CHAPTER I

INTRODUCTION

INTERNET is a useful source of information in everyone’s daily activity. Hence, this made a huge development of World Wide Web in its quantity of interchange and its size and hence increasing the difficulty of websites. Moreover, e-business and web marketing are quickly developing and significance of anticipating the requirement of their customers is obvious particularly. As a result, guessing the users’ interests for improving the usability of web mining process.

Web mining is attractive for companies because of several advantages. In the most general sense it can contribute to the increase of profits, be it by actually selling more products or services, or by minimizing the costs. In order to do this, marketing intelligence is required. This intelligence can focus on marketing strategies and competitive analyses or on the relationship with the customers. The different kinds of web data that are somehow related to customers will then be categorized and clustered to build detailed customer profiles. This not only helps companies to retain current customers by being able to provide more personalized services, but it also contributes in the search for potential customers.

There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and web browser activity tracking.
- Web graph, from links between pages, people and other data.
- Web content, for the data found on web pages and inside of documents.
1.1 DATA MINING

Data mining is also called Knowledge Discovery in Databases (KDD). It is commonly defined as the process of discovering useful patterns or knowledge from data sources, e.g., databases, texts, images, the web, etc. The patterns must be valid, potentially useful, and understandable. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization. There are many data mining tasks. Some of the common ones are supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining.

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

**Pre-processing:** The raw data is usually not suitable for mining due to various reasons. It may need to be cleaned in order to remove noises or abnormalities. The data may also be too large and/or involve many irrelevant attributes, which call for data reduction through sampling and attribute selection. Details about data pre-processing can be found in any standard data mining textbook.

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

**Post-processing:** In many applications, not all discovered patterns are useful. This step identifies those useful ones for applications. Various evaluation and visualization techniques are used to make the decision.
The whole process (also called the data mining process) is almost always iterative. It usually takes many rounds to achieve final satisfactory results, which are then incorporated into real-world operational tasks. Traditional data mining uses structured data stored in relational tables, spread sheets or flat files in the tabular form. With the growth of the web and text documents, web mining and text mining are becoming increasingly important and popular. Web mining is the focus of this research.

1.2 WEB MINING

Web mining is the application of data mining techniques to discover patterns from the web. According to analysis targets, web mining can be divided into three different types, which are web usage mining, web content mining and web structure mining. The process of getting information from web server logs is called web mining. It might include both textual data and also multimedia data. A few users are interested in text form of records, whereas some others might be looking for multimedia or animation form of data. Web usage mining data can be classified into three types. Those are web server data, application server data and application level data.

- **Web Server Data**: The web server is used to collect the users’ log. Typically the number of data can include IP address, page reference, access time and so on.

- **Application Server Data**: Application server software’s like web logic, Broad Vision, Story Server used for e-commerce have important properties in their structure. These properties will allow e-commerce applications to be built on top of them. One of the most important properties of application servers is their ability to keep track of several types of business transactions and record them in application server logs.
• **Application Level Data:** At the application server, the number of event types is increased when move to upper layers. Application level data can be logged in order to generate histories of specially defined events. This type of data is classified into three categories based on the source of information such as server side, client side and proxy side data.

![Fig 1.1: Web Mining Architecture](image)

The web mining architecture shown in Figure 1.1 is explained as follows,

**a. Data Cleaning**

Techniques to clean a server log to reduce irrelevant items are of importance for any type of web log analysis, not just data mining. The discovered associations or reported statistics are only useful if the data represented in the server log gives an
accurate picture of the user accesses of the website. Elimination of irrelevant items can be reasonably accomplished by checking the suffix of the Uniform Resource Locator (URL). For instance, all log entries with filename suffixes such as, gif (GIF), jpeg (JPEG), jpg (JPG), and map can be removed.

b. Transaction Identification

Before any mining is done on web usage data, sequences of page references must be grouped into logical units representing web transactions or user sessions. A user session is all of the page references made by a user during a single visit to a site. Identifying user sessions is similar to the problem of identifying individual users. A transaction differs from a user session in that the size of a transaction can range from a single page reference to the entire page references in a user session, depending on the criteria used to identify transactions. Unlike traditional domains for data mining, such as point of sale databases, there is no convenient method of clustering page references into transactions smaller than an entire user session.

c. Data Integration

Portions of web usage data exist in sources as diverse as web server logs, referral logs, registration files, and index server logs. Intelligent integration and correlation of information from these diverse sources can reveal usage information which may not be evident from any one of them.
d. **Pattern Discovery**

The emerging tools for user pattern discovery use sophisticated techniques from Artificial Intelligence (AI), data mining, psychology, and information theory to mine for knowledge from collected data.

