Introduction

1. INTRODUCTION

The World Wide Web (WWW), in the current era of information explosion, has become the major source of online data, which includes text, graphics, videos, sound, etc. WWW is global information medium where users read, write and communicate via computers connected to the Internet. Recent studies have estimated that the Web has more than one billion pages. It has become the defacto technology for sharing new ideas and content exchange (Inktomi, 2000). The impact of the Internet on everyday life is tremendous and it has changed the way of doing 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.

Today, the WWW and Internet has become the largest information source available in this planet. At an annual conference of the National Advertisers Association, Eric Schmidt, Google CEO had said that, from the data recorded by the search engine, the Internet is made of 5 million terabytes. According to Facebook (2010), there are more than 400 million active users and more than 23.89 billion pages available in the Web. The number of people using WWW is around 20.1 percent in Asia alone and more than 234 million websites and 126 blogs (www.thisblogrules.com/2010/07/facts-about-the-internet-infographic.html).

This statistics are expected to grow as the number of people using the internet increases. Thus, it can be concluded that the Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of users, Web service providers and business analysts. The users want to have the effective search tools to find relevant information easily and precisely.

This Tremendous growth has motivated the web service providers to predict the user's web usage behaviors to an extent that, they can

(i) Personalize the information provided

(ii) Make the websites more user friendly

(iii) Reduce the traffic load

(iv) Create or modify their website to suit different group of people.

Some tools which will help system analysts and business persons to learn user / consumer's needs, so that user requirements or demand can be solved immediately. To develop such a tool, the information available on the net has to be well understood to make it an efficient and effective tool for the analysis of web user requirements. Organizations and web developers are facing the problem of Information overload, which is a term that refers to the difficulty a person can have in understanding the issue and making decisions that can be caused by the presence of too much information (Yang *et al.*, 2003). Data mining is a technique which can be of great help in these situations. Different kinds of data mining techniques are available which are categorized according to the domain they work. Examples include text mining, image mining, database mining and web mining. This research work focuses on the problem of mining and knowledge from web data and is explained in the following sections.

1.1. WEB MINING – AN OVERVIEW

Web mining is the application of machine learning (data mining) techniques to web-based data for the purpose of learning or extracting knowledge. 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 was first proposed by Etzioni in 1996 and according to him web mining is the concept of using data mining techniques to automatically discover and extract information from WWW documents and services. WWW documents include data from hyperlinks between documents; usage statistics log data, web content, and web structure.

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 tasks, 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 to extract knowledge from web data, where at least one of the structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data).

1.2. GENERAL FRAMEWORK OF WEB MINING PROCESS

The concept of web mining started in the early 90's. In the beginning it was used to observe the user behaviour from their viewing, book marking and browsing history. The term web mining was initially used by (Etzioni, 1996) to define task oriented method and later was defined to be data oriented method (Cooley *et al.*, 1997). A general web mining scenario is given in Figure 1.1.

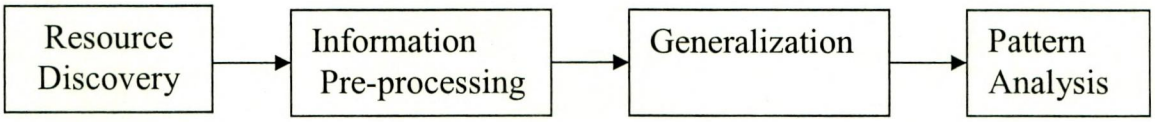

Figure 1.1: General Framework of Web Mining

According to this web mining is an activity for 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,

namely, (i) resource discovery, (ii) information pre-processing, (iii) generalization, and (iv) pattern analysis.

Resource Discovery is a task for retrieving information from web resources and documents. Web information retrieval is a process to find the 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 a transform process of the 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.

