# A Survey on Methods used in Web Usage Mining

**Rajinder Singh Rao[1], Jyoti Arora[2]**

[1] Student, Dept. Of Computer and Science Engineering, DBU, Punjab, India
[2]Assistant Professor, Dept. Of Computer and Science Engineering, DBU, Punjab, India

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *In recent years the growth of the World Wide Web exceeded all expectations (with the development of Internet technology). Today a lot of information is available in different formats; there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But retrieving interesting content from web has become a very difficult task. So in order to retrieve required information from the www mining of the web is an important task. Web usage mining (WUM) is the process of the retrieving useful information/knowledge from the server logs. Server logs contain irrelevant data which does not contribute towards extracting useful information, so these log files requires pre-processing. Then from the preprocessed files different patterns are required to be discovered in order to comprehend the behavior of the users. The found patterns required to be analyzed to form useful knowledge. The knowledge obtained from web usage mining can be used to enhance web design, introduce personalization service and facilitate more effective browsing.The various applications of web usage mining are: robots detection and removal, extracting user profiles, recommendation systems, Personalization of Web Content, Prefetching and Caching, Ecommerce etc. Web usage mining is an effective technique to extract knowledge from the unstructured data. With the help of web log data the required data can be sorted out and one can judge its popularity by deriving the interested and not interested ones. The objective of this paper is to provide a review of web usage mining.*

***Key Words*: Web Usage Mining, log files, Server Logs, Pattern Discoveries, Data Cleaning.**

## 1. INTRODUCTION

Web usage mining also known as web log mining. Web Usage Mining mines the log data stored out in the web server. Enormous development in World Wide Web enlarges the complexity for users to browse it successfully. To increase the performance of web sites better web site design, web server activities are shifted to as per users' interests. The capability to know the patterns of users' habits and interests helps the operational strategies of enterprises.

Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc.

Web mining can also be termed as the integration of the knowledge collected by traditional data mining methods and techniques with knowledge related to the web.

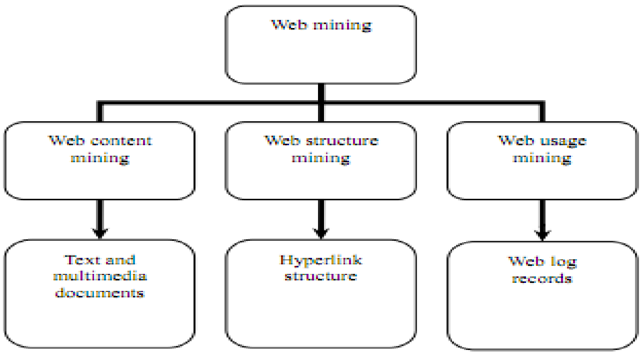
Two different approaches were taken in initially defining Web mining. First one was a 'process-centric view', which terms Web mining as a sequence of tasks. Second one was a 'data-centric view', in which Web mining is defined in terms of the types of Web data that was being used in the mining process.

There are three general classes of knowledge which can be discovered by web mining:

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

## 2. WEB MINING TYPES

Web is a collection of inter-related files on one or more Web servers. Web mining enables one to find out web pages, text documents, multimedia files, still images and other types of resources from the web.Web Mining is broadly divided into three categories as: web content, web structure and web usage mining.



Web Content Mining is the process of attain useful knowledge from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the customer. It may consist of text, still images, audio, moving images, or structured records such as lists and tables.

Web structure mining uses graph theory to analyze the node and connection structure of a web site. The aim for structure mining is to extract previously unknown relationships between Web pages. The goal of the Web Structure Mining is to generate the structural abstract about the Web site and Web page. Web Structure mining will categorize the Web pages and provide the information like similarity and relationship between different Web sites. This mining technique is performed either at the intra-page (document) level or at the inter-page (hyperlink) level.

Web usage mining is a process to keep an eye over the user findings on the internet whether they are looking for textual data or they are interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications.

## 3. WEB USAGE MINING

With the increasing demand of internet more number of websites is being involved for getting required information and thus more usage of web-based data. The data which is stored in different format types in the web log file. This log file should be maintained as these data are in unsorted manner and it is done through preprocessing. Web usage mining focuses on discovering useful information. Web log file is automatically generated by web server whenever user accesses the resource like webpage of website.

Web usage mining is the process of withdrawing the useful knowledge from the server logs. It is the application of data mining techniques to discover interesting usage patterns from Web data in order to comprehend and better serve the requirements of the Web-based applications. Web usage data note down the identity of the user and their browsing behavior at a particular Web site. Usage data can be documented in the form of log files. A Web log is a file in which the server takes the knowledge/data each time a user requests a site from a particular server. A log file can be placed in three different locations i.e. web servers, web proxy server, user's browser.

- Web Server Log files

The log file that resides in the web server notes the activity of the client who accesses the web server for a web site through the browser.

- Web Proxy Server Log files

It's the intermediate server (medium of interaction) that exists between the client and Web server. Therefore if the Web server gets a request of the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers keep a separate log file for gathering the information of the user.