For example, the WEBMINER [97] system introduces a general architecture for web usage mining. WEBMINER automatically discovers association rules and sequential patterns from server access logs.

e. **Pattern Analysis**

Once access patterns have been discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns. Examples of such tools include the WebViz [102] system for visualizing path traversal patterns. Others have proposed using Online Analytical Processing (OLAP) techniques such as data cubes for the purpose of simplifying the analysis of usage statistics from server access logs.

f. **Path Analysis**

There are many different types of graphs that can be formed for performing path analysis, since a graph represents some relation defined on web pages (or other objects). The most obvious is a graph representing the physical layout of a website, with web pages as nodes and hypertext links between pages as directed edges.

1.2.1 **Advantages of Web Mining**

Web mining is attractive for companies because of several advantages. In the most general sense it can contribute to the increase of profits by actually selling more products or services, or by minimizing the costs. In order to do this, marketing intelligence is required.
This intelligence will concentrate on marketing strategies and competitive analysis or on the relationship with the users. The different kinds of web data that are somehow related to users will then be categorized and clustered to build detailed user profiles. The web mining is not only helps companies to retain current users by being able to provide more personalized services, but it also contributes in the search for potential users.

1.2.2 Disadvantages of web mining

There are few disadvantages in web mining such as long time to explore large volume of data, loss of communication link, dynamic nature of web data and interconnection of web pages by hyperlinks which may lead to infinite loop and hidden web data cannot be located.

Web usage mining by itself does not generate issues, but this knowledge is used on data of private nature may effect concerns. The web usage mining is used for the popular criticized moral of concern, which is the attack of privacy. Privacy is well planned for missing data when information regarding a person is obtained mainly if privacy occurs without knowledge or consent.

1.3 HISTORY OF WEB MINING

Web mining started when someone with the sniffles got on the computer and googled 'flu'. The folks over at Google then compared billions of flu-related search queries (also known as 'search trends') to actual flu time series and used the best matches to create Google Flu Trends. By aggregating Google search data, Google Flu Trends can find certain terms that are fairly good indicators of flu activity, and then track such activity through time and space. Not only did the approach garner a publication in nature,
but it was expanded to predict outbreaks of other diseases, and was later applied to economic applications such as unemployment forecasting. Popular Google Apps such as Trends, Insights and Correlate quickly followed.

Web Mining or Web Scraping as it is sometimes called Google Flu Trends. New technologies, data strategies and communication flows present an exciting opportunity for scholars in the Humanities and Social Sciences – both quantitative and qualitative in orientation - who wish to harness the power of the web.

1.4 WEB MINING TAXONOMY

A high level information search device constructed to provide a means of understanding, navigating, and gaining access to intellectual capital.

![Figure 1.2: Web Mining Taxonomy](image-url)
Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. Web mining can be broadly divided into three categories:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

All of these three categories as shown in Figure 1.2 focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the web.

1.4.1 Web Content Mining

Web content mining as shown in Figure 1.3 is the established collections of multimedia credentials. The multimedia credentials such as images, video and audio are linked to the web pages. It is too fairly similar from data mining because, web data are chiefly semi-structured or unstructured. Web content mining thus requires inspired applications of data mining or text mining techniques and also its own distinctive approaches. Web content mining is also different from text mining, while text mining focuses on unstructured texts. It mainly focuses on the structure of inner-document. Web content mining might be differentiated from two points of view: agent-based approach or database approach. The first approach is used for improving the information decision and filtering. The second approach aims on modeling the data on the web into more ordered form in order to affect standard record querying method and data mining applications to analyze it.

Web content mining, also known as text mining, is generally the second step in web data mining. Content mining is the scanning and mining of text, pictures and graphs of a web page to determine the relevance of the content to the search query.
This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query.

With the massive amount of information that is available on the World Wide Web, content mining provides the results list to search engines in order of highest relevance to the keywords in the query.

Text mining is directed toward specific information provided by the customer search information in search engines. This allows for the scanning of the entire web to retrieve the cluster content triggering the scanning of specific web pages within those clusters. The results are pages relayed to the search engines through the highest level of relevance to the lowest. Though, the search engines have the ability to provide links to web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information.

Web text mining is very effective when used in relation to a content database dealing with specific topics. For example online universities use a library system to recall articles related to their general areas of study. This specific content database enables to pull only the information within those subjects, providing the most specific results of search queries in search engines. This allowance of only the most relevant information being provided gives a higher quality of results. This increase of productivity is due directly to use of content mining of text and visuals.

The main uses for this type of data mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information. This tool is imperative to scanning the many HTML documents, images and
text provided on web pages. The resulting information is provided to the search engines in order of relevance giving more productive results of each search.

Web content categorization with a content database is the most important tool to the efficient use of search engines. A customer requesting information on a particular subject or item would otherwise have to search through thousands of results to find the most relevant information to his/her query. Thousands of results through use of mining text are reduced by this step. This eliminates the frustration and improves the navigation of information on the web.

