Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S3, 2019, pp. 12-15


# Extraction of Knowledge from Web Server Logs Using Web Usage Mining

**B. Harika[1] and T. Sudha[2]**
[1]Research Scholar, [2]Professor & Head, [1&2]Department of Computer Science,
Sri Padmavati Mahila Visva Vidyalayam, Tirupati, Andhra Pradesh, India
E-Mail: harika.bommi@gmail.com

***Abstract* - Information on internet increases rapidly from day to day and the usage of the web also increases, thus there is the need to discover interesting patterns from web. The process used to extract and mine useful information from web documents by using Data Mining Techniques is called Web Mining. Web Mining is broadly classified in to three types namely Web Content Mining, Web Structure Mining and Web Usage Mining. In this paper our focus is mainly on Web Usage Mining, where we are applying the data mining techniques to analyse and discover interesting knowledge from the Web Usage data. The activities of the user are captured and stored at different levels such as server level, proxy level and user level called as Web Usage Data and the usage data stored at server side is Web Server Log, where it records the browsing behaviour of users and their requests based on the user clicks. Web server Log is a primary source to perform Web Usage Mining. This paper also brings in to discussion of various existing pre-processing techniques and analysis of web log files and how clustering is applied to group the users based on the browsing behaviour of users on their interested contents.**

***Keywords:* Web Mining, Web Usage Mining, Extraction of Knowledge, Web Server Logs, Clustering, Web Log Pre-Processing**

## I. INTRODUCTION

Today Internet plays a vital role for sharing the information from one person to another person which leads to increase the usage of web. Organizations and Companies of different Government, public and private sectors have their own websites to satisfy the needs of their individuals. Website is a collection of web pages which consists of different types of data such as text, images, audio and video. Users can access the websites to get the information of the company or organization by means of web browsers such as Internet Explorer, Google chrome and so on. With the huge amount of data available on the websites it is necessary to extract the interesting patterns from web data thus the concept Web Mining is introduced. It is a process used to extract and mine useful information & discovering knowledge from web documents by using data mining techniques is called Web Mining [2]. Web Mining [1] [3] is further classified into three types namely Web Content Mining, Web Structure Mining and Web Usage Mining.

*1. Web Content Mining:* It is the process of retrieving or extracting the valuable information from Web documents and these documents contain different types of data such as text, images, hyperlinks, metadata and structured records [1] [2].

*2. Web Structure Mining:* It is a process of discovering structured information from the websites. The structure of a graph consists of web pages and hyperlinks where the web pages are considered as nodes and the hyperlinks are edges and these are connecting between related pages.

*3. Web Usage Mining:* It is also called as Web Log Mining [4]. It reflects the user's behaviour which can catch the meaningful patterns from one or more web localities [6]. Web Usage Mining focuses on determining the users' behaviour from web log data. Web server logs play an important role to record and store the behaviour of online users. These web logs are needed to perform the Web Usage Mining. Web usage mining is a process of applying data mining techniques and application to analyse and discover interesting knowledge from the web. The following figure illustrates the classification of Web Mining.

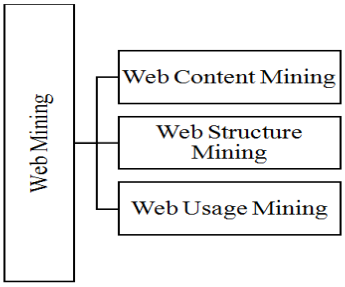


Fig. 1 Classification of Web Mining

Web mining process [7] consists of four important steps they are 1. Resource finding, 2. Data selection and pre-processing, 3. Generalization and 4. Analysis.

1. Resource finding is the process which is used to extract the data either from online or offline text resources.
2. In data selection and pre-processing step, specific information from retrieved web sources are automatically selected and pre-processed.
3. During generalization, data mining and machine learning techniques are used to discover general patterns from individual websites as well as across multiple sites.
4. Validation and Interpretation of the mined patterns are done in analysis step.

The diagrammatical representation for web mining process is as follows

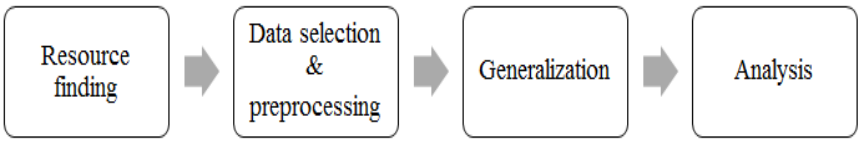


Fig. 2 Web Mining Process

Web Mining techniques are Classification, Clustering and Association Rules, which are used to understand the customer behaviour and evaluate a particular website. Web measurement or Web analytics [8] are one of the significant challenges in Web mining. The measurement factors are user clicks, page views, visits or user sessions and find the unique visitor regularly used to measure the user impact of various proposed changes.

