Volume 5, Issue 11, November 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Page Recommendation and Web Page Customization Using Fuzzy Member Clustering Algorithm

[1]K. Poorni Sri*, [2]S. Radha Priya
[1]M.Phil. Research Scholar, [2]Assistant Professor
[1, 2] PG & Research Department of Computer Science, Government Arts College (Autonomous)
Coimbatore, Tmilnadu, India

Abstract—The growth of World Wide Web is incredible as it can be seen in present days. Users find it very challenging to extract useful and relevant information from the huge amount of information. The problems can be solved by Web Usage Mining which associates preprocessing, Clustering, Feature extraction, etc. This lead to Content personalization has widely used in personalized search engines for better personalized results. For that user actions and events on the web search engine are monitored. User interaction on the page, session identification and clustering objects plays a vital role in recommender systems. This paper discusses various data preprocessing techniques, algorithm used for clustering and experiment evaluation of proposed method.

Keywords—Log file, preprocessing, content personalization, clustering, Ipaddress, server log table.

I. INTRODUCTION

Web mining refers to the use of data mining techniques to automatically retrieves, extract and analyze information for knowledge discovery from web documents and services. Web Usage Mining is a densely researching area in the field of data mining [1]. Web usage mining provides the support for the web site design, affords personalization server and other business making decision, etc. According to the differences of the mining objects, there are approximately three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their explanations. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting Patterns in web access logs [2].

II. PERSONALIZED CONTENT MANAGEMENT

Content personalization has widely used in personalized search engines for better personalized results. For that user actions and events on the web search engine are monitored. The process of making the personalized web page based on the user interest and popular data is called content personalization. User interaction on the page plays a vital role in recommender systems. Previous studies on personalization systems have mainly focused on modeling techniques and feature development, this content personalization is based on general behavior analysis algorithm. The system proposes a novel improved implicit user response and event monitoring schemes for effective content personalization [3]. For this the system proposes a new scheme named as PCM (Personalized Content Management). But user interactions in real-world Web applications are unlikely to be as ideal as those assumed by previously proposed models. The proposed system builds an online dynamic learning framework for personalized recommendation. The main contribution in this paper is an approach of personalizing users' searches to achieve better search. To achieve better item relevance estimation, the system uses the following parameters.

- Event monitoring and click behaviors from Web search,
- Personalized content optimization using FCLAM (Fuzzy clustering by Local Approximation of Membership algorithm) for effective web personalization.

Any good personalization approach starts with a fundamental understanding of your customer's behavior, needs and goals. Content personalization (or customization — take your pick) is a strategy that relies on visitor data to distribute relevant content based on audience interests and motivations.

III. PROBLEM OBJECTIVES AND METHODOLOGY

This paper focuses on providing techniques for better data cleaning and feature extraction from the web log. Log data is usually noisy and ambiguous, data preprocessing is an essential process for efficient mining process [4]. In the preprocessing, the data cleaning process involves removal of records of graphics, videos and, the records without the HTTP status code, and user identification is performed for cleaning.

The next primary goal is to learn the user's access log and their use of web resources in web usage mining. After the process of data cleaning extract features from the accessible actions in the log. Identifying the potential attributes and reducing the dimensionality of the data by not adding irrelevant attributes are the major role of feature extraction. The assignment is to change inconsistent length transactions into fixed-length feature vectors. The potential feature set extraction will lead to better understanding of the user navigation patterns in web server log files instead of taking into the consideration of whole details in the log file [5]. The steps involved in attribute extraction and techniques for user identification and session identification has been explained in the following sections. The data cleaning, Fuzzy membership and feature extraction process is explained in detail.

Data Preprocessing

Data preprocessing plays an important role in Data Mining. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends [6], and is likely to contain many errors. In this section, this has accepts the approach of web based mining for cleansing the web server log files. Web mining extracts useful information of hypertext documents. Once a user access the web pages /sites their information are recorded in a file as an entry called Web log file.

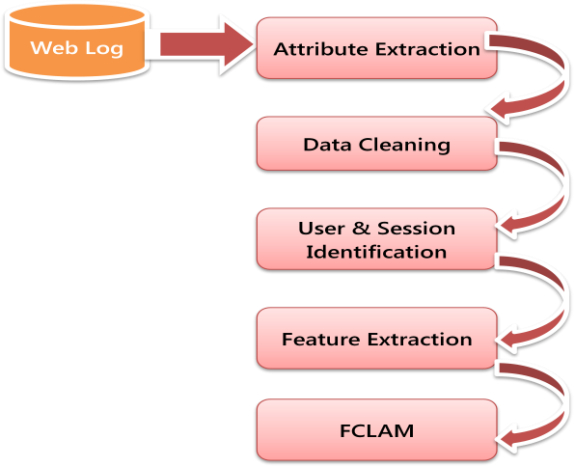

Fig.1 General flow of the preprocessing stage

Data Pre-Processing in Web Content Personalization

In Web content personalization to perform web usage mining preprocessing is mandatory, because Log file encloses noisy and unclear data which may affect results of the mining process. Some of the web log file data are irrelevant for analytical process and could affect the performance of personalization [7].