Business uses of content mining allow for the information provided on their sites to be structured in a relevance-order site map. This allows for a customer of the website to access specific information without having to search the entire site. With the use of this type of mining, data remains available through order of relativity to the query, thus providing productive marketing. Used as a marketing tool this provides additional traffic to the web pages of a company’s site based on the amount of keyword relevance the pages offer to general searches.

As the second section of data mining, text mining is useful to improve the productive uses of mining for businesses, web designers, and search engine operations. Organization, categorization and gathering of the information provided by the WWW become easier and produce results that are more productive through the use of this type of mining.
Figure 1.3: Content Based Mining
1.4.2 Web Content Mining Problems and Challenges

Web Content Mining contains problems which are recovered by using some challenges below.

a. Data/Information Extraction

Data/information extraction is that extracting a required data from web pages. Some difficult task may also taken place such as products and search results. Two main types of techniques are used to solve this problem. One is machine learning and another one is automatic extraction.

b. Web Information Integration and Schema Matching

While the web contains a large amount of information, every website represents parallel information in a different way. Identifying or matching semantically related data is a very important problem with much realistic application.

c. Opinion Extraction from Online Sources

There are many online judgment sources, e.g., client reviews of products, forums etc., the marketing intellect and merchandise benchmarking may use mining opinion.

d. Knowledge Synthesis

Concept hierarchies or ontology are valuable in many applications, but generating physically means time overriding. A few existing techniques explore that the information redundancy of the web, which are discussed in the literature.
e. Segmenting Web Pages and Detecting Noise

In many web applications this work only needs the main content of the web page without unnecessary data like advertisements, routing links and copyright notices. Mechanically segmenting web page to take out the main content of the pages is attractive and also difficulty.

1.4.3 Web Structure Mining

Web structure mining as shown in Figure 1.4 focuses on study of the link configuration of the web. The main use of this is to classify more preferable documents. The different items are connected in various manners. Web structure mining helps in discovering similarity between websites or discovering important sites for an exacting area.

Just applying the established process and assuming that the actions are free can show the way to wrong conclusions. However, the proper behavior of the associations could lead to probable correlations, and then pick up the analytical accuracy of the learned models. The goal of web structure mining is to create summary about the website and web page. Web structure mining will sort out the web pages and generate the information to facilitate information such as the parallel and relationship between different websites.

![Figure 1.4: Structure Based Mining](image-url)
a. **Hyperlink**

A hyperlink is a connecting page that the unit that connects a spot in a web page to a dissimilar site. That either within the similar web page or on a dissimilar web page. An intra-document hyperlink is that connects to a different part of the same page, and inter-document is that hyperlinks that connect two dissimilar pages. There has been a significant body of work on hyperlink analysis, of which provides an up-to-date survey.

b. **Document Structure**

In addition, the content within a web page can also be organized in a tree structured format, based on the different HTML and XML tags within the page. Mining have observed on mechanically extracting Document Object Model (DOM) structures out of documents.

1.4.4 Web Usage Mining (WUM)

It focuses on techniques that could predict the behavior of users while this works are interacting with the WWW. Web usage mining used the functional information from the derived data. It collects the information from web log records to learn user access patterns of web pages. There are several available research projects and profitable tools those for special purposes. The approaching information can be utilized in personalization, structure perfection, site adjustment, and industry brain power and usage characterization.

Web Usage Mining (analyzing user web navigation in Figure 1.5) is a process of extracting information from user how to navigate websites. Web usage mining also known as web log mining, aims to discover interesting and frequent user access patterns
from web browsing data that are stored in web server logs, proxy server logs or browser logs. Applications of Web Usage Mining are:

1. Personalization: Reconstruct the website based on user’s profile and usage behavior.
2. System Improvement: Provide help to understanding web traffic behavior. Web load balancing and data distribution or policies for web caching are benefits of such improvements.
3. Adjustment of website: Understanding visitor’s behavior in a website provides hints for adequate design and update decision.
4. Business intelligence: occupying the application of intelligent techniques in order to help certain businesses, mainly in marketing.
5. Advertisement: Valuing the effectiveness of advertising by analyzing large number of access behavior patterns.
6. Improving the design: An e-commerce website according to user’s browsing behavior on site in order to better serve the needs of users.