The following lists some major challenges in web mining

1. Web data sets are huge in size; it takes ten to hundreds of terabytes to store on the database.
2. Difficulty in finding relevant information
3. Extracting useful interesting patterns or new knowledge from the web

## II. WEB USAGE MINING

In this paper our focus is on mainly on Web Usage Mining, the term Web Usage Mining was introduced by Cooley *et al.,* in 1997 and in accordance with their definition; Web Usage Mining or Web Log Mining concerned with web logs, is the automatic discovery of user access patterns from web servers or web logs. The process of discovery and analysis of interesting user access patterns focuses on web usage data. The browsing behaviour of users are captured and stored in web server logs at server side called web usage data used for web usage mining. Web Usage Mining [5] process includes the phases as shown in below figure.

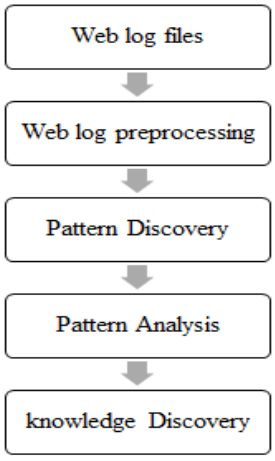


Fig. 3 Web Usage Mining process

Web log files are the primary source for Web Usage Mining, where the Web log files records or captures the activities of the user on the Website. Log files contain information about User Name, IP address, Access Request, Number of bytes transferred, Result Status, URL that referred and User Agent. These Log files can be stored at three places such as Web Servers, Web proxy servers and client browsers.

1. *Web Server Log Files:* These log files records the actions of the user in accessing the Website from the Web Server by means of Web Browsers such as Google Chrome, Internet Explorer and Mozilla Firefox.
2. *Web Proxy Server Log Files:* Proxy Server is an intermediate server between the client and the web server. When the client sends a request to a web server through a proxy server, then the entries to the log file will be from the proxy server. These proxy servers will maintain a separate file for gathering the information of the user.
3. *Client Browser Log Files:* The log files that are stored at the client side are called client Browser log files. Even though these log files present at client side, the entries to it are made by web server.

*A. Web Log Pre-Processing:* Pre-processing techniques are needed to apply on the data collected from web log files because the data in log files may be incomplete, noisy and inconsistent. To improve the quality of the later phases in Web Usage Mining such as pattern discovery and pattern analysis several pre-processing techniques are applied in order to mine useful knowledge from the web log data. Pre-processing of Web log data include Data cleansing, User identification, Session identification, path completion etc. The following figure represents the steps in web log pre-processing.

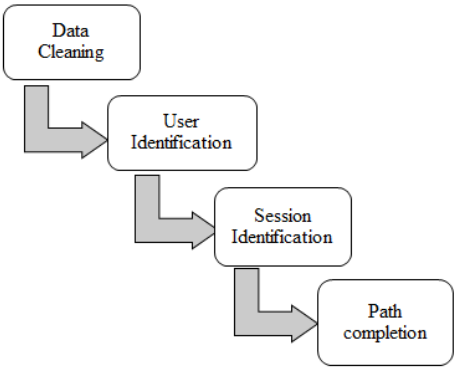


Fig. 4 Steps in Web Log processing

Web Log pre-processing is first and an essential step in Web Usage Mining and it is prerequisite phase to the pattern discovery phase. Due to large amount of irrelevant data in web log files they are not suitable to produce

accurate results from web usage mining process, thus there is the need to remove the irrelevant data from web log files. Pre-processing techniques are applied on data stored in web server logs to improve the quality of the data, make it easier to mine for knowledge. The following are the steps in pre-processing.

