Access Patterns in Web Log Data: A Review

Mohammed Hamed Ahmed Elhiber and Ajith Abraham*

Sudan University of Science and Technology, Faculty of Computer Science, Khartoum, Sudan
batch_30@hotmail.com
*Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence
Washington, USA
ajith.abraham@ieee.org

Abstract: The traffic on World Wide Web is increasing rapidly and huge amount of information is generated due to users interactions with web sites. To utilize this information, identifying usage pattern of users is very important. Web Usage Mining is the application of data mining techniques to discover the useful, hidden information about the users and interesting patterns from data extracted from Web Log files. It supports to know frequently accessed pages, predict user navigation, improve web site structure etc. In order to apply Web Usage Mining, various steps are performed. This paper discusses the process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has also presented several approaches such as statistical analysis; clustering, association rules and sequential pattern are being used to discover patterns in web usage mining. The pattern analysis phase means applying data mining techniques such as SQL and OLAP on the pattern discovery data to filter insignificant information to obtain the valuable information.

Keywords: World Wide Web, data mining, Web Usage Mining, OLAP.

1. Introduction

World Wide Web is a major source of information and it creates new challenges of information retrieval as the amount of information on the web increasing exponentially. Web Mining is the application of data mining techniques to extract and analyze useful information from Web data. Based on kind of data to be mined Web Mining can be classified into three different categories: Web Content Mining, Web Structure Mining and Web Usage Mining [1].

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files. In Web Usage Mining, data mining techniques are applied to preprocessed web log data in order to find interesting and useful patterns. The web log files on the web server are major source of data for Web Usage Mining. Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user click streams stored in Web log files. The process of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis [2].

Pre-processed and cleaned data could be used for pattern discovery, pattern analysis, Web usage statistics,

Clustering, Association Rules and sequential rules. Clustering is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Pattern analysis is the last step in the overall Web Usage mining process. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The most common form of pattern analysis consists of:

- A knowledge query mechanism such as SQL.
- Another method is to load usage data into a data cube in order to perform Online Analytical Processing (OLAP) operations.
- Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data.

The paper is organized as following: the next section briefly describes some of the main concepts and definitions of the Web Mining followed by log file formats in Section 3. Section 4 presents review Data Pre-processing on Web Server Logs and Section 5 depicts a review of the different techniques in Pattern Discovery. Finally the conclusions of this paper are provided.

2. Web Mining

The term Web-mining (web data-mining), was first mentioned by Etzioni [3], who suggested that traditional data mining techniques for finding hidden patterns in huge databases, can be applied to web-based information. Web mining is an emerging methodology in education research, assisting instructors and developers in improving learning environments and supporting decision-making of policymakers.

Web Mining is use of Data Mining techniques to automatically discover and extract information from web data [4]. Web mining is the general name of the data mining technique used in an attempt to make content analysis from the online web sites. Web mining has the