Figure 1.5: Web Usage Mining
In general, there are mainly four kinds of data mining techniques applied to the web mining domain to discover the user navigation pattern. Those are association rule mining, sequential pattern, clustering and classification.

a. Association Rule mining

Association rule mining is the discovery of association relationships or correlation among a set of items. Association and correlation is used to find the frequent item set among large datasets. Association rule for a given dataset is very large and it generally in (if any) value. The main task of association rule mining is to find sets of binary variable that co-occur together frequently in a transaction database. Association rule holds many algorithms like Apriori, Coeur d’Alene (CDA), Delhi Development Authority (DDA) and Interestingness Measure. Association rules are if then statements that search uncovered relationship between unrelated data in the relational database. Many association rule mining techniques exist on the basis of literature such as multilevel association rule, multi dimensional association rule and quantitative association rules.

i) Problems with the association rule mining

Single minsup: It assumes that all items in the data are of the same nature and/or have similar frequencies.

Not true: In many applications, some items appear very frequently in the data, while others rarely appear.

- E.g., in a supermarket, people buy food processor and cooking pan much less frequently than buy bread and milk.
b. Sequential Pattern Mining

The finding of order patterns in web server access logs allows web-based organizations. That organization is used to expect user steering patterns and helps in targeting promotion. By analyzing this information, the web mining system can determine temporal relationships among data items.

c. Clustering

Clustering as the name suggests is the process of grouping data into classes, so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other cluster. Dissimilarities have been observed on the basis of attribute value describing the objects often distance used. It is the process of identification of similar classes of object. It is an unsupervised learning technique that shows the natural groupings in data. Clustering has been frequently used in data mining applications for discovering patterns in huge datasets.

d. Classification

Classification is a supervised learning technique. It maps the data into predefined groups. It is used to develop a model that can classify the population of records at large level. Classification algorithm requires that the classes be defined based on the data attribute value. It describes these classes according to the characteristics of the data that is already known to belong to the classes. The classifier training algorithm uses these pre-defined examples to determine the set of parameters required for proper discrimination. Some of the useful data mining techniques such as Decision Tree, Neural Networks, Bayesian Classification etc.,
1.5 TEXT MINING USED IN WEB MINING

The text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. The combination of text mining and web mining are presented in Figure 1.6.

- **Information Retrieval (IR) Systems** equivalent a user’s doubt to documents in a collection or database. The first step in the text mining practice is to locate the body of credentials that are associated to the research question.

- **Natural Language Processing (NLP)** analyzes the text in structure based on person speech. It allows the computer to achieve a grammatical study of a sentence to “read” the text.

- **Information Extraction (IE)** involves structuring the data that the NLP system generates.

- **Data Mining (DM)** is the process of identifying patterns in large sets of data, to find that new knowledge.

1.5.1 Uncover Information Hidden in Text

- Application of data mining to unstructured or less structured text files.

- Attempts to categories textual data and not to understand its contents

- Entails the production of important mathematical indices from the shapeless text and then processing these indices using a variety of data mining algorithms

- Organize a knowledge for a classification problem with a binary output (i.e. whether or not a document is about a specific topic)
1.5.2 Text Mining Algorithm Involvements

- Train – Build a quality dictionary where the attributes characterize words from related work to an exacting subject matter. Choose only words that take place a less number of times.

- Filter – Take out general vocabulary known to be useless in differentiating works.

- Classify – Make sure every text to be classified for being there and regularity of the selected attributes. Categorize the document under a particular area if it contains a set of least number of references to the selected attributes for the topic.

The text mining and web mining definitions and types are shown in Figure 1.6.

---

**Figure 1.6: Text and Web Mining**
1.5.3 Challenges of Textual Data Mining

- More tricky textual mining problems occupy the study of free-form text initiate in e-mail documents and recorded telephone transcripts.
- The challenge of textual data mining is managing ambiguities such as spelling and grammar errors.
- Text contains acronyms, abbreviations, misspellings. (e.g. customer, cust, customer, csmr)
- Semantic analysis: understanding the meaning of words (i.e. book = to reserve vs book = a manual)
- Syntax analysis: understands a sentence’s structure and the roles of words (i.e. subject, verb, preposition, noun)

1.5.4 Applications of Text Mining

- Automatic detection of e-mail spam or phishing through analysis of the document content.
- Mechanical processing of messages or e-mails to route a message to the most suitable party to process that message.
- Analysis of guarantee claims, facilitate desk calls/reports, and so on to classify the most common problems and related responses.
- Filter & match resumes to open positions
1.6 KEY MODELS DEVELOPED IN WEB MINING

In Figure 1.7(a) illustrates page rank concept clearly, by showing how the page rank of the page P is calculated. This can be seen as depicted in Figure 1.7(b), where the nodes on the left represent the hubs and the nodes on the right represent the authorities. The hub and authority scores computed for each web page indicate the extent to which the web page serves as a “hub” pointing to good “authority” pages or as an “authority” on a topic pointed to by good hubs. The hub and authority scores for a page are not based on a single formula, but are computed for a set of pages related to a topic using an iterative procedure called Hyperlinks Induced Topic Search (HITS) algorithm. Figure 1.7(c) depicts a high level view of this model. Figure 1.7 (d) shows the web communities by using maximal flow model.
Figure 1.7: Key Models Developed in Web Mining

