ISSN:0975-9646



# Web Log Based Analysis of User's Browsing Behavior

**Ashwini ladekar**
*B.E Student, Department of Computer,*
*JSPM's BSIOTR ,Wagholi,Pune, India*

**Pooja Pawar**
*B.E Student, Department of Computer,*
*JSPM's BSIOTR ,Wagholi,Pune, India*

**Dhanashree Raikar**
*B.E Student, Department of Computer,*
*JSPM's BSIOTR ,Wagholi,Pune, India*

**Prof.Jayashree Chaudhari**
*Department of Computer,*
*JSPM's BSIOTR,Wagholi,Pune,India*

***Abstract*— The paper discusses about browsing behavior of the user and the user's interest, the technology of web mining, the information source from the system i.e. web log data and the apriori algorithm which is used. Web graph is generated by incorporating the behavior of user's browsing. The privacy is therefore not affected as the user's feedback is not collected. The paper also tells us how user's estimation of behavior is done based on studying web logs.**

***Index Terms*—Web log data, apriori algorithm, data mining, page interest estimation.**

## I. INTRODUCTION

With ample amounts of information present on the World Wide Web (WWW), issues relating to acquiring useful data from the Web has mounted the attention among researchers in the field of knowledge discovering and mining of data. In today's agonistically business environment, Web services have become an implicit need for the organizations for discovering patterns. Knowledge acquired from the data which is collected helps in developing strategies for business. To create faithful customers and gain militant advantage organizations are implementing value added services. By providing personalized products and services, the companies are creating long-term relationships with users. By focusing on each individuals need this type of personalization can be achieved. Web mining helps to retrieve such know-ledge for personalization and improved Web services. Web mining pertains to Knowledge Discovery in Data (KDD) from the web. That means it is the process of application of data mining technique to retrieve useful information from immense amount of data available from web. Web mining and data mining objective being same both try searching for variable and useful information from web log and databases.

In retrieval of data from the Web, the personalized search carried out by a particular user forms an important research for personalized search engine. Commonly, there are two ways to collect user interest. The first approach is to take a feedback from the user in terms of his interest level. But all the user's are not interested to give the feedback so this approach is a bit inconvenient so we use the second approach of user interest based on his browsing behavior. Without users knowledge, the degree of his interest is evaluated. The second approach has be-come one of the important approach for collecting the interest of the user. This paper deals with web usage mining and analyzing the user's browsing behavior.

## II. WEB MINING

When we are surfing on the internet we come across data which is not completely useful as per our needs. Mining is the technique of excavation for discovering knowledge. The process of extraction of knowledge from data is known as data mining. While the process of extracting information and data and patterns from web is known as web mining. Three methods as shown in fig. 1 are relevant for web mining-

*A. Web Content Mining*

*B. Web Structure Mining*

*C. Web Usage Mining*

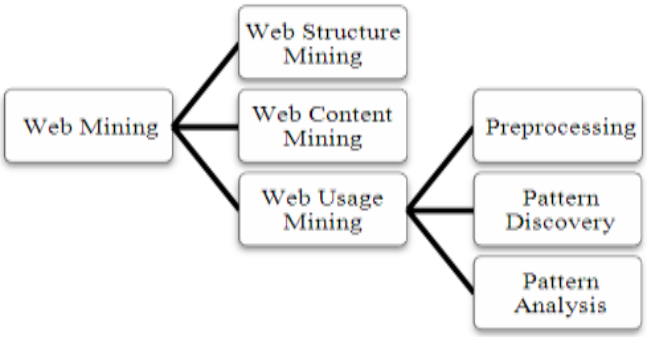


Fig 1. Classification of Web Mining

## III. WEB USAGE MINING

By determining frequent access behavior for a user the web usage mining can be used to make search significant, the future access can be improved by identifying the needed links. Web usage mining is the purpose of data mining for discovering usage patterns from Web data to recognize the user needs. Web usage mining is contains three parts namely preprocessing, pattern discovery and the analysis of the pattern. Web usage mining is mainly applied to the log files

ie. the data stored in the log files. The user requests various data and this is the information that is stored in the log file. Web usage mining is used to give recommendations to the visitors [9]. In web usage mining the main aim is to discover and retrieve attractive patterns from dataset. The web mining is the application of the data mining technique and it is used to retrieve information in web data, in which the structure/usage data is made use in the process of mining. The various application areas of web usage mining are web pre-fetching, reorganization of the site, personalization of the web and prediction of the link. Finding useful pattern by making use of the web log data is the most important phase of web usage mining[10].