1.3. PARADIGM OF WEB MINING

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined (Figure. 1.2):

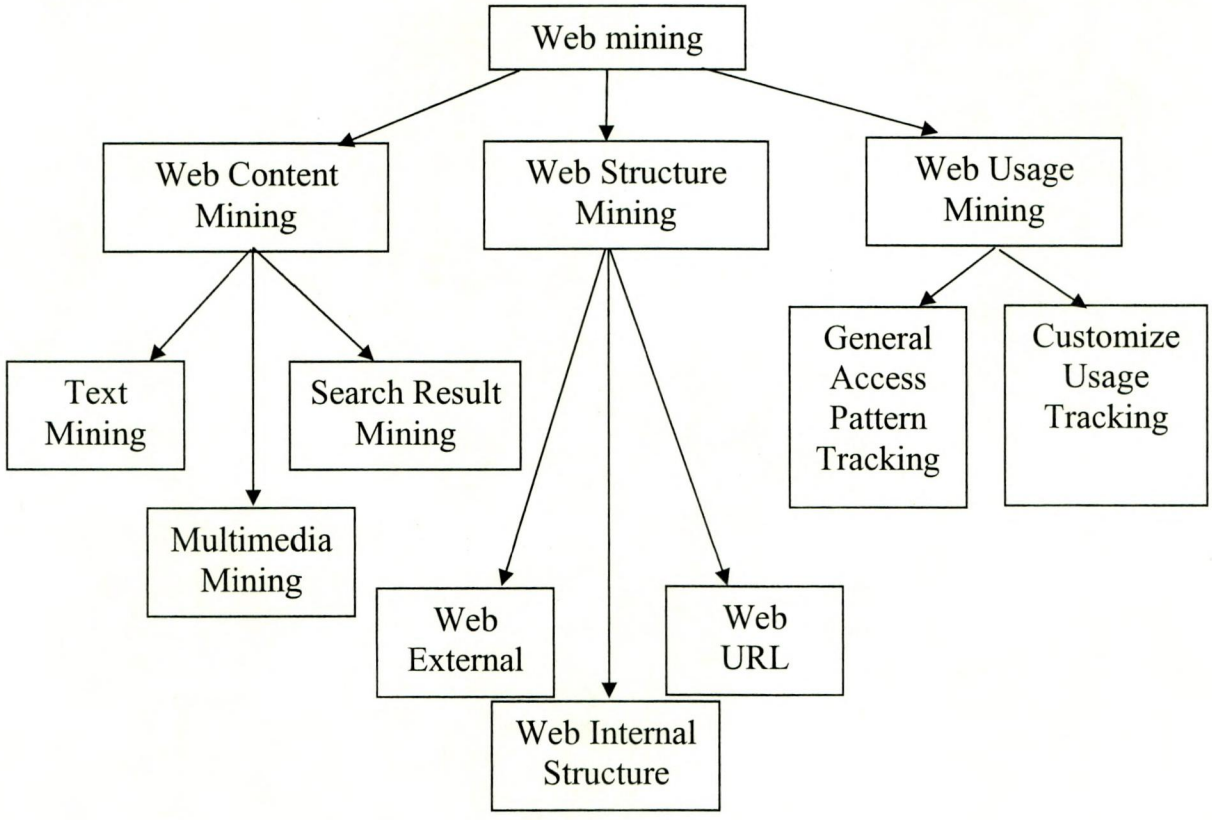

Figure 1.2: Paradigm of Web Mining

- **Web Content Mining**

Web Content Mining is the process of extracting knowledge from web contents. Web content mining can be divided into text mining (including text file, HTML, DHTML, XML document, etc.), multimedia mining and the Search Result Mining. The main categories of Web text mining are text categorization, text clustering, association analysis, trend prediction and so on. Multimedia mining deals with the extraction of implicit knowledge, multimedia data relationships, or other patterns not explicitly stored in multimedia files. Multimedia mining is more than just an extension of data mining, as it is an interdisciplinary endeavor that draws upon expertise in computer vision, multimedia processing, multimedia retrieval, data mining, machine learning, database and artificial intelligence (Kotsiantis *et al.*, 2004). Search Result Mining is concerned with extracting information from the search result page. It is closely related to text mining and some application include topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP).

- **Web Structure Mining**

Web structure mining helps the user to retrieve the relevant documents by analyzing the link structure of the Web (Kumar *et al.*, 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. This type of mining can be further divided into three kinds based on the kind of structural data used, web external, web URL and web internal structure. Another application of web structure mining is to discover the structure of the web document itself, so that user navigation comparison between different web pages can be analyzed. This type of structure mining

will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema.