(a) Page Rank

\[ PR(P) = \frac{d}{N} + (1 - d) \left( \frac{PR(P_1)}{OutDeg(P_1)} + \frac{PR(P_2)}{OutDeg(P_2)} + \frac{PR(P_3)}{OutDeg(P_3)} \right) \]

(b) Hubs and Authorities

(c) Information Scent

(d) Maximal Flow Model for Web Communities
1.7 WEB LOG PREPROCESSING

Web log preprocessing involves a great deal of work including web log refining, users identification, conversation recognition and path added (optional). Data refining deletes mainly the useless web logs and irrelevant data. User identification refers to a process of labeling users with their visiting pages web logs. According to the IP address and user agent, guests will be classified consequently. Due to the existence of cache, proxy server (including cafe, etc) and firewall network, this step could be very complicated and time consuming; scholars have put forward some heuristic rules to identify users: (1) different IP address represents with different user. (2) When the IP address is as same as the others’, the defaults of different operating systems or browser represent different users. (3) With the same IP address and operating system and also the same browser, judging whether there is a direct link between the requiring page and all the pages visited previously, if so, then there is only one user, if not, then different users.

In order to solve the identification problem, currently there are many methods, such as using cookies, embedding user ID and client software agent, making register for use, etc. But in reality users may close cookies considering security problem or register with false information because it is unwilling to reveal privacy. So only the analysis of such as IP address, agent to identify different customers is actually feasible. Session refers to a series of activities, when a user first log into the website till the user leaves it. The goal to identify a session is to get meaningful visiting sequence during specific time. The web log preprocessing may include the following steps:
- Data cleaning
- User identification
- Session identification
- Path completion
- Transaction identification

1.7.1 Data Cleaning

According to the purposes of dissimilar mining applications, unrelated records in web access log will be eliminated throughout data cleaning is shown in Figure 1.8. Since the goal of web usage mining is to get the user’s travel patterns, following two kinds of records are avoidable and should be removed:

Figure 1.8: Data Cleaning in Web Mining
1. The records of graphics, videos and the format information

The records have filename extension of GIF, JPEG, CSS and so on, which can be found in the URI field of the every record, can be removed. This extension files are not actually the user interested web page, rather it is just the documents embedded in the web page. So it is not necessary to include in identifying the user interested web pages. This cleaning process helps in discarding unnecessary evaluation and also helps in fast identification of user interested patterns.

2. The records with the failed HTTP status code

By investigating the position fields of each evidence in the web access log, the records with status codes over 299 or below 200 are unconcerned. It should be pointed out that dissimilar from most other researches, records having value of POST or HEAD in the technique field are reserved in present study to acquire more exact referrer information.

1.7.2 User and Session Identification

The job of client and session classification as shown in Figure 1.9 is to find out the dissimilar user sessions from the original web access log. The difficulties to achieve these steps are given by using proxy server, e.g. dissimilar users may have similar IP address in the log. A referrer-based method is proposed to solve these problems in this study.
Figure 1.9: User and Session identification

The rules adopted to distinguish user sessions can be described as follows:

1. The different IP addresses differentiate the different users.

2. If the IP addresses are same, the different browsers and operating systems indicate different users.

3. If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URL field is checked, and a new user session is identified if the URL in the Refer URL field has not been accessed earlier, or there is a
huge distance (usually more than 10 seconds) between the accessing instance of this evidence and the previous one if the Referral URL field is empty.

4. The discussion known by rule 3 may contain more than one visit by the same user at different time, the time-oriented heuristics is then used to separate the dissimilar visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm must be used for acquiring the whole user contact path. The method of path completion will be completely discussed in the next section.

1.7.3 Path Completion

Path completion is the procedure of adding the page accesses that are not in the weblog but that have really occurred. If a user uses an amount of pages to reach to the final page, then the last page before the final page becomes the source page and it is referred to as the referrer area.

A different complicated step in data preprocessing is path completion. There are some reasons that outcome in path's incompletion, for example, local cache, agent cache, "post" method and browser's "back" button can outcome in some essential accesses not recorded in the access log file, and the number of URL recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy server’s caching without leaving any record in server's access log. As a result, the user access paths are partly preserved in the web access log. To realize user's travel model, the missing pages in the user access path should be appended. The reason of the path completion is to finish this task. The better results of data preprocessing, and will get better the mined patterns worth and save algorithm's running time. It is particularly
essential to web log files, in respect that the construction of web log files are not the similar as the data in database or data warehouse. This work is not ordered and complete due to different causations. So it is particularly compulsory to preprocess web log files in web usage mining. Through data preprocessing, web log can be changed into another data structure, which is easy to be mined.