*A. Web usage mining process :*

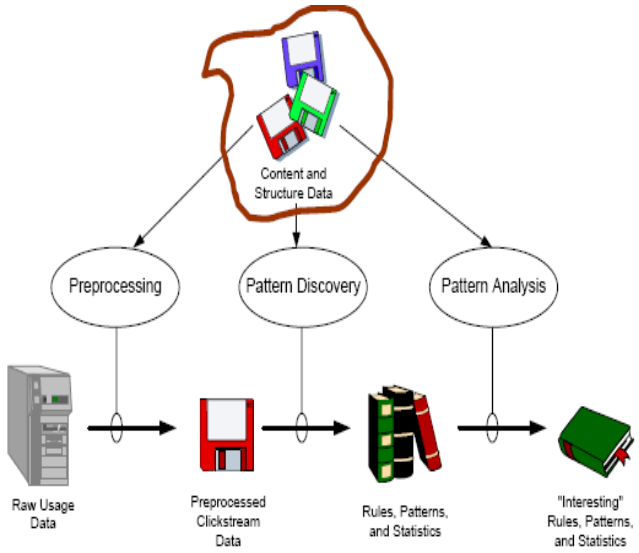


Fig 2. Web Usage Mining Process

*B. Hyperlink :*

The interaction between two pages is the hyperlink, like it may be within the same or different web page. There are two types of association present in hyperlink : one which connects the various parts of same page and it is called as Intra Document Hyperlink and the second is Inter Document Hyperlink which connects two different pages.

*C. Document Structure :*

The web page content is set in tree structured format, it is dependent on the HTML and XML within that particular page.

*D. Web Server Log Data :*

The web acts as an interface for extracting logical data. The data log needs to trail any transaction of conversations. It examines the users actions over a longer time. Server log offers technical data like the requests that are successful, marketing assistance and expansion of site and activities related to marketing. Below is an example of frequently collected transfer log. This is the NASA web server logs data.

EXAMPLE:

198.72.81.55 - [01/Jul/1995:00:00:01 -0400] "GET /history/apollo/ HTTP/1.0" 201 6245

unicomp6.unicomp.net - [01/Jul/1994:00:00:06 -0400] "GET /shuttle/countdown/ HTTP/2.0" 200 3985

198.130.110.21 - [01/Jul/1995:00:00:09 -04001] "GET /shuttle/missionss/sts-73/mission-sts-73.html HTTP/1.0" 210 4085

The server log consists of several attributes such as-

- **Date:** The date is the date of the transactionon which it was tracked. Eg : from above 1995-07-01.
- **Time:** It is the time of the transaction. The time format is as HH:MM:SS eg. From above 00:00:01.
- **Client IP Address:** Client IP is the computer who made request for the website.
- **User Authentication:** Few websites have an authentication facility provided to the users which requests the user to enter username and password.
- **Server IP Address:** The Internet Service Provider gives the Server IP address it is static. This IP is useful for accessing the information from the server.
- **Server Port**: This port is used for broadcasting the data.
- **Server Method (HTTP Request)**: The word requests reference to an image, document and more. The below instance indicates that folder00.gif was the item accessed.
- **URL:** URL is a pathway from the host. It is structure of the website. For instance:/tutor1/images1/icons1/folder00.gif.
- **Agent Log:** The Agent Log generates information on browser of the user, the browser version, and the operating system. This is the worthy data that stores data such as the browser type and the operating system and tells what a user is able to access on a website.

*E. Pattern Discovery and Pattern Analysis :*

The three phases of web usage mining are : Data preprocessing, Pattern discovery and Pattern analysis.

**Creation of Log file :**

The quality of patterns which is discovered in web usage mining process is highly dependent on quality of the information that is used in the process of mining [1]. When browser tracks the web pages and stores the log file of the server. Web usage data consists of data about the Internet addresses of various web users along with their navigational behavior [2].

**1. Web Server Data:**

Whenever an user agent hits a URL in the domain, the detail of data related to that particular operation is stored in a log file. The session information of all users can be obtained by preprocessing the web log data. Information stored in access log file at server side contains the session[2,5] data of various users. The records have seven common fields as follows :

- IP address of the user
- Accessed time and accessed date
- Request method (GET/ POST),
- URL of the page which was accessed
- Transfer protocol (HTTP 1.0, HTTP 1.1,),
- Success of return code.
- Number of bytes that were transmitted.

