# A STUDY ON WEB USAGE MINING

**G.Vijaiprabhu[1], Dr.K.Meenakshisundaram[2]**

Ph.D. Research Scholar [1], Associate Professor[2]
Department of Computer Science[1,2]
Erode Arts and Science College, Erode[1,2]

***Abstract*** - Expedite advancements in the communication and information technology increases immense use of web and the behaviors of the learners. The web has become an information hub that is available in a pervasive manner. Now-a- days web is an admirable platform for handing and acquiring needed records for everyone. Web mining is an application of data mining extracts beneficial data or pattern from the web facts. Web mining comprises of three categories web content, usage and structure mining. Amongst them web usage mining uses data mining strategies to extract understanding pattern from weblog or blog. Web log data are the crucial sources to portray user intent and access behaviour. This technique is referred to as user behaviour pattern analysis and it assist to investigate data, user behavior and it aids in future developments. Digging knowledgeable information and analyzing the log files is very tough as data over the web log is not in readable text. The most common used technologies in web mining are user access pattern analysis, clustering, classification and statistics filtering. The supremacy in today's era has changed the concern from the traditional discovery of knowledge to the operational knowledge of decision support system. This paper presents the perspective of web mining and various data mining statistical classification techniques used to gain information of web usage.

***Keywords:- Data Mining, Web Mining, Web usage mining..***

## I. INTRODUCTION

Mining the data is the method to [4] discover information by analyzing massive set from various perception and extracting useful information via procedures. This is the most motivated research area to find of variant patterns. The main goal is to discover the knowledge hidden in data. Due to enormous growth of data, mining is at drastic level in each field. Getting the right information from data is the most challenging task. Many academicians and industry researchers are engaged on the process of knowledge mining due to abundance of data. It is the core step of Knowledge discovery procedure The recent aspects and development promotes the rapid growth of KDD and DM.

### *A. Stages in Mining includes*

### *1) Analysis and Selection*

The decision makers need to formulate goals, problem and objectives must be clearly defined. The process could not be proceedwithout the idea of the outcome. Selection includes finding the best source databases for the requirement.

### *2) Pre processing*

While creating the data store or warehouse, it is integrated from various sources. So there is a possibility for missing data, data conflicts and data ambiguity. To avoid this circumstances data is cleaned in this stage.

### *3) Transformation*

Data are transformed from one format to another format, which is more appropriate for mining.

Some techniques are smoothing, Aggregating, Normalization etc.

### *4) Data Mining*

The sample of data is put against relevant techniques in data mining. Classification , clustering or Association rule mining are applied until a suitable methods is selected for further exploration, testing and validation.

### *5) Interpretation/evaluation*

Explaining the results to the decision makers is an important step in the mining process. For each technique the results are evaluated and their significance is interpreted. Most data mining tools has visualization modules. These tools communicate the results with more than two dimensions.

## *B. Web Mining*

Web mining is the application [5][20] of data mining toreveal patterns from the world wide web.It is the largest data base, isgrowing in unsystematic way. The pages arelinked each other, but are not organizedlogically. During some course of period millions of webpages are added to web and undergochanges daily. This leads to informationoverloading. So in this situation, getting desiredinformation or particular details is burden.Therefore a very efficient and effectivetechnique is needed to extract or access the requiredinformation. The main challenge behind is theextraction of information with less attempt and time. Another major issue is therelevancy of the information.Proper management of dataimprove the retrieval efficiency. Web Mining is the widelyaccepted method for this. To satisfy the requirements of web crawlers one ofthe most used functional techniques in webmining is to analyze the user browsing patterns through web usage mining. The three broad areas in web mining are: Web content, Web Structure and Web Usage mining.

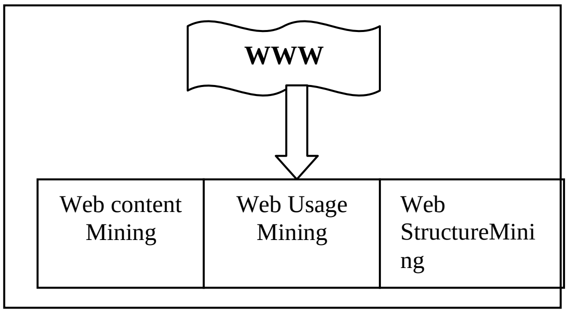

**Figure 1. Web Mining**

Figure 1. shows the three categories of web mining namely content, usage and structure mining.

### *1) Web Content Mining*

The extraction of desired [14][19] information from the unstructured raw data is referred to as Web content mining. A set of information extraction tools is used to identify and collect content items. It is the process of extraction of information fromweb document that may be in any of the format as video, audio, text or structured records.

### *2) Web structure mining*

This is the process of assessing the nodes and connection structure of a site through the use of graph theory. There are two concerns that can be obtained here. One is the structure of a website and how it is connected to other sites and the document structure of the website that how each page is connected.