facility of utilization in two different areas, the first is the analysis related to the content of the pages presented and the second is the analysis based on the user interaction. According to the differences of the mining objects, there are roughly three knowledge discovery domains that pertain to web mining[1][5][6][7]: Web Content Mining, Web Structure Mining, and Web Usage Mining (see Figure 1).
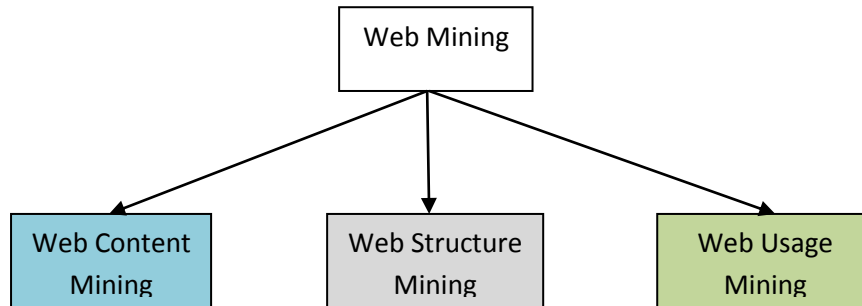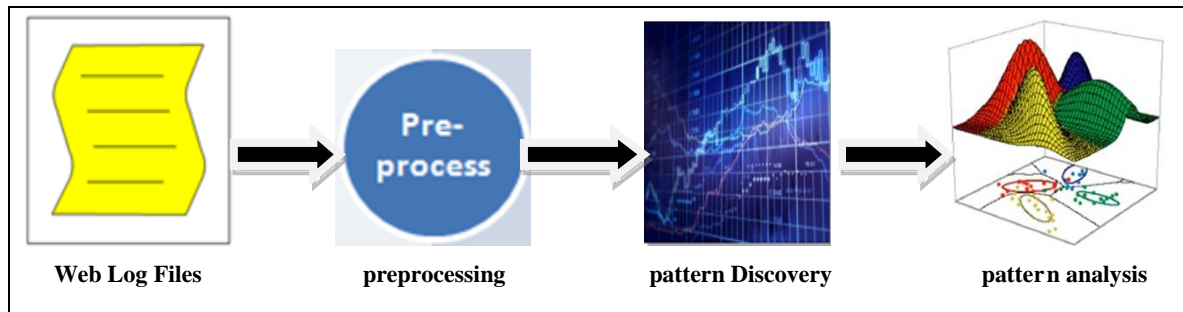

Figure 1: Web mining taxonomy

Figure 2: Phases of Web Usage Mining

2.1. The Usage Mining on the Web

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files. It mines the secondary data (web logs) derived from the users' interaction with the web pages during certain period of Web sessions. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis [8][9][10][11][12], (see Figure 2). The goal of web usage mining is to get into the records of the servers (log files) that store the transactions that are performed in the web in order to find patterns revealing the usage the customers [13][14]. WUM has become an active area of research in field of data mining due to its vital importance [15].

2.2. Analysis of Mined Knowledge

The output of knowledge mining algorithms is often not in a form suitable for direct human consumption, and hence there is a need to develop techniques and tools for helping an analyst better assimilate it. Issues that need to be addressed in this area include usage analysis tools and interpretation of mined knowledge [16].

2.3. Web access patterns

Web access pattern mining is an application of sequence mining on web log data to generate interesting user access behavior on World Wide Web. Mining of web access patterns generated by the users' interaction with the World Wide Web is thrust area of research.

3. The log file: what is it and how do we store information?

3.1 The log file definition

A log file is defined as "a file that lists actions that have occurred" [17]. These files are generated by servers – a computer or a device on a network that manages network resources and contain a list of all requests made to the server by the network's users. A Web log file [18] records activity information when a Web user submits a request to a Web Server. The main source of raw data is the web access log, which we shall refer to as log file.

3.2 The way to store information on a log file

As it is the rule for every file, information in the log file has to be written in a specific format; that is in a specific

sequence and in a certain way that will facilitate the analysis of the file and 'instruct' the computer as to how to read and use [17]. These Log files can be located in 3 places [19][20].

- **Web Servers-** A web server is dispenses the web pages as they are requested.
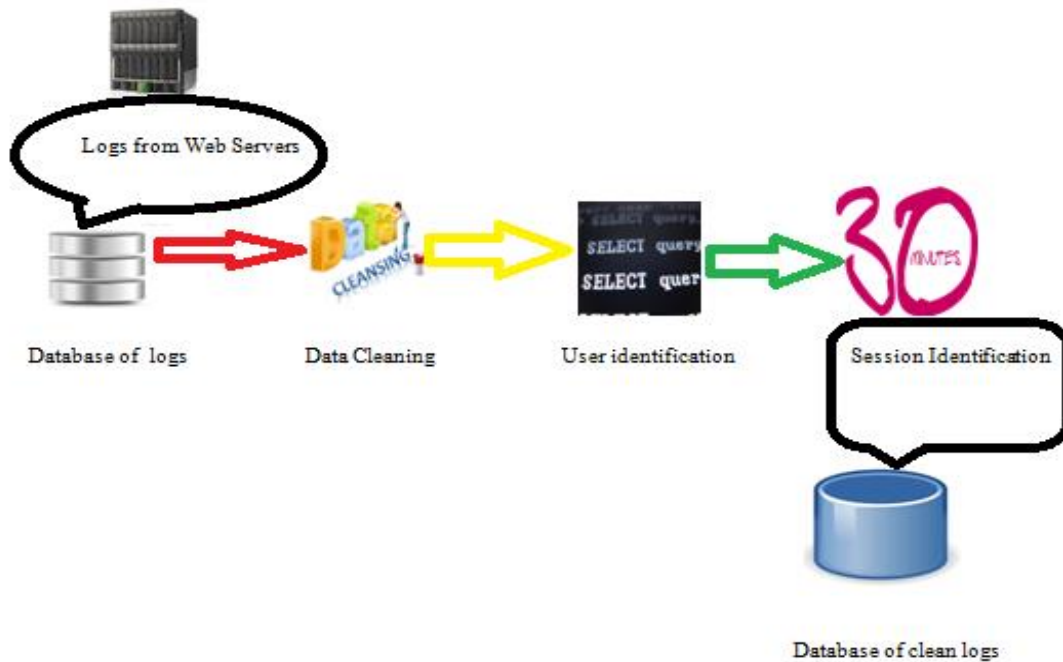