- **Web Usage Mining**

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 tendencies 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 is aiming 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 display the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns.

1.4. WEB USAGE MINING

Web usage mining, popularly known as web log mining. 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, online and off-line with respect to the Web server activity (Baraglia et al., 2007; Frias_martinez et al., 2003; Jalali *et al.*, 2008; Yan *et al.*, 1996). The off-line component is aimed at building the knowledge base by analyzing historical data, such as server access log files, that is then used in the online component.

There are three generic types of Web applications:

- Revolutionary applications: They have emerged with the Web and have no counterpart in the pre-Web era.
- Innovative applications: They have emerged with Information Technology. The capabilities and particularities of the Web have a major impact on them. Applications like E-learning belong to this category.
- Web-empowered conventional applications: They were transferred in the Web context; the Web revolutionized the way of doing them. Examples include marketing of products and literature search.

1.4.1. Application Areas

As shown in Figure 1.3, usage patterns extracted from Web data have been applied to a wide range of applications. They are

- Personalization
- System Improvement
- Site Modification
- Business Intelligence
- Usage characterization

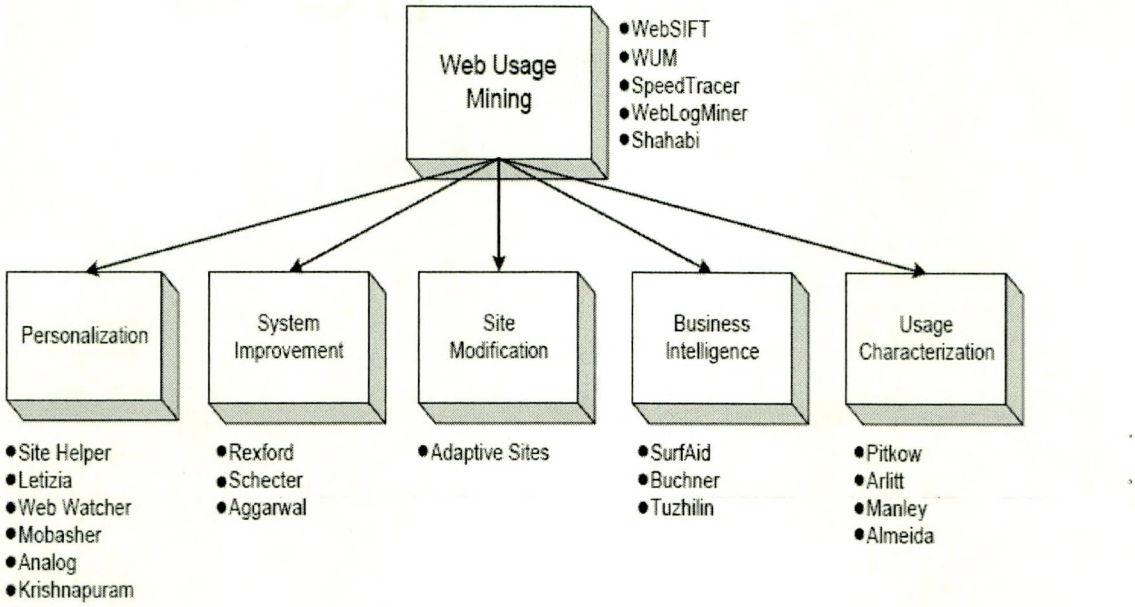

Fig. 1.3: Applications Areas

Personalizing the Web experience for a user is very advantageous for many Web-based applications e.g. 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, e.g. 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.

Performance and other service quality attributes are crucial to user satisfaction from services such as databases, networks, etc. 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 et al., 1999). Web usage mining can also provide patterns which are useful for detecting intrusion, fraud and attempted break-ins.

The attractiveness of a Web site, in terms of both content and structure, is crucial to many applications, e.g. a product catalog for e-commerce. Web usage mining provides detailed feedback on user behavior. In a similar fashion, information on how customers are using a Web site is crucial information for the marketers of e-tailing businesses. The result of web usage mining depends on the web usage data and is explained in the next section.