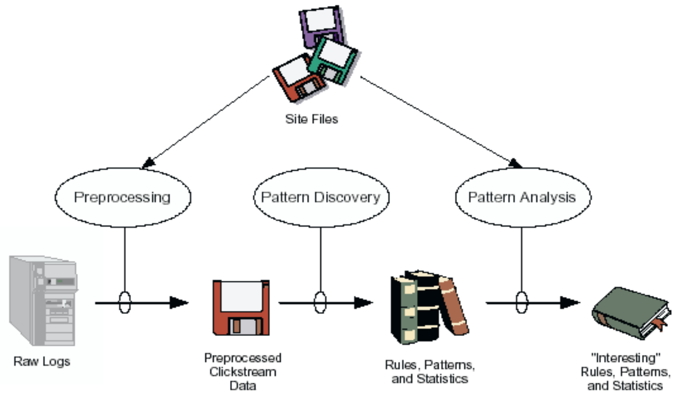
- Client/User Browsers Log files

These log files can be made to reside in the client's browser window itself. A number of software's are there that can be downloaded by the user to their browser window. Even then the log file is present in the client's browser window, the entries to the log file is done only by the Web server.

## 4. WEB USAGE PHASES

Web Usage Mining consists of four basic steps, Data Collection, Data Preprocessing, Pattern Discovery and Pattern Analysis.

**1 Data Collection**: This is the first step in which user's log data is collected from various sources. This includes only the relevant data that is to be collected. Data source can be gathered at the server-side, client-side, proxy servers, or obtain from an enterprise's database, which contains business data or consolidated Web data.


Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

## 2 Data Preprocessing:

Some databases are insufficient, inconsistent and including noise. The data pre-treatment is to carry on a unification transformation to those databases and the database will become integrate and consistent, thus results the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion. Basically, Data Preprocessing extracts text format data form log file and store clean data into database.

**Data Cleaning:** It removes the irrelevant and redundant log entries and it also corrects or removes corrupt records from the database files. There are three kinds of inapt or redundant data to be removed.

**User and Session Identification:** Its task is to find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The aim of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The troublesome to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log file.

**Path Completion:** Another critical step in data preprocessing is path completion. Thereare some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of URL's recorded in log may be less than the real one. Using the local caching and proxy servers also provide the drawback for path completion since users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log, in reaction the user access paths are incompletely sustained in the web access log. To uncover user's travel pattern, the lost pages in the user access path should be included. Alike with user identification, the heuristic presume that if a page is requested that is not directly linked to the previous page accessed by the same user, the referrer log can be referred to see from which page the request has came. If the page is in the user's recent click history, it is presumed that the user browsed back with the "back" button, using cached sessions of the pages.

## 3 Pattern Discoveries

After the change of the data in the log file into a formatted data, the pattern discovery process is under gone. Pattern Discovery Tools apply techniques from data mining, machine learning, statistics and pattern recognition etc. In other words,

Pattern Discovery finds pattern, Classify data by applying mining techniques.

**Association rule:** Association rule learning is a process in which we search for relationships between variables. In web usage mining, association rules are used to find out which pages are frequently visited together in a single server session with a view to discover which websites and which sectors are frequently visited together. These pages may not be directly linked to one another via hyperlinks.

**Sequential pattern mining:** This finds sets of data items that occur together frequently in some sequences. Sequential pattern mining, which extracts frequent subsequences from a sequence database, has attracted a great deal of interest during the recent data mining research because it is the basis of many applications, such as: web user analysis, stock trend prediction, DNA sequence analysis, etc.

**Clustering:** In this, without using the known structures in the data we discover groups and structures in the data that are in some way or another similar. It is a technique to group users in clusters based on their common characteristics like browsing pattern, keyword selection etc. Clustering algorithms learn in an unsupervised way wherein their own classes and subsets of related objects in the training set are discovered. Clustering of web pages helps in grouping with related content, information much useful for internet search engines and web assistance providers.

**Classification:** The main objective of classification in web domain is to develop a profile of users belonging to a particular class or category. Contrary to clustering, classification is a supervised way of learning wherein the data items are mapped into one of the several predefined classes. It can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifiers, support vector machines etc.

**4 Pattern Analyses**

This is the final stage of Web Usage Analysis. Pattern analysis finds knowledge from the discovered pattern Of the interesting patterns by eliminating the irrelevant patterns. Pattern Analysis involves the validation and interpretation of the mined patterns. Validation can be used to remove the irrelevant patterns and to extract the interesting patterns from the output of the pattern discovery process. The output result is in mathematic form which is not suitable for direct human interpretations. So, Visualization techniques are used to interpret the results. The most general ways of analyzing user access patterns are either by using a knowledge query mechanism on a database such as SQL or data cubes to perform OLAP operations. Visualization techniques, such as graphing patterns are used for an easier interpretation of the results.

## 5. WEB USAGE MINING ADVANTAGES AND DISADVANTAGES

**Advantages**

Web usage mining has many advantages which makes it more appealing to enterprises including the government agencies.

