multiple HTTP requests to Web server to display a single Web page. This is because a Web page not only needs the main HTML document, it may also need additional files, like images and JavaScript files. The main HTML document and additional files all require HTTP requests, and
- Each Web server has its own log file format.

1.5.2. Preprocessing

Data preprocessing is responsible for converting the usage, content and structure information contained in the web log file into a format that is suitable for pattern discovery. It is the most difficult step in web usage mining. The reason behind this difficulty is the incompleteness of the available data. Typical challenges faced during preprocessing are :

- Single IP address/Multiple Server Sessions – Internet service providers (ISPs) typically have a pool of proxy servers that users access the Web through. A single proxy server may have several users accessing a Web site, potentially over the same time period,

- Multiple IP address/Single Server Session - Some ISPs or privacy tools randomly assign each request from a user to one of several IP addresses. In this case, a single server session can have multiple IP addresses,

- Multiple IP address/Single User - A user that accesses the Web from different machines will have a different IP address from session to session. This makes tracking repeat visits from the same user difficult, and

- Multiple Agent/Singe User - Again, a user that uses more than one browser, even on the same machine, will appear as multiple users.

The aim of the preprocessing step, otherwise known as cleaning of data, is to solve the above mentioned situations either by manual processing or by using some tool. The result of this step is fed as input to the next step of the usage analysis.

1.5.3. Pattern Discovery

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. All of these techniques work on the same objective, that is, to search

and extract information patterns from the web log file. The knowledge or pattern discovered plays a vital role in the interpretation or evaluation. The techniques used are presented below.

(i) Statistical Techniques

The most common and widely used knowledge extracting technique is the statistical techniques. These techniques analyze the session file and perform different kinds of descriptive statistical analyzes like frequency, mean, median, etc., on variables such as page views, viewing time and length of a navigational path. The aim of such web traffic analysis tools is to produce a report that contains statistical information like the most frequently accessed pages, average view time of a page or average length of a path through a site. The report may also include limited low-level error analysis such as detecting unauthorized entry points or finding the most common invalid URI. Despite lacking depth in its analysis, this type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task and providing support for marketing decisions.

(ii) Association Rules

From the web mining perspective, association rule generation can be used to relate pages that are most often referenced together in a single server session. The association rules refer to sets of pages that are accessed together with a support value exceeding some specified confidence and support. These pages may not be directly connected to one another via hyperlinks.

For example, association rule discovery using the Apriori algorithm (Agrawal and Srikant, 1994a), (or one of its variants), may reveal a correlation between users who visited a page containing electronic products to those who access a page about sporting equipment. Apart from being applicable for business

and marketing applications, the presence or absence of such rules can help Web designers to restructure their web site. The association rules may also serve as a heuristic for pre-fetching documents in order to reduce user-perceived latency when loading a page from a remote site.

(iii) Clustering

Clustering is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered :

(a) Usage clusters and
(b) Page clusters.

Clustering of user trends is used to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-commerce applications or provide personalized web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for Internet search engines and Web assistance providers. In both applications, permanent or dynamic HTML pages, that suggest related hyperlinks to the user according to the user's query or past history of information needs, can be created.

(iv) Classification

Classification is the task of mapping a data item into one of several predefined classes (Fayyad *et al.*, 1994). In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. Classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-

nearest neighbor classifiers, Support Vector Machines, etc. For example, classification on server logs may lead to the discovery of interesting rules such as 30% of users who placed an online order in /Product/Music are in the 18-25 age group and live on the West Coast.

(v) Sequential Patterns

The technique of sequential pattern discovery attempts to find inter-session patterns, such that, the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. Other types of temporal analysis that can be performed on sequential patterns include trend analysis, change point detection or similarity analysis.

(vi) Dependency Modeling

Dependency modeling is another useful pattern discovery task in Web Mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the Web domain. As an example, one may be interested in building a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (i.e., from a casual visitor to a serious potential buyer). There are several probabilistic learning techniques that can be employed to model the browsing behavior of users. Such techniques include Hidden Markov Models and Bayesian Belief Networks. Modeling of Web usage patterns will not only provide a theoretical framework for analyzing the behavior of users, but is potentially useful for predicting future Web resource consumption. Such information may help develop strategies to increase the sales of products, offered by the web site or improve the navigational convenience of users.