1.5. 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. Web logic [BEA], Broad Vision [BV], Story Server [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.

1.5.1. Data Sources

The web usage data can be collected either from server side, client side, and 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, multi-site access patterns. The potential data sources are explained in this section and are illustrated in Figure 1.4.

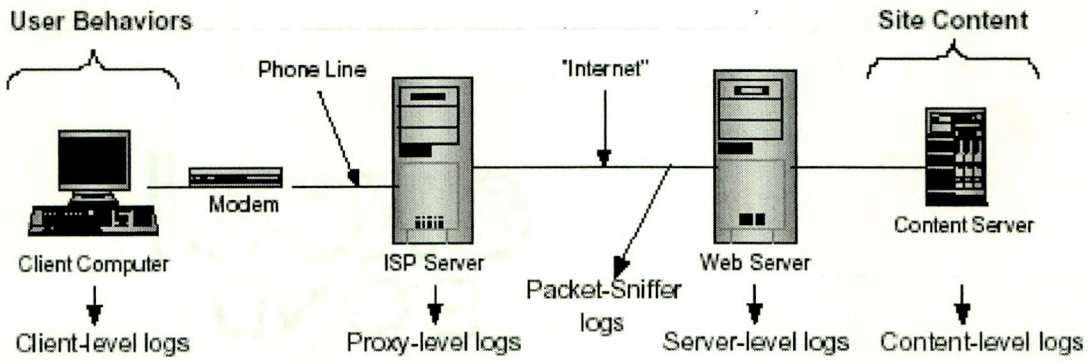

Figure 1.4: Potential Data Sources

- **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. An example of extended log format is given in Figure 1.5. 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.

- **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 cooperation, either in enabling the functionality of the Javascripts and Java applets, or to voluntarily use the modified browser.

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	209.456.78.2	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	209.456.78.3	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1660	A.html	Mozilla/3.04 (Win95, I)

Fig. 1.5: Sample Web Server Log

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 All Advantage 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.

1.6. GENERAL WEB LOG MINING SYSTEM

A general web log mining system consists of three steps, namely, preprocessing, pattern discovery and pattern analysis (Figure 1.6).

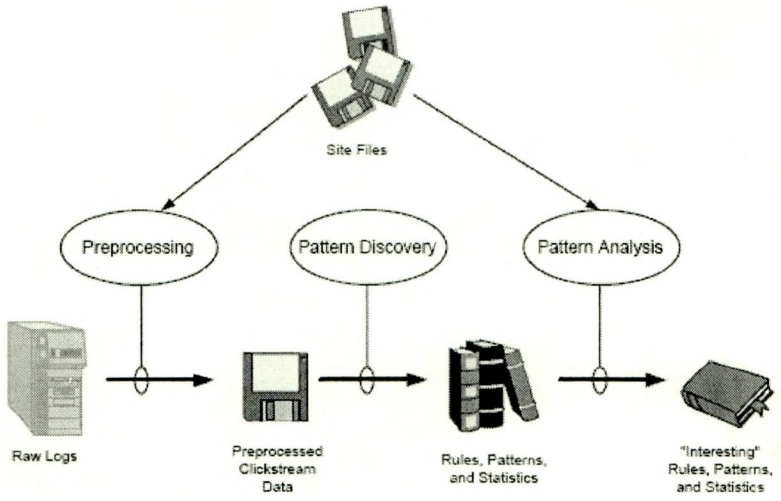

Fig. 1.6: Knowledge Discovery

1.6.1. Web Log File

Web log file is a log file created automatically and maintained 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 about who was visiting the site, where they came from, and exactly what they were doing on the web site. A typical web log format along with a partial sample file is shown in Figures 1.7 and 1.8.

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

Figure 1.7: Format of Web Log File

```
203.30.5.145 www.acr-news.org - [01/Jun/1999:03:09:21 -0600] "GET /Calls/OWOM.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/OWOM.html"
"Mozilla/4.5 [en] (Win98; I)"
```

Figure 1.8: Sample Web Log File

Typically a web log files stores information about

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;
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.
- Each Web server has its own log file format.

