# Web Usage Mining and User Behaviour Prediction

Prof. Ms. Asiya N. Khan, Prof. Ms. Pallavi S. Pandhare
Assistant Professor, CSE dept., BMIT , Solapur , Maharashtra, India.
*masiya.khan@gmail.com, pallavideshmane26@gmail.com*

***Abstract*****:** Today, Internet is playing such a significant role in our day-to-day life. We have witnessed the evermore- interesting and upcoming publishing medium is the World Wide Web (WWW). The rapid growth in the volume of information available over the WWW and number of its' potential users' has leads to difficulties in providing effective search service for users', resulting in decrease in the web performance. Web Usage Mining is an area, where the navigational access behaviour of users' over the web is tracked and analyzed. So that websites owner can easily identify the access patterns of its users'. By collecting and analyzing this behaviour of user activities, websites owner can enhance the quality and performance of services to catch the attention of existing as well as new customers. This research paper intends to provide an overview of past and current evaluation in users' future request prediction using Web Usage Mining.

***Keyword:*** *Web Usage Mining, Pre-processing, Clustering, Pattern Mining, Link Recommendation.*

*****

## I. INTRODUCTION

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Web Usage Mining can be described as the discovery and analysis of user accessibility pattern, during the mining of log files and associated data from a particular Web site, in order to realize and better serve the needs of Web-based applications. Web usage mining itself can be categorised further depending on the kind of usage data considered they are web server, application server and application level data. This Research work focuses on web use mining and specifically keeps tabs on running across the web utilization examples of sites from the server log records.

Web usage mining is the process of finding out what users are looking for on the internet. Few users might be looking at only documented data, whereas some others might be interested in multimedia data. It is the submission of facts and figures mining techniques to find out interesting usage patterns from World Wide Web facts and figures in alignment to realise and better serve the desires of Web based applications. Usage facts and figures hold the persona or source of World Wide Web users along with their browsing demeanour at a World Wide Web site.

## II. LITERATURE REVIEW

Web mining helps the web designers in discovering the knowledge from the information available in the web. Also it helps the users in getting the fast retrieval of the information they are looking for. Three major areas of web mining are:

Web Content Mining: - Trying to get useful information from the text, images, audio and video in web pages.

Web Structure Mining: - Trying to understand the link structures of the Web which will help in categorization of Web pages.

Web Usage Mining: - Trying to get useful information from the server logs to understand what the users are looking for. It also helps in personalization of web pages.

Though all the three categories of web mining are interlinked, in this research we were going to discuss about the web usage mining. Web usage mining helps the web masters to understand what the users were looking for so that they can develop the strategies to help the user to get the required information quickly.

Web mining is generally implemented by using the navigational traces of users which give the knowledge about user preferences and behaviour. Then the navigational traces were analyzed and the users were grouped into clusters. The classification of navigational patterns into groups helps to improve the quality of personalized web recommendations. These web page recommendations were used to predict the web pages that are more likely to be accessed by the user in near future. This kind of personalization also helps in reducing the network traffic load and to find the search pattern of a particular group of users.

Clustering is a data mining technique used to extract interesting and frequent patterns from the information recorded in web server logs. These patterns were used to understand the user needs and help the web designers to improve the web services and personalization of web sites.

### Web Access Sequence:-

Generally the web usage mining will be done based on the navigation history stored in the logs of the web server. This navigation history is also called as Web Access sequence which will contain the information about the pages that a user visit, the time spent on each page and the path in which the user traverse within the website. So the web access sequences will contain all the details of the pages that a user visited during a single session. This data that we get from the log files will be subjected to various data mining techniques to get the useful patterns which can describe the user profile or behaviour. These patterns will act as the base knowledge for developing the intelligent online applications, to improve the quality of web personalization, web recommendations, etc. The web mining can be generally classified into two categories online mining and offline mining. In offline mining we use the data stored in the log files to find the navigational patterns while in online mining the requests of users in his current active session will be used. Current user profile will be decided by matching the recommendations from both the online and offline methods.