(vii) Dependency detection

The deviation detection class, a sub area of dependency modeling, contains techniques aimed at detecting unusual changes in the data relatively to the expected values. Such techniques are useful, for example, in fraud detection, where the inconsistent use of credit cards can identify situations where a card is stolen. The inconsistent use of a credit card could be noted if there were transactions performed in different geographic locations within a given time window.

(viii) Summarization

The summarization techniques aim at inferring a compact description of a large data set. A common example is the application of the association rules technique to a big database of sales transactions. The inferred association rules show the items which have high probability of being brought together in a transaction.

1.5.4. Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process. The motivation behind pattern analysis is to filter out uninteresting data or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. In this research, the usage of association rule mining is used for pattern analysis.

Pattern analysis is an evolving and growing area of research and development both in academia as well as in industry that uses mining techniques to discover patterns. Hence, the pattern analysis step is also called as Pattern Mining (Karin, 2004). It involves interdisciplinary research and development encompassing diverse domains. New techniques and directions are being proposed in the literature every day (Kriegel *et al.*, 2003a; Nierman and Jagadish, 2002).

Pattern mining is a process that has become one of the most actively researched topics in web usage mining and knowledge discovery in web log databases (Aggarwal *et al.*, 2009). These techniques analyze historical data to discover previously unknown interesting and useful patterns.

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks, namely, descriptive data mining tasks and predictive data mining tasks (Banerjee and Ghosh, 2002). Descriptive data mining tasks describe the general properties of the existing data. They find human-interpretable patterns that describe the data. Examples include association rule discovery, sequential pattern discovery, clustering, characterization, etc.

Predictive data mining attempts to predict based on inference on available data. They use some variable to predict unknown or future values of other variables. Some predictive data mining techniques are, classification, regression, outlier detection, change/evolution analysis, etc. There are a number of mining methods that are available for this purpose to business today (Figure 1.10).

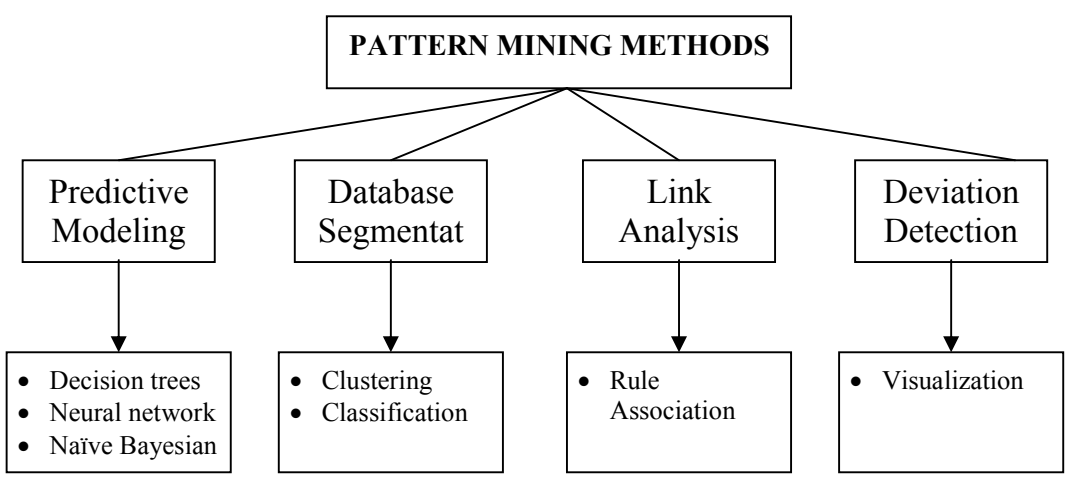

Figure 1.10 : Data Mining Methods

The various methods include data characterization, data discrimination, association analysis, classification, prediction, clustering, etc. Pattern analysis can be applied to different types of data respiratory like text data (both numbers and alphabets), web data (web links, web contents), multimedia data (images, audio, video) and transaction data (market basket, dump data). In this research work, rule association or association rule mining method is used to analyze patterns from web log data for next page prediction and is analyzed in the following section.

1.6. MOTIVATION AND OBJECTIVES

Business enterprises that years ago started collecting and storing large amounts of data, generated as a result of their operational activities, are the ones that currently are capitalizing on their data assets. Such historical data have buried within them patterns, relating to the effectiveness of their various business processes. Data mining has been successfully applied in many industries as a practical tool for knowledge discovery.