### *3) Web usage mining*

This is the process of finding patterns and information from server logs to have the idea on the user activity including where the users are from, how many users clicked on which site and the types of activities being done.

### Objectives of Web mining

- To get the knowledge hidden in the log files.
- Identification of user loyalty and interest,
- Enhancing design and usability of web sites,
- Improving search efficiency by modifying the linkage structure.

## II. WEB USAGE MINING

Web usage mining refers to the automatic detection and analysis of patterns in blog data and it relatesthe data collected and[13] generated as a result of learners interactions with resources like websites and blogs. The aim is to capture, model, and assess the behavioral patterns and profiles of the users on the basis of interacting with the website or blogs. The revealed patterns are finally represented as collections of pages, objects and resources with frequently accessed groups of users with common desires.

### *A. Phases in Web Usage mining*

1. Data Collection,
2. Data Pre processing,
3. Pattern Discovery,
4. Pattern Analysis.

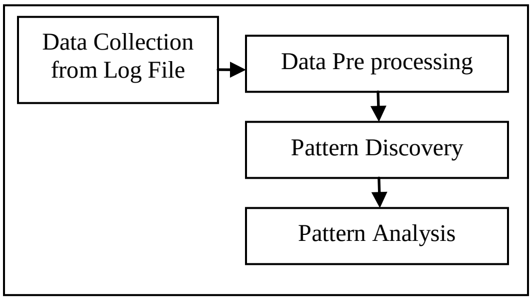

**Figure 2. Web usage mining process**

Figure 2. explains the Web usage mining process from data collection to pattern analysis.

### *1) Data Collection –*

A Web log files records information when a user submits a request to a server. It is a text file which is created automatically, when a user requests a page. It is a file on which the server records information every time a user requests a resource from a site. When a user sends request to the server, the databases will be retrieved. At the same time , the user session including the URL, Client's IP address, accessing date and time , query stem will be recorded in the logs.

A log file resides at three different areas:

### *a) Web Server Log*

These log files resides inside the web server and keeps track of the activity of the user browsing site. The four types of web server logs are agent logs, access logs, referrer logs and error logs.

### *Types of Web server log files*

Access Log – Stores the information of requested filefrom user.

Example: 120.236.0.14 -20011-12-12

Referrer Log – Stores information of the path or 'Uniformresource Locater' of the pages from other sites that link tothe pages.

Example: http://myblaze.sez.html>/library/lectures/news.gif

Agent Log – Stores information about the web clientsthat sends request to the server.

Example: Internet Explorer/5.0(win 7 ;)

Error Log –Stores information about [18] failed requests on the server.

### *Top five error logs*:

The five most common HTTP errors on Google are –

- HTTP Error 500 – Internet server error.
- HTTP Error 403 – Forbidden
- HTTP Error 404 – Not Found.
- HTTP Error 400 – Bad Request.
- HTTP Error 401 – Unauthorized.

### *b) Web Proxy Server Log*

These log files contains information about the proxy server on which the user request arise to the web server.

### *c) Client browser Log Files*

These log files resides in client's browser anda special software are used to store the details.

***Fields in Log files***

- User IP address.
- Time stamp.
- Mode of Request,
- Host IP address,
- Requested URL.
- Status Code.
- Content size in Bytes.
- Agent type.
- Remote URL.

***Example File formats in Log file***

- **Apache common Logging**

A java based logging [17] facility. It provides Application Program interface, log and wrapper implementations.

- **Common Log File system**

A general purpose logging facility accessible to kernel and user mode. It was used with windows operating system for data and event logging.

- **Common Log Format**

A standardized text files facility used by servers when generating log files. The files can be readily analyzed by web analysis programs.

- **Extended Log Format**

An extended version of common log format with more information and flexibility.

***2) Data Preprocessing***

Data preprocessing is a technique that integrate databases and make raw [12] information to be understandable and consistent . The information stored in web logs is processed as it has insufficient and noisy data. It is done in early step by removing redundancy, useless, error, incomplete, inconsistency. There are many e-sources and web usage mining analyze data logs, site address, login information, access logs, cache, cookies etc. It includes methods like cleaning and User, session identification.

***3) Pattern Discovery***

Pattern discovery is the key component in web usage .[15] [16]Itincludes the algorithms and procedures from data mining, machine learning and pattern recognition. A variety of methods are used to find hidden data information on Web server logs. It includes methods like association rules, statistical classification clustering and sequential patterns.

***4) Pattern Analysis***

In this stage, repeated patterns are eliminated and relevant and meaningful patterns are found using Structured Query Language a knowledge query mechanism and On-line Analytical Processing a multi-dimensional data cube,Usability analysis a modeling technique to accessing the behavior of user on the web site and Visualization Technique a graphical method makes the result into suitable and readable format. Using these methods, the output obtained at previous phase is structured.

### *B. Web Personalization*

Web personalization is [7] the process of customizing a Web site as per the needs of users and taking hidden knowledge from the analysis of the user's navigational behavior combined with other information collected from the web logs.