Figure 3: Pre-processing step

- **Proxy Server**- A proxy server is a intermediary compute that acts as a computer hub through which user requests are processed.
- **Web Client**- A Web client is a computer application, such as a web browser, that runs on a user local computer or workstation and connects to a server as necessary.

3.3. Web Server Log File

A server log file is a log file that automatically creates and maintains the activities performed in it. This file is used to record each and every hit to a web site [21].

It maintains a history of page requests, also helps us in understanding how and when your website pages and application are being accessed by the web browser. These log files contain information such as the IP address of the remote host, content requested, and a time of request.

3.4. NCSA Log Formats

National Centre for Supercomputing Application (NCSA) established in 1986 developed a web server called httpd at its centre.

3.5. NCSA Common Log or Access Log Format

Stores basic information about the request received.
Syntax

Host IP address, Proprietor, Username, date: time, request method, status code, byte size.

3.6. NCSA Combined Log Format

Stores all common log information with two additional fields: referrer and user_agent.
Syntax

Host IPaddress, Proprietor, Username, date: time, request method, status code, byte size, referrer, User_agent.

4. Data Pre-processing on Web Server Logs

The data collected in web log file is incomplete and not suitable for mining directly. Pre-processing is necessary to convert the data into suitable form for pattern discovery [22]. We begin **Pre-processing** phase by data extraction then data cleaning and data filtering because the origin web logs data sources are blended with irrelevant information. Data preprocessing plays an important role in Web usage mining. It is used to filter and organize only appropriate information before using web mining algorithms on the web server logs. The original server logs are cleaned, formatted, and then grouped into meaningful sessions before being utilized by WUM. This phase contains three sub steps: Data Cleaning, User Identification, and Session Identification (see Figure 3).

4.1 Data Cleaning

The data cleaning process removes the data tracked in web logs that are useless or irrelevant for mining purposes. The request processed by auto search engines, such as Crawler, Spider, and Robot, and requests for graphical page content. Thus the data cleaning step removes the following entries from the original log file[23].

- The entries having suffixes like .jpg, .jpeg, .css, .map etc.,
- Entries having status code failure.
- Remove all records which are not contain method " GET".
- Remove navigation sessions performed by Crawler ,Spider, and Robot.

4.2. User Identification

User identification is the process of identifying each different user accessing Web site. Goal of user identification is to mine every user's access characteristic, and then make user clustering and provide personal service for the users. Each user has unique IP address and each IP address represents one user. But in fact there are three conditions: **(l)** Some users has unique IP address. **(2)** Some user has two or more IP addresses. **(3)** Due to proxy server, some user may share one IP address. Rules for user identification are:

- Different IP addresses refer to different users.
- The same IP with different operating systems or different browsers should be considered as different users.
- While the IP address, operating system and browsers are all the same, new user can be determined whether the requesting page can be reached by accessed pages before, according to the topology of the site.
- Users is uniquely Identified by combination of referrer URL and user agent

4.3 User session

After identify the users; we need to identify the sessions. To do this we can divide the access of the same users in sessions. It's difficult to detect when one session is finish and start another. To detect sessions is common use of time between requests; if two requests are called in of time frame, we can suppose that these requests are in the same session; in other way below of time frame we can consider two different sessions. A good time frame is 30 minutes.