1.7.4 Transaction Identification

A most important operation detection technique as shown in Figure 1.10 is to find the session prior to the path Maximal Forward References (MFR), each MFR is a transaction. MFR is defined as a group prior to the browsed page. The request page is not the visited page; “back” refers to the accessed page in the past of the client conference prior to the stay. There will be innovative pages additional to the traversal path, while “back” does not increase to the user's contact records.

Figure 1.10: Transaction Identification
1.8 COMMON WEB ACCESS LOG FORMAT

The Apache log format is used in this research, e.g. one record in web access log is written as: 219.144.222.253 - - [16/Aug/2004:15:36:11 +0800] “GET /images/1 r3 c2.jpg HTTP/1.1” 200 418“http:// 202.117.16.119:8089 /index.html” “Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)” and each field of the log is explained in Table 1.1.

Table 1.1 Web Access Log Format

<table>
<thead>
<tr>
<th>Field</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>219.144.222.253</td>
<td>Users’ IP address (UIP)</td>
</tr>
<tr>
<td>[16/Aug/2004...</td>
<td>The date and time of the request (Date)</td>
</tr>
<tr>
<td>GET</td>
<td>The method of the request (Method)</td>
</tr>
<tr>
<td>/images/1 r3...</td>
<td>The URL of the current request (URI)</td>
</tr>
<tr>
<td>HTTP/1.1</td>
<td>The version of transport protocol (Version)</td>
</tr>
<tr>
<td>200</td>
<td>The HTTP status code returned to the client (Status)</td>
</tr>
<tr>
<td>418</td>
<td>The content-length of the page transferred (Bytes)</td>
</tr>
<tr>
<td><a href="http://202.11">http://202.11</a>...</td>
<td>The URL requested just before (Refer URI)</td>
</tr>
<tr>
<td>Mozilla/4.0 (...)</td>
<td>Browser &amp; OS (Browser OS)</td>
</tr>
</tbody>
</table>

1.8.1 Extended Common Log Format (ECLF)

The following Table 1.2 describes how to configure its logging capabilities, and how to understand what the logs contain.
Table 1.2 Extended Common Log Format (ECLF)

<table>
<thead>
<tr>
<th>Extended Common Log Format ECLF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Remote host domain</strong></td>
</tr>
<tr>
<td><strong>rfc931domain</strong></td>
</tr>
<tr>
<td><strong>Authorized user domain</strong></td>
</tr>
<tr>
<td><strong>Date domain</strong></td>
</tr>
<tr>
<td><strong>Request domain</strong></td>
</tr>
<tr>
<td><strong>Status domain</strong></td>
</tr>
<tr>
<td><strong>Bytes domain</strong></td>
</tr>
<tr>
<td><strong>Referrer domain</strong></td>
</tr>
<tr>
<td><strong>User Agent domain</strong></td>
</tr>
</tbody>
</table>

1.9 USER SESSION IDENTIFICATION BY LOG DATA PREPROCESSING

User session identification is the method of analyzing procedure data in order to take out valuable information concerning user navigational actions by structuring the requests contained in the web log files. Access log files of a website consist of manuscript files where the server stores all the accesses made by the users in
chronological order. According to the Common Log Format, every log access includes: the user’s IP address, the ask for date and time, the demand method, the URL of the accessed page, the data transmission protocol, the revisit code representing the status of the demand, the size of the visited page in terms of number of bytes transmitted. Based on such information, this work can begin to identify the user sessions, i.e. the sequence of URLs that each user has accessed all through their visit. To identify user sessions, and used LODAP (Log Data Preprocessor), a software tool that previously presented. LODAP performs this task into three principal steps: data cleaning, data structuration and data filtering. Precisely, web log data are cleaned from useless and irrelevant information in order to retain only the entries corresponding to the effective user requests. Mainly, requirements with an contact process different from “GET”, failed and corrupt requests, requests for multimedia objects (such as images, videos, sounds, etc.), visits made by web robots are removed from log files. Next, log entries are structured into user sessions. Here, a user session is defined as the finite set of URLs accessed by a user within a predefined time period (in this work, 25 minutes). While the information regarding the user login is not presented, user sessions are well-known by combining the requirements originating from the same IP address throughout the traditional time period. Finally, data are filtered in order to maintain only the most related pages and user sessions. At the end of preprocessing, then obtain a collection of ns sessions denoted by the set S= <s_1, s_2,...,s_{ns}>. Every session contains information regarding accesses to pages in the session time. Precisely, a user session is formally described as a triple s_i= <u_i, t_i, p_i> where u_i represents the user identifier, t_i is the access time of the whole session, p_i is the set of all pages (with corresponding access information) requested during the i^{th} session. Namely:
\[ P_i = \langle (p_{ij}, t_{ij}, N_{ij}), (p_{ij}, t_{ij}, N_{ij}), \ldots \rangle \]

With \( p_{ij}, P \), where \( N_{ij} \) is the number of accesses to page \( p_{ij} \) during the \( i^{th} \) session and \( t_{ij} \) is the total time spent by the user on that page during the \( i^{th} \) session. Data preparation is typically conducted in a given order. The order may vary slightly because of private preference and the environment of available computer programs.