Several systems have been designed to implement the web usage mining. Among many Analog is one of the first systems developed for Web Usage Mining. It has two components online component and offline component. The offline component will reformat the data available in the log file. Generally the web server log will contain the information like IP address of the client, the time in which the web page is requested, the URL of the web page, HTTP status code, etc. Then the data available will be cleaned by removing the unwanted information after which the system will analyze the user's activities in the past with the information available in the log files of the web server and classify the user's session into clusters. Then the online component will classify the active user sessions based on the model generated by the offline component. Once the user group is found then the system will give a list of suggestions to each user request. The suggestions will depend on the user group to which the user belongs.

**Clustering:-**

One of the important portions of web usage mining is the process of clustering the users into groups based on their profile and search pattern. The clustering of the user's session can be done in several ways. Christos et al. represents each page as a unique symbol which makes the web access sequence to a string. Consider S as the set consisting of all possible web access sequences. Then the web mining system will process this set S in offline as a background process or during the idle time to group the pages into clusters such that similar sequences were in the same cluster. The formed clusters were represented by means of weighted suffix tree. The clustering is done by constructing a similarly matrix which when then is given as input to k windows clustering algorithm to generate the clusters with very similar length then the global alignment has to be taken into account rather than the local alignment. Also the scores were calculated for both the local and the global alignment. A simple way to calculate the scores is to assign a positive value to a matching sequence and a negative value for a mismatch.

Two web access sequences were said to be similar if they have the maximum alignment in their sequence. Sometimes the web pages listed in the sequence may be unimportant to the user. C.Suresh et al had proposed an approach in which the clusters were identified based on the distance based clustering methods also they had developed a framework to compare the performance of various clustering methods based on the replicated clustering. In traditional methods, the distance between two user sessions will be calculated using the Euclidean-distance measure. But experiments show that the Sequence Alignment Method is better in representing the behavioural characteristics of web users than the Euclidean-distance method. Cadez et al categorizes the users session as general topics and the behaviour of each particular topic is represented by morkov chain. Fu et al. uses Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm for clustering at page level. BIRCH is a distance-based hierarchical algorithm and it is used for clustering the web user sessions. It has been noticed that the increase in the number of pages is diminishing the performance of the BIRCH algorithm. Since each website contains hundreds of pages considering each page as a separate state will make the clustering unmanageable. To overcome this difficulty the authors proposed an approach to generalize the sessions using attribute-oriented induction. In this new approach the clustering of pages will be done at the category level. It is has been always a difficult job to map a particular page to a specific category but it can be done by using clustering algorithms.

## III. PROPOSED SYSTEM

The aim of the proposed system is to recognize usage pattern from web monitor files of a website. Apriori and FP Tree Algorithm is used for this. Both are prominent algorithms for mining frequent item sets for Boolean association rules. In computer science and data mining, Apriori is a typical algorithm for understand associatIOn rules [10]. Apriori Algorithm follows "bottom-up" technique, used to design to operate on databases containing transactions.

### A. WEB USAGE MINING:

Web usage mining is a regular detection of patterns in click streams and linked data collected or generated as an outcome of client communications with one or more Web sites. The purpose is to scrutinize the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are usually symbolized as collections of sheet, objects, or resources that are commonly accessed by collection of users with common interests.

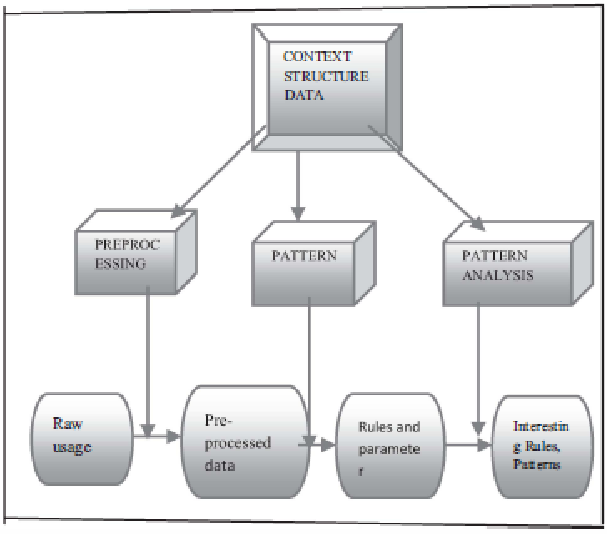

**Fig1: WEB USAGE MINING PROCESS**

**Algorithm for new improved fp -tree web using mining :**

Variables used in this algorithm are as follows:

- URI Stem: is the field in the log that corresponds to the address of a web-page.
- AtEndOfLog: tells us, whether the log record come to an end.
- Token: is a variable that is initially set to a value of 0
- SID: is the session ID of the record that has been retrieved from the log record.
- Write: function that writes the instructed value in a file.