5. Pattern Discovery

After data preprocessing phase, the pattern discovery method should be applied. This phase consists of different techniques derived from various fields such as statistics, machine learning method mainly have Association Rules, data mining, pattern recognition, etc. applied to the Web domain and to the available data [5]. Several methods and techniques have already been developed for this step [8]. Some of the frequently used solutions are statistical analysis, clustering and association rules.

5.1. Statistical Analysis

Statistical analysis is the most common method to extract knowledge about visitors to a Web site. We can compute various kinds of descriptive statistics measurements like (frequency, mean, median, etc) on variables such as page views, viewing time, or length of the navigation path. Although the statistical analysis useful for improving system performance, enhancing system security, or facilitating site modification. For example, we can detect unauthorized entry points to our Web site.

5.2 Clustering

Clustering has been widely used in Web Usage Mining to group together similar sessions among large amount of data based on a general idea of distance function which computes the similarity between groups. [24] Clustering means the act of partitioning an unlabeled dataset into groups of similar objects. Each group, called a 'cluster', consists of objects that are similar between themselves and dissimilar to objects of other groups. In the past few decades, cluster analysis has played a central role in diverse domains of science and engineering [25], [26]. Two types of clusters can be found in web usage mining: user clusters and page clusters. User clusters will discover users having same browsing patterns whereas page clusters will discover pages possessing similar content [27]. Here we will briefly describe some techniques to discover patterns from processed data. Commonly used clustering algorithms are: K-means, Fuzzy C-means, k-Nearest Neighbor and Neural Network.

5.2.1 k-Means Clustering

The k-means method partitions the data set or to classify your objects based on attributes into positive k cluster in which each observation belongs to the cluster with the nearest mean [28]. The classification is done by minimizing the sum of squared distance in each cluster. Thus, the strength of K-means algorithm lies in its computational efficiency and the nature of easy to use. The procedure follows a simple way to classify a student's data set:

- The basic step of k- means clustering is simple. In the beginning we determine number of cluster k and assume the centroid or center of clusters.

- We can take random objects as the initial centroid or the first k object which serve as initial centroid.
- Then the k means algorithm will do its steps until convergence. Iterate until stable (= no move group) to group the object based on minimum distance.

3. Real Time Operation: ANN computations may be carried out in parallel, and special
hardware devices are being designed and
4. Manufactured which take advantage of this capability.
5. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of.

5.2.2 Fuzzy C-Means Clustering

The fuzzy C-means (FCM) algorithm is one of the most popular techniques used for clustering. It is a kind of partitional clustering algorithm, which aims at partitioning a given data set into disjoint subsets so that specific clustering criteria are optimized[29].The algorithm is an iterative clustering method that produces an optimal c partition by minimizing the weighted within group sum of squared error objective function JFCM [30]:

$$J_{FCM} = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^q \, d^2(x_k, v_i)$$

where $X = \{x1, x2, \cdot \cdot \cdot, xn\} \subseteq Rp$ is the data set in the p-dimensional vector space, n is the number of data items, c is the number of clusters with $2 \leq c < n$, uik is the degree of membership of xk in the ith cluster, q is a weighting exponent on each fuzzy membership, vi is the prototype of the centre of cluster i, d2 (xk, vi) is a distance measure between object xk and cluster centre vi.

5.2.3 Neural Network based Clustering.

The use of neural network is an efficient technique for web usage mining to extract hidden knowledge from web-data in easy and efficient manner. An advantage of ANNs is their ability to represent both linear as well as non-linear relationships. These networks are data mining techniques that have been inspired by the desire to develop artificial systems capable of performing 'intelligent' computations similar to those performed within the human brain. An ANN acquires its knowledge through repeated presentations of data. It 'learns' by adjusting the weights of the network connections, which is similar to adjusting the synaptic weights within the inter-neuron connections within the human brain [31].Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. Neural Network has other advantage [32]:
1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
2. Self-Organization: An ANN can create its own organization or representation of the
information it receives during learning time.