**Principal elements includes:**

- Categorization and preprocessing of Web data,
- Extraction of relations between variant data,
- The recommendation of the actions taken for better designing of site.

***Process of usage-based Web personalization -***

It consists of five modules:

***1) User profiling:*** The process of gathering information of each visitor. It includes demographic information, user interests and behavior when browsing a site. This information is used to customize the content and structure of a Web site to meet the visitor's needs.

***2) Log analysis and Web usage mining****:* The data stored in Web server logs is processed by applying data mining techniques toextract statistical information, cluster the users into groups and discover relations between Web pages and user groups.

***3) Content management*:** The process of classifying the content of a site to make information retrieval and presentation easier for the users. This is very important for the sites whose content is increasing on a daily basis, such as news sites.

***4) Website publishing:*** Presents the content stored locally in a uniform way to the end-user. Variant technologies are used to publish data on the Web.

***5) Information acquisition and searching:*** Not only the information is stored in the server provided by a site. In the case of a Web portal users are interested in searching the same information from various Web sources. So it is the duty of the web site editors to search the Web for content of interest. Searching and relevance ranking techniques is used for the acquisition of relevant information data to each group of users.

### *C. Applications of Web usage Mining*

- Personalized marketing has enabled e-commerce which eventually results in higher trade volumes.
- Government agencies are benefited by classifying threats and fight against terrorism.
- Companies can establish better customer relationship by understanding the needs of the customer and satisfy customer needs faster.
- In business information are gathered to improve customer attraction, retention, sales marketing and better advertisement.
- Identifying common access user behaviors can beused to improve the actual designof Web pages and to make other changes to the site as per the customer needs .

## III. CONCLUSION AND FUTURE SCOPE

Web mining is the application of data mining finds a pattern on web oriented tasks. It includes web content, usage and structure mining. The user visited sites are normally stored in Web log which is a data source for the web usage mining. Web usage mining discovers needed pattern based on users behavior and interest on web pages. This paper focuses on Web usage mining, its applications and process which are used in Pattern discovery.

In future, the Pattern discovery in web usage mining is done by the statistical pattern recognition methods with appropriate web log data.

## REFERENCES

[1] Amit DipchandjiKasliwal, Dr. Girish S. Katkar, “Web Usage mining for Predicting User AccessBehaviour”, International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015.

[2] AnandanBellie, “Web Usage Analysis of University Students to Improve the Quality of Internet Service”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 4 Issue 5, May 2015.

[3] Anupama Prasanth, “web personalization using web usage Mining techniques”, International Journal Of Current Engineering And Scientific Research, vol 3, issue 3, 2016.

[4] Arun K. Pujari,“Data mining techniques”,University Press, First 2001.

[5] Bing Liu, “Web Data Mining: Exploring Hyperlinks,Contents, and Usage Data”, second edition, Springer.

[6] Dharmaraajan, M.A. Dorairangaswamy, “Analysis of FP-Growth and Apriori Algorithms on Pattern Discovery from Weblog Data”, IEEE International Conference on Advances in Computer Applications (ICACA), 2016.

[7] MagdaliniEirinaki, Michalis Vazirgiannis, “Web mining for Web personalization”, ACM Transactions on Internet Technology, Vol. 3, No. 1, February 200, uploaded in 2016.

[8] Romil V Patel, Dheeraj Kumar Singh, “ Pattern Classification based on Web Usage Mining using Neural Network Technique”, International Journal of Computer Applications (0975 – 8887) Volume 71– No.21, June 2013.

[9] Shiva Asadianfam& Masoud Mohammadi,“Identify navigational patterns of web Users”, International Journal of Computer-Aided Technologies (IJCAx) Vol.1,No.1,April 2014.

[10] Sunil, Doja M.N, “Web Usage Mining Techniques to Improve the Capabilities of E-learning Websites and Blogs”, International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017.

[11] VedpriyaDongre,JagdishRaikwal,“An Improved User Browsing Behavior Prediction using Regression Analysis on Web Logs”, International journal of computerapplications, Volume 120 – No.19, June 2015.

[12] Webb A., Statistical Pattern Recognition, England: John Wiley & Sons Ltd., 2002.

[13] en.wikipedia.org/wiki/Web_mining#Web_usage_mining

[14] en.bitcoinwiki.org/wiki/Web_mining

[15] en.wikipedia.org/wiki/Association rule learning .

[16] tutorialpoint.com/data mining/dm cluster analysis.htm

[17] en.wikipedia.org/wiki/Category:Log_file_formats

[18] royal.pingdom.com/the-5-most-common-http-errors-according-to-google/

[19] www. techopedia.com/definition/web-mining

[20] Zdravco, D. T. Larose, "Data Mining the Web uncoveringPatterns in Web Content, Structure and Usage. Published by John Wiley & Sons, Inc., Hoboken, New Jersey.