1.6.2. 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.

- Multiple Agent/Single Users - 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.

1.6.3. 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

(i) Statistical techniques
(ii) association rules
(iii) classification
(iv) clustering
(v) Sequential patterns
(vi) dependency modeling
(vii) dependency detection and
(viii) Summarization

(i) Statistical Techniques

The most common and widely used knowledge extracting technique is the statistical technique. These techniques analyze the session file and perform different kinds of descriptive statistical analysis 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.

(ii) Association Rules

An association rule is a statement of the form, $A \Rightarrow B$, where A and B are attributes and are characterized by two important parameters

(i) Support and

(ii) Confidence.

The support of a rule is the percentage of all transactions which contains all the attributes in A and B. The confidence of a rule is given by the percentage of the transactions containing A that also contains B.

From 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 et al., 1994) (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. Aside 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 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 users tends 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, static or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs.

(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 groups and live on the West Coast.

(v) Sequential Patterns

The technique of sequential pattern discovery attempts to find inter-session patterns such as the presence of set of items 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 to build 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 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 which items have high probability of being bought together in a transaction.

1.6.4. 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 rules or patterns from the set found in the pattern discovery phase. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

Thus a web usage mining system consists of three main areas, namely, preprocessing, pattern discovery and pattern analysis. In the present research work, pattern discovery is used to detect the web users' navigation pattern and is performed using clustering technique.

1.7 CLUSTERING AND CLASSIFICATION

Clustering is a technique that groups together users or data items (pages) with the similar characteristics. It is a method that divides a set of objects into subsets, called clusters, such that the objects in any cluster are similar to those

inside it and different from those outside it (Ivancsy et al., 2006). The objects are described with numerical or nominal variables.

In web usage mining, traditional clustering techniques, such as distance-based methods cannot handle web data (Mobasher *et al.*, 2000). The reason is, instead of using URLs the transactions must be used as features, whose number is in tens to hundreds of thousands in a typical application. Furthermore, dimensionality in this context may not be appropriate, as removing a significant number of transactions as features may lose too much information.

Several computer scientists have proposed novel and successful approaches for solving problems in web usage pattern discovery (Labroche *et al.*, 2003). One technique that is commonly used is ant-based method. In this method, artificial ants are considered as agents, which do not communicate directly with each other, but influence themselves through the configuration of objects on the floor. Thus the agents construct groups of similar objects or construct clusters.

Classification is another area of research which has gained equal importance in the applications related to WWW. Predicting user request is an area that is in great demand today. Classification, which is defined as assigning a predefined label to an incoming input, is used for this purpose. The goal of classification in web usage mining is to follow user navigation and to track its behavior during a session so that an adequate help could be provided to the user online. There have been some studies for defining navigation behaviors' in WWW (Gallinari et al., 2003; Liu et al., 2007).

1.8. MOTIVATION AND OBJECTIVES

The volume of information transaction from web servers and the number of requests from web users are continuously growing in WWW. Providing web administrators with meaningful information about user access behaviour and usage patterns has become a necessity to improve the quality and performance

of web services. They help to engage new customers, maintain current customers, and track customers who are leaving web site, and so on. Usage information can be extracted to increase web server efficiency by pre fetching and caching strategies. Thus, the hidden knowledge from web server traffic and user access pattern could be applied directly for marketing and management of E-applications like E-business, E-services, E-searching E-education, etc.

The knowledge in WWW lacks an integrated structure or schema which makes it very difficult to access relevant information efficiently. 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' demands. Following that, the analyzed results can be seen as knowledge to be used in intelligent online applications, refining web site maps, web based personalization system and improving searching accuracy when seeking information. A clustering algorithm discovers groups in the set of documents such as documents within a group or more similar than document across groups. It is a classic area of machine learning and pattern recognition. It is considered challenging in web mining because of the hypertext domain. One basic problem is the number of attributes or features available, which is very large and finding similarity between them is very difficult.