1. Data Cleaning [9] is a process applied on web log files to remove irrelevant, noisy, redundant and inconsistent data. It is concerned with removal of useless data such as requests made by web robots, requests for graphical or multimedia objects (e.g. gif, jpeg, jpg, css, mp3 and so on) and failed and corrupted requests (status with value of 200 represents a succeeded request and status with different from 200 represents a failed request and status with 404 indicates that the requested file was not found).
2. User Identification is the second step in web log pre-processing. Identification of user [10] from the log data is a complex task because many users may use the same computer and the same user may use different computers. For websites, needed user registration the log file contains the login details of the user used for user identification. When user login is not available for a website, in that case IP address is used to identify a unique user. First we consider each IP address as a different user. Secondly, if the IP addresses are same but with different browsers and operating systems indicate different users, and thirdly if the IP addresses, browsers version and operating systems are same then we check whether there is a connection between the pages request to access and pages which have been accessed, if we found that there is no direct link between them then we consider as multiple users accesses the websites.
3. Session Identification is the task to find different user sessions from web log files. Session [11] is group of activities made by a single user with a unique IP address during a specific period of time on website. If the time between page requests exceeds 30 minutes then it will consider that the user is starting a new session. Finally the login and logoff represent the logical start and end of the session.
4. Path Completion [12] step is used to acquire the complete user access path. It is a process of adding the page accesses that are not in the web log, out those which have actually occurred. These missing page references occur due to caching and can be completed through path completion.

Pre-processing techniques improve the quality of the data, thereby increasing the accuracy and efficiency of the subsequent mining process. Web Log Pre-processing is an important step in the knowledge discovery, because quality decision is based on quality data. Once the pre-processing is completed on web log data collected from web log servers then the data is ready to perform the next phase in web usage mining process.

*B. Pattern Discovery:* In this phase, patterns are discovered by making use of various techniques like Statistical analysis, Clustering, Association rules and so on. Pattern discovery is the key success to web usage mining which covers the various algorithms and techniques from several research areas such as data mining, machine learning and statistics. This paper brings in to discussion of how clustering is applied in order to discover interesting patterns from web log data.

*1. Clustering*

Clustering is a data mining technique used to analyse huge data sets by making clusters of those data where as a cluster [13]group the objects of "same" and are "different" from the objects of other clusters. Clustering [2] is a technique applied to group items with similar properties. Clustering can be applied on web log servers to group users having similar browsing behaviour, Websites pages group that has similar content, group of users visit similar websites and group of users based on their interested contents.

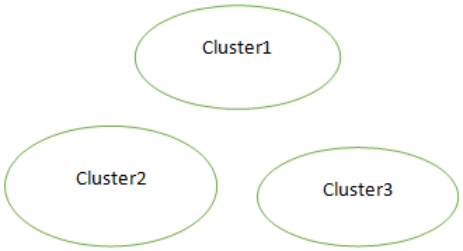


Fig. 5 clustering of similar data

Clustering technique [14] allows one to group together users or data items that have similar characteristics. In Web Usage Mining, clustering techniques are mainly used to discover two kinds of useful clusters, namely user clusters and page clusters. Clustering is a technique applied on pre-processed web log data to group users with same browsing behaviour or same type of interested contents called as user clustering, whereas page clustering is to group pages that are related to another based on users perception. Such knowledge is useful for many applications such as Web Personalization and Market Segmentation in E-Commerce.

*2. Types of clustering*

The following diagram [13] describes the types of clustering

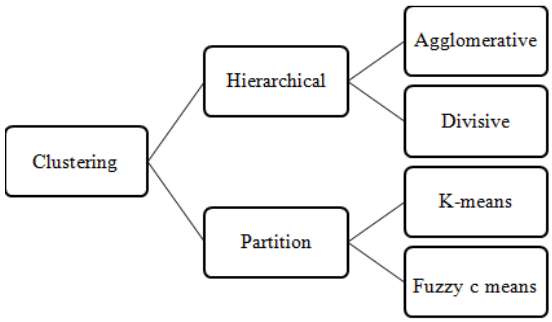


Fig. 6 Types of Clustering

In this paper, we will discuss how K-means clustering algorithm is used in web usage mining to mine or to extract the interesting patterns from web log servers which captures the browsing behaviour of users. K-means clustering is a popular partition clustering algorithm used to divide the pre-processed web log data into different clusters based on their similar characteristics. Web log data [14] in web servers is in the form of CLF (common log file format) or ELF (Extended Log file format).ELF format has two extra fields at the end, which are referrer URL (Uniform Resource Locator) and User Agent. The following is the form of ELF.

<ip_addr><base_url><date><method><file><protocol><code><bytes><referrer><user-agent>

Where ip_addr – user IP address
base_url- requested resource path
date-access date and time
method-HTTP request type
file- html file requested by the user
protocol- protocol used for transmission
code-status code
bytes-number of bytes transferred
referrer-previously visited site by the user
user agent-type and version of the browser

Once the web log data is pre-processed, the irrelevant and useless data are removed from the server logs and the data is ready to apply the K-means clustering algorithm in order to produce the clusters which groups the users based on the browsing behaviour of their interested contents. Let we take the data from the web server logs of any website which consists of journals or articles of different subjects, and by applying the K-means clustering algorithm we try to group the users based on their browsing interested subjects such as the group of users who are interested in Electronics, group of users interest in Commerce, group of users interest in Science and Technology and so on.