First an array arr is maintained where a number of unique session ids are stored.

```
While Not At End Of Log
Read Log Record
Token = 1
If SID = arr then
Token = 0
End if
If Token = 1 then
arr (k)= SID
k=k+l
end
if
Wend
Session Distribution
Session x Starts
For I = 0 to n
While not At End Of Log
Read Log Record
/* only those files that have either .asp or .html
extension name are being selected */

If SID = arr (i) and right (URI Stem, 4) = ".asp" or right
(URI Stem, 5) = "html "and

/* repeated occurrence of URI Stem is ignored */

URI Stem # uri then
write URI Stem
uri = URIStem
End if
Wend
Next
Session x Ends
```

Thus a log file that has mega bytes of data can be reduced to a few bytes. The above Algorithm works for a single session, this can be repeated for a desired number of times which is equal to the number of sessions required to analyze.

## IV. CONCLUSION

Web usage mining is the procedure of finding out which users are looking for the internet. It can be described as the sighting and scrutiny of user ease of access pattern, during mining of files and its connected data from a Web site, in order to recognize and better serve up the desires of Web-based applications. This algorithm is used in the present Research work to generate association rules that associates the usage pattern of the clients for a website. The output of the system was in terms of memory usage and speed of producing association rules. In future the algorithm can be extended to web content mining, web structure mining, etc. The work can also be extended to extract information from image files.

## V. REFERENCES

[1] K .S .K .D. Association Rules Mining: A Recent Overview, GTS International Tran on Computer Science, Vol.65 (1), 2006, pp.45-65

[2] A R "Fast Algorithms for Mining Association Rules", Sep 12-15 1994, Chile, 487-99, pdf, 1-55860-153-9.

[3] Mannila H,"Efficient algorithms for discovering association rules mining." conference Knowledge Discovery in Databases (SIGKDD). 181-83.

[4] Tan, P. N., M. St., V. Kumar, "Introduction to web Mining", Addison-Wesley, 2013, 769pp.

[5] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementation, 2nd ed. San Mateo.

[6] Huang, H., Wu, X .. Association analysis with one scans of web data bases. Paper submitted at the IEEE On Data Mining, Japan.

[7] R. Jin "An Efficient Implementation of Apriori Association web mining," Proc. Workshop on High Performance Data webMining, Apr. 2011.

[8] J. H and M. Kaber, "association mining:" 2014.

[9] Han J "Mining frequent patterns without candidate rules mining technique," in the national seminar of the international web of data, ACM Press, pp. 4-11- 2004

[10] E-H. Han, G. Caryopsis "Scalable Data web mining for Association web Rules," IEEE Trans. Eng., vol. 12, no. 3, July 2012.

[11] Brin S., R. Mot, J.D. Ullman, "web item set counting and implication rules

[12] Association mining in data base", in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp.289-294, 1999.

[13] Masseglia F., "Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure language", In ACCM Web Letters, Vol. 10 No. 9, pp.13-19, 2011.

[14] Sturnme.A G., Hotho A.H and Berendt B. "Semantic Web Mining-A web survey"MIIT press in Delhi, No.2, pp.124-143, 2014.

[15] Pei J. and Han J. "Constrained frequent pattern mining: a pattern growth view" in SIGKDD Explorations, Vol. 4, No. 1, pp. 31-39,2004.

[16] Antunes C. and Olivei A.L.G "Generalization and association web Pattern-Growth for Sequential attern web Mining with Gap Constraints" in Int'l Conf Machine Learning and Data in published 2012.