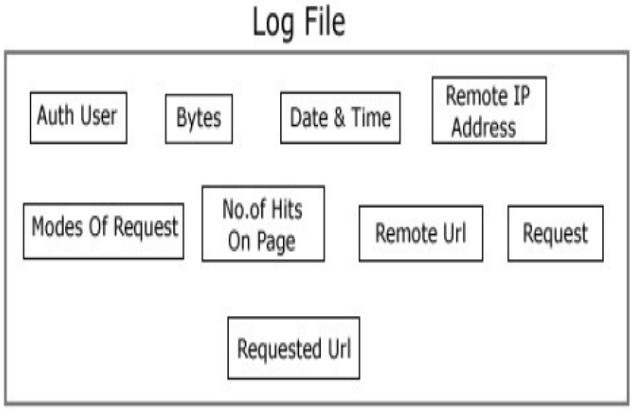


Fig 3. Log File Functional Diagram

**2. Preprocessing:**

Data Cleaning [6, 7], includes integrating the various usage logs, and parsing the data from usage logs. By the detection of files which has suffixes like text and hyperlink the data cleaning process can be performed. When selecting the proper algorithm for clustering the nature of information to be clustered plays an important role.

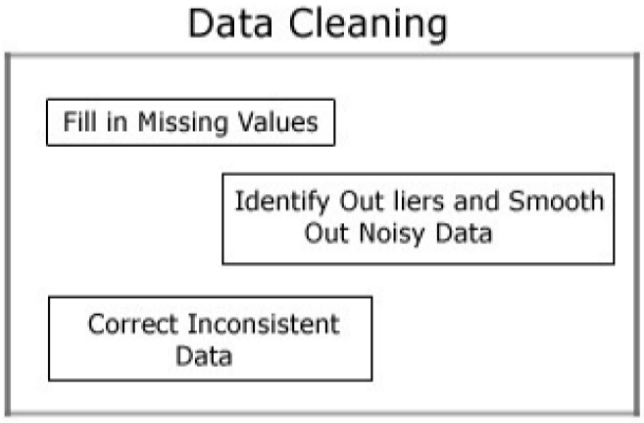


Fig 4. Data Cleaning Diagram

**3. Pattern Analysis Phase:**

Pattern discovery is the main issue in both web usage mining and data mining [6]. As the lengths of patterns to be discovered increases the search space increases exponentially. The discovered patterns need to be interpreted and logical information is extracted from it.

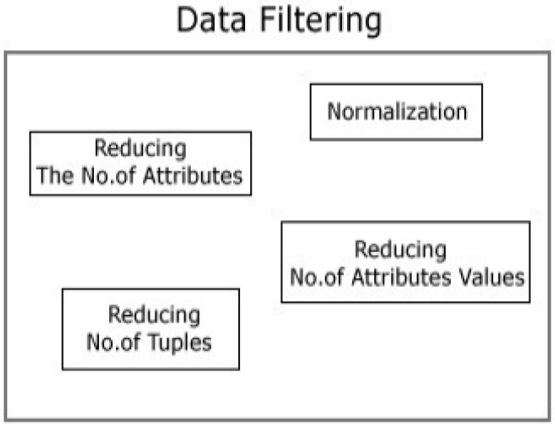


Fig 5. Data Filtering Diagram

**Location of a Log file:**

The web server writes information to the web log file each time a user requests a site from server. Various places where a log file is located are as follows:

- The Web Servers
- Web Proxy Server
- The Browser of the Client

Following figure depicts the example of raw log file. The raw log file consists of 546 file where each file consists of the information collected during one hour from the activities of the users in a Web store. Each row of log contains the following :

- IP address
- URL
- The time of request.
- The time of response.
- The time of difference.

The log file identification is done in sessions, the web page sequences the same sessions are already being identified in the log files, the same sessions are collected only in preprocessing step.

*D. Apriori Algorithm :*

In the field of mining of data and computer science, Apriori algorithm is an algorithm typically used for learning association rules. This algorithm functions on databases which consists of the transactions. It uses breadth first search and the tree structure for counting item sets of candidates. Length k candidate item sets are produced from item sets of length k-1. Pruning is done on uncommon sun pattern of candidates. The candidate consists of the frequent k-length item sets according to the downward lemma. Then the examination of transaction database is done for determining repeated item sets amongst the candidates. The key concepts of the algorithm are :-

- **Repeated Item sets**: The item set that has minimal support.
- **Apriori Property**: Subset of any item set should be frequent.

- **JoinOperation:** To find out Ak, set of candidate k-itemsets is generated by joining Ak-1 with itself.

The advantages of using Apriori algorithm are:-

- Makes use of large item set property.
- Simple for parallising.
- Simple to implement.

The Apriori algorithm is useful for finding the repeated item sets. The Apriori algorithm is as follows.