These historical data, generated by both humans and intelligent machines, consists of unstructured information, which is difficult to analyze for knowledge discovery. Thus, taking up a structured approach to control this information exchange and extract meaningful knowledge from them is gaining more and more importance in the field of information technology.

The e-commerce companies want to evolve the way to predict the users' behavior and personalize information to reduce the traffic load and design the web site suited for the different groups of users. The business analysts wish to identify tools to learn the user or consumers' needs. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the problems encountered on the Web. Therefore, web mining becomes a popular and active field, and is taken as the research topic in this work.

The usage of WARS is multi-fold and is recapitulated as follows. It can be used to improve the effectiveness of the web sites by adapting the information structure of the sites to the user behaviour. The ease and speed, with which business transactions can be carried out over the web, has been a key driving force in the rapid growth of e-commerce applications. The ability to track user browsing behavior down to individual mouse clicks will bring the vendor, and end customer closer and in turn will improve the business.

The advancement in WARS has made it possible to discover user access patterns from the log file and construct an overall model about the user. This model can be used to perform web surfing on the behalf of the user by pre-fetching pages included in the user model. With the use web usage mining techniques, it is now possible for a vendor to personalize product message for individual customers at a massive scale, a phenomenon that is referred to as mass customization.

Despite these advantages, the field still has various issues which are unresolved. These issues are mainly concerned with the accuracy and speed of predicting a page to the user, which is due to the following challenges.

- The major reason that WARS has attracted a great deal of attention in the e-commerce industry years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The amount of data accumulated each day, about users navigation patterns, around the world is alarming,

- Existing techniques that analyze and discover knowledge are found inadequate while searching for complex multifactor dependencies in data,

- Unlike data mining, the records (= web documents) are not structurally identical and the records are not statistically independent, which makes the

analysis more difficult. The knowledge in WWW lacks an integrated structure or schema which makes it very difficult to access relevant information efficiently, and

- At the same time, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Modeling and analyzing web navigation behavior is helpful in understanding online users demand. Nevertheless, an online navigation behavior grows each passing day and hence extracting information intelligently from it is a difficult issue.

In order to answer the above issues, the key objective of the research work is to develop a WARS for user next web page prediction that will help both web administrators and users to find and extract desired information and resources in an e-commerce website in an easy, fast and accurate manner. To attain the above key objective, the following aims were formulated :

1. To design and develop preprocessing techniques that improve the quality of web log database through various steps including cleaning of web log data, user identification and enhanced session identification techniques, so as to improve the performance of page prediction,

2. To design and develop techniques that reduce the size of web log data by identifying only the potential users from the preprocessed web log data, and

3. To design and develop web page prediction system, using association rule mining classification, that considers customer emotions and interest in the webpage.

1.7. CHAPTER FORMULATION

The underlying objective of this research work is to develop an efficient tool for predicting user's next page during browsing from web log files. This chapter (**Chapter 1, Introduction**) provided the introductory materials covering data mining, web data mining with emphasis on web usage mining. The objectives of the research work were also outlined. The rest of the thesis is organized as follows.

The literature review is a critical look at the existing research that is significant to the work that is carried out. In case of data mining, several researchers have addressed the problem of web usage mining and next page prediction. A critical look at the various available literatures related to the present research work is given in **Chapter 2, Review of Literature. Chapter 3, Methodology** presents the research methodology and identifies the different steps of the proposed WARS. The various methods and techniques used in each of these steps are introduced in this chapter.

Chapter 4, Preprocessing Algorithms, Chapter 5, Potential User Identification and Chapter 6, Web Log Associative Classification respectively describe the techniques used to preprocess the web log data to potentially identify users and sessions, method to reduce the size of web log database through potential user identification and predict next page using associative classification.

Chapter 7, Results and Discussion, tabulates and analyzes the various results obtained while testing the proposed algorithms in each step of the proposed WARS. The findings of the study are summarized along with future research directions in **Chapter 8, Summary and Conclusion**. The work of several researchers are quoted and used as evidence to support the concepts explained in this dissertation. All such evidences used are listed in the **Bibliography** of the thesis.

1.8. CHAPTER SUMMARY

This chapter provided a brief introduction to the research work along with the research objectives. To achieve the objectives outlined in this chapter and in order to develop an enhanced next web page prediction system, a review of the previous research works was performed. The next chapter, **Review of Literature**, presents details regarding the literature study conducted.