5.3. Association Rules

Association rule mining are one of the major techniques of data mining and it is the most common form of local-pattern discovery in unsupervised learning systems. It serves as a useful tool for finding correlations between items in large database [33]. This method can be used to find group of pages which are frequently accessed together with support exceeding a threshold. It is not necessary that the pages are connected directly. For example, Apriori algorithm can be used to find relation between users who access a faculty page of a college and those who access a syllabus download page.

5.4 Prediction

Prediction and classification are used for predicting users' future requests systems. To provide prediction in Web Usage Mining system we describe most popular approaches for classifying user navigation patterns for predicting users' future requests.

5.4.1. Decision Tree

Decision trees are one of the most popular approaches for both classification and predictions. They are generated based on specific rules. Decision tree is a classifier in a tree structure. Leaf node is the outcome obtained. It is computed with respect to the existing attributes. Decision node is based on an attribute, which branches for each possible outcome for that attribute. Decision trees can be thought as a sequence of questions, which leads to a final outcome. Each question depends on the previous question hence this case leads to a branching in the decision tree. While generating the decision tree, the main goal is to minimize the average number of questions in each case. This task provides increase in the performance of prediction.

5.4.2. Support Vector Machine
The SVM algorithm was developed by [34] and has become a powerful method in machine learning, applicable to both classification and regression. Our motivation to use the SVM algorithm in our approach is that this algorithm has been used in real-world applications and is well known for its superior practical results [35]. It is an elegant tool for solving pattern recognition and regression problems. Over the past few years, it has attracted a lot of researchers from the neural

network and mathematical programming community; the main reason for this being its ability to provide excellent generalization performance. SVMs have also been demonstrated to be valuable for several real-world applications. [36]. Support Vector Machines (SVMs) are popular machine learning method for classification and other learning tasks [37].

5.4.3 K-NN

Nearest Neighbor (also known as Collaborative Filtering or Instance-based Learning) is a useful data mining technique that allows you to use your past data instances, with known output values, to predict an unknown output value of a new data instance. So, at this point, this description should sound similar to both regression and classification. k-Nearest Neighbor is one of the most popular algorithms for text categorization . Many researchers have found that the kNN algorithm achieves very good performance in their experiments on different data sets [38].The general principle is to find the k training samples to determine the k nearest neighbors based on a distance measure. Next, the majority of those k nearest neighbors decide the category of the next instance.

5.5. Sequential Patterns

The technique of sequential pattern discovery can be applied to web server logs. It attempts to find intersession patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions [22]. Using sequential pattern discovery, useful user trends can be discovered, website navigation can be improved and adopt web site contents to individual client requirements or to provide clients with automatic recommendations that best suit customer profiles.

5.6. Pattern Analysis

This is the final step in the WUM process. It helps to filter insignificant information to obtain the valuable information. The Pattern Discovery of web usage patterns, through techniques described above, would not be very useful unless there were mechanisms and tools to help an analyst better understand them. The pattern analysis phase means applying data mining techniques on the pattern discovery data. The patterns are analyzed using several techniques. The most common form of pattern analysis consists of Structured Query Language (SQL), Online Analytical Processing (OLAP)[18]. In OLAP techniques, the result of pattern discovery is loaded into data cube and then OLAP operations are performed. After this, to interpret the results, visualization techniques are used [22], such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. The result of pattern analysis helps to improve the system performance and to modify the web site. It helps to attract the visitors and to give the personalized services

to regular user. The result of such analysis might include: most recent visit per page, who is visiting which page, the frequency of use of each hyperlink, and most recent use of hyperlinks.

6. Conclusions

Web usage mining model is a kind of mining to server logs. It is plays an important role in realizing, enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. WUM has been a potential technology for understanding behavior of the user on the Web.

There are several techniques proposed by different researchers for the web usage mining. This paper discussed about various techniques available for web usage mining. This paper mainly discusses about three vial steps in WUM such as preprocessing, pattern discovery and pattern analysis.