- *Ak*: Set of repeated item sets of size *k* (with minsupport)
- *Ck*: Set of candidate item set of size *k* (po-tentially repeated item sets)

*L1* = {repeated items};
**for**(*k* = 1; *Ak*!=; *k*++) **do**
*Ck+1* = candidates generated from*Ak*;
**for each** transaction *t* in database do
increase the count of all candidates in
*Ck+1* that are contained in *t*
*Ak+1* = candidates in *Ck+1* with min_support
**return***k Ak*;

**Proposed System and its flow :**

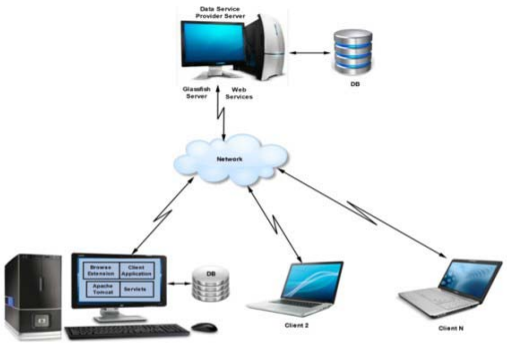


Fig 6. The proposed system

**Steps on client and server side :**

**Client:**

- The client machine's browser tracks user's behavior of browsing and the time details that are relates.
- The extension makes use of servlets for logging data on database(local)by making use of apache tomcat.
- The client application supports the feature of storing the browsing history(local), the filtering data and it applies user statistics to the cloud.
- The user can thus estimate its time that was needed to open a particular page or site.

**Server:**

- Server obtains log which consists of user's action from the client machines.
- Server applies mining algorithms so that better business intelligence solutions can be found out.
- Server application serves with a graphical analysis of user's usage patterns.

## IV. CONCLUSION

Estimating the interest of a user as he/she visits Web pages has gained an importance as Web-based activities have increased. With the tremenduous growth of World Wide Web, the study of modeling and predicting a user's access on a Web site has become more important. Web usage mining is an application of data mining technique to discover usage patterns from Web data. It helps to understand and serve the need of user. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests.

Web usage mining basically has three stages, namely preprocessing, pattern discovery, and pattern analysis. One of the algorithms which is very simple to use and easy to implement is the Apriori algorithm. Web usage mining refers to the automatic discovery and analysis of patterns in user access stream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site.

## ACKNOWLEDGMENT

We would like to express gratitude to our staff, family and friends for guiding us throughout this paper publication process and providing us with excellent support.

## REFERENCES

[1] Ai-Bo Song , Zuo-Peng Liang, Mao-Xian Zhao, Yi-Sheng Dong, Mining Web Log data based on Key path, Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002.

[2] R. Vaarandi, A Data Clustering Algorithm for Mining Patterns from Event Logs, in Proceedings of the 3rd IEEE Workshop on IP Operations and Management. Kansas City, MO, USA: IEEE Press, October2003, pp. 119 – 126.

[3] F. Masseglia, P. Poncelet, and M. Teisseire, Using data min-ing techniques on web access logs to dynamically improve hypertext structure. In ACM SigWeb Let-ters, 8(3): 13-19, 1999. Web Site Link: http://portal.acm.org/citation.cfm?id=951440.951443.

[4] V elasquez, Bassi J D, YasudaA. Mining Web data to create online navigation recommendations. Data Mining, 2004:166-172. Proceedings of the Fourth IEEE International Conference on Data Mining(ICDM'04) 0-7695-2142-8/04 IEEE.

[5] James H. Andrews, Member, IEEE, and Yingjun Zhang, General Test Result Checking with Log File Analysis , 0098-5589/03/ @ 2003 IEEE Published by the IEEE Computer Society.

[6] Junjie Chen and Wei Liu, Research for Web Usage Mining Model, International Conference on Computational Intelligence for Modelling Control and Automation,and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06) 0-7695-2731-0/06 ©2006 IEEE

[7] Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R.W., & Musen, M.A. (2001), Creating Semantic Web Contents with Protégé-2000. IEEE Intelligent Systems 16(2), 60-71.

[8] Mike Perkowitz, Oren Etzioni, Adaptive Web Sites: Automatically Synthesizing Web Pages, Department of Computer Science and Engineering, Box 352350 University of Washington, Seattle, WA 98195, 1998, American Association for Artificial Intelligence(www.aaai.org).

[9] Sanjay Kumar Malik, Nupur Prakash, S.A.M. Rizvi Ontology and Web Usage Mining towards an Intelligent Web focusing web logs 2010 International Conference .

[10] Han J., Pei J., Yin Y. and Mao R., Mining frequent patterns without candidate generation: A frequent-pattern tree approach Data Mining and Knowledge Discovery, 2004.