The following are the steps in K-means clustering algorithm
We have input data from web log servers x1, x2, ----and assume value of k where k is the number of clusters to form
*Step 1:* Pick random points as cluster centers called centroids
*Step 2:* Assign each data item $X_i$ to nearest cluster by calculating its distance to each centroid
*Step 3:* Locate new cluster center by taking the average of the assigned points
*Step 4:* Repeat step 2 and step 3 until none of the cluster assignments change.
The successful applications of K-means clustering algorithm depend on the value of K. By selecting good K value the clustering algorithm produce meaningful clusters.

## III. CONCLUSION

With the wide and huge volumes of data on the World Wide Web, it is necessary to mine the web in order to discover or to extract the interesting knowledge from web thus the concept of web mining is introduced. In this paper we discussed web usage mining, where it extracts the interesting patterns from web log data in order to discover the navigation behaviour of users. To get the meaningful patterns from web usage mining, first we have to pre-process the web log data to remove useless data from web log servers. Then on the pre-processed data, a popular clustering algorithm is applied to discover interesting patterns from the data and to obtain knowledge from the pre-processed web log data.

## REFERENCES


[1] J. Srivastava, P. Desikan and V. Kumar, "Web Mining: Accomplishments and Future Directions", *Proc. USNatl Science Foundation Workshop on Next-generation data mining* (NGDM) Nat Science Foundation, 2002.

[2] Tawfiq A. Al-Asdi and Ahmed J Obaid, "An Efficient Web Usage Mining Algorithm Based on Log File Data", *Journal of Theoretical and Applied Information Technology*, Vol. 92, No.2, 2016, pp. 215-223

[3] R. Kosala and H. Blockeel, "Web Mining Research: A Survey", *ACM SIGKDD explorations*, Vol. 2, No.1, pp. 1-15, 2000.

[4] AmitPratap Singh and Dr. R.C. Jain, " A Survey on different phases of Web Usage Mining for Anomaly user behaviour Investigation", *International Journal of Emerging Trends & Technology in Computer Science* (IJETTCS) Vol. 3, No. 3, May-June 2014.

[5] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance improvements", in 5th*International Conference Extending Database Technology*, AVignon, France, pp.13-17, March 1996.

[6] C. Gomathi and M. Moorthi, "Web Access Pattern Algorithms in Education Domain" *Computer and Information Science Journal*, Vol.1, No.4, Nov. 2008.

[7] Mr. Dushyant, B. Rathod, and Dr. SamratKhanna, "A Review on Emerging Trends of Web Mining and its Applications", ISSN: 2321-9939.

[8] Dr. S. Vijiyarani and E. Suganya, "Research issues in Web Mining", *International Journal of Computer Aided Technologies (IJCAX)*, Vol.2, No.3, July 2015.

[9] Vijayashri Losarwar and Dr. Madhuri Joshi, "Data Pre-processing in Web Usage Mining", *International Conference on Artificial Intelligence and Embedded Systems* (ICAIES'2012) July 15-16, Singapore 2012.

[10] Sheetal A. Raiyani and Shailendra Jain, "Efficient Pre-processing Technique using Web log mining", *International Journal of Advancements in Research & Technology*", Vol. 1, No. 6, 2012.

[11] Manisha Valera and Kirit Rathod, "A Novel Approach of Mining Frequent Sequential Pattern from Customized Web Log Pre-processing", *International Journal of Engineering Research Applications(IJERA)*,Vol.3, No. 1, pp. 269-380, Jan-Feb 2013.

[12] K. Sudheer Reddy, G. ParthaSaradhi Varma, and M. Kantha Reddy, "An Effective Pre-processing Method for Web Usage Mining", *International Journal of Computer Theory and Engineering*, Vol. 6, No.5, October 2014.

[13] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques", *International Journal of Science Engineering and Technology Research (IJSETR)*, Vol. 2, No. 4, Apr 2013.

[14] J. Srivastava, Robert Cooley, M. Deshpande and Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations ACM SIGKDD*,Vol.1,Issue 2.Jan 2000

[15] R. Suguna, and Dr. D. Sharmila, "User Interest Level Based Pre-processing Algorithms using Web Usage Mining", *IJCSE*, ISSN: 0975-3397, Vol.5, No.9, Sep. 2013.