In this paper various algorithms of pattern discovery techniques like Fuzzy C-Means and techniques of clustering, DT and SVM techniques of classification etc. are described for user future request prediction. The patterns are analyzed using several techniques. This paper described the most common form of pattern analysis consists of Structured Query Language (SQL), Online Analytical Processing (OLAP). This paper gives a detailed look about web mining, web using mining, web server log file and its format. Further we plan to apply WUM process for extracting knowledge or pattern from the SUST server logs.

Acknowledgements

References:

[1]   Arvind   Kumar   Sharma,Dr.   P.C. Gupta, "Exploration of Efficient Methodologies for the Improvement In Web Mining Techniques: A Survey", *International Journal of Research in IT & Management* Vol 1, Issue 3, pp.85-95 , July 2011.

[2]   A. Al-qwaqenah and M. Al-kabi, "Discovering the Web Usage in three Jordanian Universities," *The International Conference on Information and Communication Systems*, pp. 1–8, 2011.

[3]   P. Toth, "Applying Web-mining Methods for Analysis in Virtual Learning Environment," *Budapest Tech Polytechnical Institution, Centre for Teacher Training and Engineering Education.*

[4]   K. B. Patel, J. A. Chauhan, and J. D. Patel, "Web Mining in E-Commerce : Pattern Discovery , Issues and Applications," *International Journal of P2P Network Trends and Technology*, vol. 1, no. 3, pp. 40–45, 2011.

[5]   A. Upadhyay and B. Purswani, "Web Usage Mining has Pattern Discovery," *International Journal of Scientific and Research Publications*, vol. 3, no. 2, pp. 1–4, 2013.

[6]   M. Pratap, S. Dohare, P. Arya, and A. Bajpai, "Novel Web Usage Mining for Web Mining Techniques," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 1, pp. 253–262, 2012.

[7]   A. Kusmakar, "Web Usage Mining : A Survey on Pattern Extraction from Web Logs," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 9, pp. 834–838, 2013.

[8]   K. Etminani, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method," *IFSA-EUSFLAT*, pp. 396–401, 2009.

[9]   R. Gupta and P. Gupta, "Application specific web log pre-processing,"   *Int.J.Computer Technology & Applications*, vol. 3, no. 1, pp. 160–162, 2012.

[10]   R. Gupta and P. Gupta, "Fast Processing of Web Usage Mining with Customized Web Log Pre-processing and modified Frequent Pattern Tree," *International Journal of Science and Communication Networks*, vol. 1, no. 3, pp. 277–279, 2011.

[11]   and D. S. Singh, Arun, Avinav Pathak, "Web Usage Mining : Discovery Of Mined Data Patterns and their Applications," *International Journal of Computer Science and Management Research*, vol. 2, no. 5, pp. 2423–2429, 2013.

[12]   M. S. Kamat, J. W. Bakal, and M. Nashipudi, "Improved Data Preparation Technique in Web Usage Mining*," International Journal of Computer Networks and Communications Security*, vol. 1, no. 7, pp. 284–291, 2013.

[13]   S. K. Pani , L.Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi," Web Usage Mining: A Survey on Pattern Extraction from Web Logs", *International Journal of Instrumentation, Control & Automation* , Volume 1, Issue 1,pp.15-23, 2011.

[14]   Sawan Bhawsar, Kshitij Pathak, Sourabh Mariya, Sunil Parihar," Extraction of Business Rules from Web logs to improve Web Usage Mining", Vol 2, Issue 8, Aug,pp.333-340, 2012.

[15]   P. Nithya and P. Sumathi, "A Survey on Web Usage Mining: Theory and Applications," *International Journal*, vol. 3, no. August, pp. 1625–1629, 2012.

[16]   H. Yogish, D. Raju, and T. Manjunath, "The Descriptive Study of Knowledge Discovery from Web Usage Mining," *IJCSI International Journal of Computer Science Issues*, vol. 8, no. 5, pp. 225–230, 2011.

[17]   G. K. Lekeas, "Data mining the web: the case of City University's Log Files," 2000.