Etminani *et al.* (2009) used Ant-based clustering method to discover and extract user's navigational pattern from web log files. The classic ant colony clustering algorithm takes use of the characteristics of positive feedback of the ant colony. Such algorithm is robust, good convergence, and parallel. However, it is also with the disadvantage of long time convergence, easy stagnation and local optimization. In the present work, Etminani *et al.* (2009) hereafter referred to as "Base Model", is enhanced by following the ant-based clustering method by a classification method called LCS (Longest Common Subsequence) method. The LCS method was proposed by Jalali *et al.* (2008)

and is used to improve the accuracy of pattern discovery. The enhanced Base Model hereafter will be referred to as "Proposed Model".

To answer the above challenges the present research work focus on the area of web usage mining from log data to understand user navigation pattern through the use of clustering. The aim is to develop intelligent tool for user navigation pattern information which will help both web administrators and users to find and extract desired information and resources in an efficient manner. To attain the above aim, the following objectives were formulated:

- To develop user-navigation pattern discovery and analysis tool and an ant-based clustering algorithm for pattern analysis from web log files.
- To preprocess server logs files from the Web servers for determining and discovering the user navigation pattern through clustering.
- To implement the tool based on ant-based clustering algorithm.
- To enhance the tool with Longest Common Subsequence (LCS) algorithm
- To analyze and compare the output of both ant-based tool and LCS-based tool in their efficiency.

To attain the above objectives, the proposed system is designed as in Figure1.9.

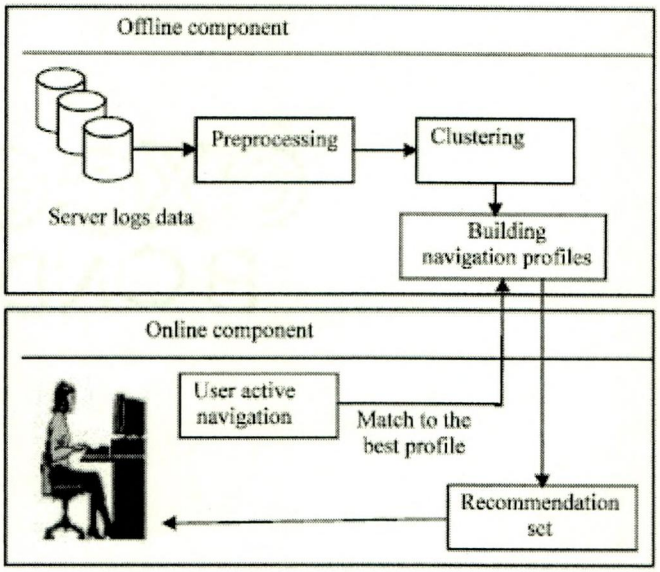

Figure 1.9: Proposed System

The system is divided into two stages, the online and offline stage. During the offline stage, the server log files goes through a series of preprocessing steps to identify accurate transactions, which is clustered to identify the navigation patterns. In online, a classification algorithm is used to analyze the historical behaviour and predict the future request of a user.

1.9. CHAPTER FORMULATION

The underlying objective of this research work is to develop an efficient tool for discovering user navigation pattern from web log files. In the present context, the work of Etminani *et al.* (2009) is enhanced by the use of a classification technique called LCS. The algorithm and results obtained are reported in this dissertation. The dissertation is organized as follows.

Chapter 1 provides a brief introduction to web usage mining and its applications for knowledge discovery. This chapter also outlined the objective of the research work.

The review of literature is a critical look at the existing research work and it is very significant to the current work. In the case of data mining, several researchers have addressed the problem of web mining. A critical look at the

various literatures related to the present research work is given in Chapter 2, Review of Literature.

The main components of the web mining tool developed in this research are clustering algorithm and classification algorithm. These two algorithms are explained in Chapter 3. The base system and the proposed system were compared and various results obtained are tabulated and discussed in Chapter 4. The conclusion of the research work is summarized along with future research direction in Chapter 5.

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 reference section of the dissertation.

1.10. CHAPTER SUMMARY

This chapter provides a brief introduction to the research work and also highlights the objectives. To achieve the outlined objectives, a review of the previous research work was studied and the scrutinized works are summarized in the next chapter, Review of Literature.