**1.10 WEB SESSION CLUSTERING**

The session parallel technique described in the preceding section can be useful to compute the decision between each pair of sessions, and build a comparison matrix. Suitable clustering algorithms are useful to this similarity matrix to locate the session clusters. An essential issue is how to estimate the worth of clusters in the result. Clustering confirmation is a field where attempts have been made to get rules for quantifying the worth of a clustering result. This issue, however, is a complicated one and usually people assess clustering results visually or compare to known automatically clustered data.

Visually inspecting cluster for unconditional data such as web session data is inflexible and has not been done. The method to envision comparison of similarity and dissimilarity between clusters. For this ordered the resulting clusters according to their descending sizes on two axes of a 3 dimensional graphs. Sessions within clusters are also controlled with the same ordering on both axes. The third axis just represents the level of comparison between sessions. Following Figure 1.11 shows an unrealistic example with 1000 sessions in 3 clusters. In this idealistic case where the cross parallel between each pair of sessions within a same cluster is 1:0, and cross similarity between each pair of sessions from two different clusters is 0:0, this can see that only the diagonal has some values on the similarity dimension. The presence of the diagonal indicates good
clustering while high similarity values outside the diagonal would indicate inadequate clustering. This testing session set used in experiments have 1000 arbitrarily selected sessions from a tangible e-learning system.

Both Jaccard similarity and dynamic programming based similarity methods were used to provide comparison matrices for the given session set. ROCK, CHAMELEON and TURN were applied on the comparison matrices to each constructed clustering result.

From the clustering results found that ROCK tends to find bigger cluster with lower average similarity. CHAMELEON and TURN can find cluster with high internal cross similarity. The difference between the two is that TURN can identify outliers while CHAMELEON cannot. Rare sessions dissimilar to most other sessions are identified by TURN, while CHAMELEON forces them to belong to a given cluster. Using the Jaccard Coefficient as a similarity measure for sessions tends to give more clusters than dynamic programming based similarity measure. In general when evaluated manually, the cluster

**Figure 1.11: Session clustering visualization**

From the clustering results found that ROCK tends to find bigger cluster with lower average similarity. CHAMELEON and TURN can find cluster with high internal cross similarity. The difference between the two is that TURN can identify outliers while CHAMELEON cannot. Rare sessions dissimilar to most other sessions are identified by TURN, while CHAMELEON forces them to belong to a given cluster. Using the Jaccard Coefficient as a similarity measure for sessions tends to give more clusters than dynamic programming based similarity measure. In general when evaluated manually, the cluster
quality between clusters using the dynamic programming based correspondence measure was better than when using the Jaccard Coefficient similarity measure.

The clusters were just additional important, which an estimated result is since it took in reflection the series of clicks in a session. However, this work do not presently have the means to compute quantitatively this cluster quality, and it would be very complicated to automatically assess and evaluate the quality of the clusters resulting from the dissimilar parallel measures when the dataset is very huge. Nevertheless, the technique scales well with the size of the dataset to group, and self-assured, given opening tests with the 1000 session set, that the web session clustering with series alignment would always yield more important results than the commonly used approximation of sessions with sets.

1.11 EXPOSE THE VARIOUS TECHNIQUES AVAILABLE FOR WEB USAGE MINING

Web mining for usage pattern is the key to discover marketing intelligence in e-commerce. It helps tracking of general access pattern, personalization of web link or web content and customizing adaptive sites. It can disclose the properties and inter-relationship between potential customers, users and markets, so as to improve web performance, on-line promotion and personalization activities. Web Log Mining uses KDD techniques to understand general access patterns and trends to shed light on better structure and grouping of resource providers. For e.g., Web miner discovers association rules and sequential patterns automatically from server access logs. Clementine and DB2 Intelligent Miner for Data are two general-purpose data mining tools, which can be used for web usage mining with suitable data pre-processing.
1.12 SESSION CLUSTERING MODEL CONCEPT

Web usage mining inspects the direction finding patterns in web access logs and extracts earlier indefinite and useful information. This may lead to strategy for different web oriented applications like website reorganize, recommender system, web page calculation and so on. The current work demonstrates clustering of user sessions of uneven lengths to notice the access pattern by proposing an expanse method to group user sessions is shown in Figure 1.12.

- Session storage: wherever sessions are stored, this can be either memory, or disk-based storage space with a tunable memory gathering.