[18]   Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang Ning Tan "Web usage mining: Discovery and Applications of usage patterns from web data" *SIGKDD Explorations*- vol-1, issue-2,pages 12-33, Jan 2000.

[19]   K.   Suneetha   and   R.   Krishnamoorthi, "Identifying user behavior by analyzing web server access log file," *IJCSNS International Journal of Computer Science and Network Security*, vol. 9, no. 4, pp. 327–332, 2009.

[20]   S. G. Langhnoja, M. P. Barot, and D. B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for," *International Journal of Data Mining Techniques and Applications*, vol. 02, no. 01, pp. 141–150, 2013.

[21]   O. CU and P. Bhargavi, "Analysis of Web Server log by web usage mining for extracting users   patterns,"   *http://www.   tjprc. org/view_archives. Php*, vol. 3, no. 2, pp. 123–136, 2013.

[22]   K. B. Patel, "Process of Web Usage Mining to find Interesting Patterns from Web Usage Data," *International Journal of Computer Applications & Technology*, vol. 3, no. 1, pp. 144–148, 2012.

[23]   P. Nithya and P. Sumathi, "An Effective Web Usage Analysis using Fuzzy Clustering," *ARPN Journal of Science and Technology*, vol. 3, no. 7, pp. 693–698, 2013.

[24]   Y. Xie, V.V. Phoha, "web user clustering from access log using belief function", in: proceedings of the first international conference on knowledge capture (k-cap 2001), ACM press, , pp. 202–208, 2001.

[25] S. Das, A. Abraham, and a. Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 38, no. 1, pp. 218–237, Jan. 2008.

[26] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Systems with Applications,* vol. 38, no. 3, pp. 1835–1838, Mar. 2011.

[27] S. Dhawan and S. Goel, "Web Usage Mining: Finding Usage Patterns from Web Logs," *American International Journal of Research in Science, Technology, Engineering & Mathematics*, pp. 203–207, 2013.

[28] I. Frades and R. Matthiesen, "Overview on techniques in cluster analysis," *Bioinformatics Methods in Clinical Research*, pp. 81–107, Springer, 2010.

[29] X. Lou, J. Li, and H. Liu, "Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density Related Work," *Journal of Computational Information Systems*, vol. 2, no. January, pp. 72

[30] Pal N.R., Pal K., Keller J.M., Bezdek J.C., A possibilistic fuzzy c-means clustering algorithm, IEEE Transactions on Fuzzy Systems 13(4), 2005, pp. 517–530.

[31] Patel, Pretesh B., and Tshilidzi Marwala. "Caller behaviour classification using computational intelligence methods*." International journal of neural systems* 20.01 (2010): 87-93.

[32] A. Sharma, "Web usage mining using neural network," *International Journal of Reviews in Computing*, vol. 9, no. April, pp. 72–78, 2012.

[33] G. Pandey, S. Patel, V. Singhal, and A. Kansara, "A Process Oriented Perception of Personalization Techniques in Web Mining," *International Journal of Science and Modern Engineering (IJISME)*, vol. 1, no. 2, pp. 26–30, 2013.

[34] Boser, B. E., Guyon, I. M., and Vapnik, V. N. "A training algorithm for optimal margin classifiers", in D. Haussler, editor, *5th Annual ACM Workshop on COLT (*ACM Press, Pittsburgh, PA), (1992) 144-152.

[35] Burges, C. "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, (1998),2(2), 121-167.

[36] Shevade, Shirish Krishnaj, et al. "Improvements to the SMO algorithm for SVM regression." *Neural Networks, IEEE Transactions on* 11.5 (2000): 1188-1193.

[37] Vishwanathan, S. V. M., and M. Narasimha Murty. "SSVM: a simple SVM algorithm." *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on.* Vol. 3. IEEE, 2002.

[38] Li Baoli, Chen Yuzhong, and Yu Shiwen, "A Comparative Study on Automatic Categorization Methods for Chinese Search Engine", *Proceedings of the Eighth Joint International Computer Conference, Hangzhou: Zhejiang University Press,* pp.117-120. 2002.