- This technology has started the e-commerce to do personalized marketing, which eventually turn out to great in trade volumes.
- Government agencies are using this technology to categorize threats and fight against terrorism.
- The predicting capability of mining applications can be useful for society by recognising criminal activities.
- The companies can establish better customer relationship by proving them exactly what they need.
- Companies can understand the requirements of the customer better and they can respond to customer needs faster.
- The companies can find, attract and retain customers; they can also save the production costs by using the acquired insight of customer needs.
- They can make more profit by target pricing based on the profiles created.

**Disadvantages**

This Web Usage Mining when used on data of personal results to some concerns:

- The Main issue with web usage mining is the invasion of Privacy. Privacy is lost when information concerning a User/individual is obtained, used, or diffused, especially if this occurs without their knowledge or consent.
- Another Major concern is that the companies collecting the data for some specific purpose but might they use the data for some totally different purpose and this essentially violates the user's interests.
- Some techniques may use controversial attributes like sex, race, religion, or sexual orientation to classify individuals. These practices might be against the anti-discrimination Law. The applications make it hard to detect the use of such controversial attributes, and there is no hard rule against the usage of such techniques/algorithms with such attributes. This task could end up into a privilege to an individual based on his race, religion or sexual orientation.
- The growing trend of selling personal data as a commodity encourages website owners to trade personal data obtained from their site. This trend has increased the amount of data being captured and traded grow up the likeliness of one's privacy being violated. The companies which buy the data make it anonymous and these companies are considered authors of any specific release of mining patterns.

## 6. CONCLUSIONS

Data mining is the study of exploring patterns in huge volumes of raw data. The term Web mining has been used to refer to techniques that help us to find content of web and retrieve the user's interest and needs. This paper focuses on the comparison of mining algorithms. In this survey paper all the main Phases of web usage mining have been addressed. These include Data Collection, Data Preprocessing, Pattern Discoveries and Pattern Analyses .The literature consulted so far reveals that the existing recommendation systems are capable of giving recommendations to a user on the basis of the only that user's discovered pattern. In the first step, the existing systems, the preprocessing operations are performed on the web log in order to remove the useless data from it that id gathered through data collection technique and to reduce its size. In the next step, the resulted web log is used to discover usage patterns. In the last Step, On the basis of the discovered patterns of the particular user or single user session the recommendations are suggested. The existing systems are not capable of recommending on the basis of the discovered patterns from all the users log history. Also in the existing system, the patterns are discovered by using complex methods and the discovered patterns of every user are too complex for the server administration to Analyze.

## ACKNOWLEDGEMENT

Firstly I thank Almighty God for the blessing endowed onto me. I wish to express my sincere thanks to my guide Ms .Jyoti Arora, (Assistant Professor of Computer and Science Engineering department, Desh Bhagat University) and other Faculty members for their guidance and encouragement in carrying out this Project.

## REFERENCES

[1] Modi H.Y and Narvekar M, "Enhancement of Online Web Recommendation System Using a Hybrid Clustering and Pattern Matching Approach", (ICNTE-2015).

[2] Bhargav A and Bhargav M, "Pattern Discovery and Users Classification Through Web Usage Mining", International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT-2014).

[3] Gupta A, Arora R, Sikarwar R and Saxena N. "Web Usage Mining Using Improved Frequent Pattern Tree Algorithms", International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT-2014).

[4] R. Thiyagarajan and K. Thangavel, "Usage Profile Based Recommendation System", International Conference on Intelligent Computing Applications (2014).

[5] Sharma, Manohar M and Bala A. "An Approach for Frequent Access Pattern Identification in Web Usage Mining", International Conference in Advances in Computing, Communications and Informatics (ICACCI-2014).

[6] Sha H and Liu T, "EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", (ITQM-2013).

[7] Yogish H K and G. T. Raju, "Clustering of Preprocessed Web Usage Data Using ART1 Neural Network and Comparative Analysis of ART1, K-means and SOM Clustering Techniques", 5th International Conference on Computational Intelligence and Communication Networks (2013).

[8] Singh N and Jain A, "Comparison Analysis of Web Usage Mining Using Pattern Recognition Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.3, No.4, July 2013.

[9] V. Sujatha and Punithavalli, "Improved User Navigation Pattern Prediction Technique From Web Log Data", International Conference on Communication Technology and System Design (2011).

[10] P.Nithya and Dr.P.Sumathi, "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots", (NCCCS-2012)

[11] Dario M, Sarria D and Leon-Guzman E, "A recommendation-based web usage mining model for a university community", Eighth Latin American Web Congress (2012).

[12] Aye T.T, "Web Log Cleaning for Mining of Web Usage Patterns", IEEE 2011.

[13] Sneha Y.S and Prakash M, "An Online Recommendation System Based On Web Usage Mining and Semantic Web Using LCS Algorithm", IEEE 2011.

[14] Zhang Y, Dai L and Zhou Z.J, "A New Perspective Of Web Usage Mining: Using Enterprise Proxy Log",International Conference on Web Information Systems and Mining (2011).

[15] V.Chitraa and Dr. Antony Selvdoss Davamani, "A Survey on Preprocessing Methods for Web Usage Data", (IJCSIS), Vol. 7, No. 3, 2010.