- Zend Session Manager (SCD): the Session cluster daemon, that transfers sessions from session storage to the PHP engine and from isolated nodes

- mod_cluster: the PHP session manager that communicate with the session clustering daemon

![Figure 1.12: Session Clustering](image.png)
The session clustering unit employs burly locking and data consistency mechanism to make sure that sessions are not at all degraded. Session clustering is also tunable with two dissimilar session storage space model: write through or delayed write, allowing organizations to decide their favored storage space model. The session clustering module can be incorporated into any existing PHP relevance that uses PHP’s native session expansion - without shifting any code. Session clustering implements a local PHP session unit, and switch among the existing resolution and session clustering is simply a matter of changing a php.ini file directive.

1.13 MOTIVATION

Rapid growth of information sources available on the World Wide Web has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses. A web site is the most direct link a company has to its current and potential customers. In web log data preprocessing is a difficult procedure and also takes 80% of total mining process. Log data is pretreated to get consistent data. The aim of data preprocessing is to choose important features from data by removing inappropriate records and finally transform raw data into sessions. By stimulating the web robot for correct detection of user attracted patterns through supply of the relevant web logs. Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining. Clustering of documents, web log data, customer data in marketing or other applications are very essential and makes system easy for their further proceedings. The main motivation of this research is to achieve good clustering accuracy. Then the execution time is to be very low in the proposed system.
1.14 PROBLEM SPECIFICATION

Presently there are varieties of problems linked with the existing web usage mining approaches. Obtainable web usage mining algorithms bear from problem of practical applicability. So, a novel research is very much required for the perfect calculation of future performance of web users with hasty execution time. Web usage mining consists of preprocessing, model discovery and model analysis. Log data is normally noisy and unclear, so preprocessing is an essential process for successful mining process. In addition, endurance of plentiful data in the network and the unstable and various nature of the web, web searching has become a tricky procedure for the majority of the users. This makes the users feel puzzled and at time lost in stuffed data that persist to increase.

1.15 OBJECTIVES OF THIS RESEARCH

Web usage mining improves the web interaction drastically. It has the following main objectives for any application.

- To optimize user’s web surfing experience and to know the users attention and navigational clustering performance of transactions.
- To improve the performance of a website by the elimination of Local and Global Noise.
- To find associations among web pages that frequently appears together in users’ sessions.
To improve the design of a website and eliminating web robot-generated record entry not only simplifies the withdrawal task that will follow, but it also removes unexciting sessions from the log file.

To provide faster and easier data processing and it also helps in saving time and it resources.

To make recommendation, personalize Web sites, and for other uses such as targeting users for advertising.

To preprocess the data and divide the web users into clusters that can be used to classify future web users.

1.16 SCOPE OF THIS RESEARCH

Different from other implementations records are cleaned effectively by removing local, global noise and robot entries. An efficient web usage mining can be developed for enhancing web design, enhancing satisfaction of customer, marketing analysis, target advertisements and increasing the competitive strength of enterprises etc. The preprocessing of web log data is to reorganize and correct original data in web access log for preparing the reliable stuff for web usage mining calculations. Dividing the web users into groups that can be used to classify future web users. The fast and efficient web mining can able to process large web log datasets with complex structures too. The best hope is to identify appropriate measure to session the data, due to the fact that cache server is used to access the most recent page request by the client. The plan of Fuzzy Cluster is to determine cluster centers that reduce a difference function.
1.17 ORGANIZATION OF THE THESIS

The thesis is organized as follows.

Chapter 1 deals with the introduction of web mining and types and techniques used in web mining.

Chapter 2 reviews the previous work done and literature survey in the areas of web mining.

Chapter 3 describes the first method work for *Novel Preprocessing Technique for Web Log Mining by Removing Global Noise and Web Robots*.

Chapter 4 discusses the second approach to improve the clustering accuracy for *An Effective Web User Analysis Using Fuzzy Clustering*.

Chapter 5 describes the third technique proposed to obtain efficient result for *Hybrid Clustering for Web User Analysis Using Proposed Rough and Fuzzy Possibilistic C-Means*.

Chapter 6 presents the findings and discussion of the proposed methodologies. The performance of the proposed approach is evaluated based on the web mining.

Chapter 7 provides the concluding remarks and the future enhancements of the proposed approaches.

The works of several researchers are quoted and used as evidence to support the concepts explained in the thesis. All such evidences used are listed in the reference section of the thesis.
1.18 SUMMARY

This chapter completely deals with the brief explanation of web mining and techniques, also provides the detailed overview of the web session clustering techniques and their algorithms. The process of web log pre-processes and web session clustering are discussed. The data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the unsuccessful HTTP status code and eventually robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries are discussed. It also describes the web log mining for removing noise.