Any real time data mining project usually spends 80% of the time on the data pre-processing step. The ground work of web data mining is done at preprocessing phase. Data preprocessing use log data as input then process the log data and produce the decisive data. The goal of data preprocessing is to choose cardinal features then remove irrelevant information and finally transform raw data into sessions [8]. In Fig. 1 shows the steps involved in data preprocessing. To achieve its goal Data preprocessing is divided into Data Cleaning, user identification, and Session Identification, Feature extraction and last step to clustering algorithm.

Attribute Extraction

Each user entry is represented as a single line of the log file. The log entry contains many attributes as discussed in the earlier section which has to be separated out for further processing. The filed extraction is the process of separating the attribute from the single line of the server log file. The server used different characters which work as separators. The most used separator character is ',' or 'space' character.

Using the attribute extractor, attributes are separated using the delimiters. Then the data cleansing process is applied for filtering the unwanted and irrelevant data entry to increase the quality of data. Here Delimiter based Attribute Extraction is used in the attribute extraction method to get the extracted web log files for further cleansing process.

Data Cleaning

Data cleaning is to remove all the useless data used in data analysis and mining. Data cleaning is necessary for increasing the mining efficiency [9]. Cleaning of data, unwanted data will be deleted. Examples of unwanted data include requests for images, flash animations, video, etc. These data are not required for user navigation and hence are deleted from the log file. The identification of users through the IP address is the most frequently used method as it is simple, easy to capture and is never empty [10]. Session Identification is performed using session timeout value.

User Identification

User's identification is, to identify who access web site and which pages are accessed. The aim of this process is to retrieve every user's access characteristics, then make user clustering and provide recommendation service for the users [11]. There are three conditions to identify the user:

- Some user has a unique IP address
- Some user has two or more IP addresses
- Some user may share one IP address due to the proxy server.

Distinct User Identification analyses more factors, such as user's IP address, Web site's topology, browser's edition, operating system and referrer page. This algorithm possesses preferable precision and expansibility. It can not only identify users but also identify session. Session identification will be discussed in next section [12]. This method shows comparison not only based on User_IP somewhere same User_IP may generate the different web users, based on path which chosen by any user and access time with referrer page we find out the distinct web user.

Input: Given a clean and filtered web log file and record set web log file
Step1: input Log file
Step2: Distinct User identification base
Step3: Read URL, ipaddress, operating system, status, session
Step4: Read log DB of user
Step5: Check new user or existing user
Step 6: End
Output: Identified User

Session Identification

It defines the number of times the user has accessed a web page. It takes all the page reference of a given user in a log and divides them into user sessions. These sessions can be used as an input data vector in classification, clustering, prediction and other tasks [13]. Based on a uniform fixed timeout a traditional session identification algorithm is used. A new session is identified when the interval between two sequential requests exceeds the timeout. The session identified contains more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions the obtained cleaned datasets will be used for further personalization process [14]. The Cleansed data has taken for the next step for further clustering process for neighbour graph, Cluster Objects and membership by FCLAM algorithm is described below.

Fuzzy Clustering by Local Approximation of Memberships

Fuzzy clustering by Local Approximation of Memberships (FCLAM) defines clusters in the dense parts of a dataset and performs cluster assignment based on the neighbourhood relationships among objects. The FCLAM constructs k-Nearest Neighbours graph to identify the cluster centers and outliers. Proteins with the highest local density called Cluster Supporting Objects (CSO) and proteins with a local density lower than a threshold are called outliers. CSOs are assigned with full membership to represent it as cluster centers. Outliers are assigned with full membership to the outlier group. Fuzzy memberships are then assigned to remaining proteins with varying degrees of memberships to the cluster supporting objects. There is no need to specify the predefined number of clusters [15]. It automatically determines the numbers of cluster and outliers. FCLAM requires the number of k-Nearest Neighbours and threshold value for outliers as initial parameters.

The FCLAM algorithm is mainly divided into three steps:

1. Extraction of the structure information from the weblog dataset:
    1. Construct a neighbourhood graph to connect each object to its K-Nearest Neighbours (KNN);
       Based on the IP, the neighbourhood graph will be constructed.
    2. Estimate a density for each object based on its neighbours to its KNN; number of neighbours for each user based on IP and query.
    3. Users and Objects are classified into 3 types:
        i. Cluster Supporting Object (CSO): object with density higher than
            a. all its neighbours;
            b. Based on the weblog, the system finds highest frequency link.
        ii. Cluster Outliers: object with density lower than all its neighbours,
            a. and lower than a predefined threshold;
        iii. The rest.

2. Local/Neighbourhood approximation of fuzzy memberships:
    1. Initialization of fuzzy membership:
        i. Each CSO is assigned with fixed and full membership to itself to represent one cluster;
        ii. All outliers are assigned with fixed and full membership to the outlier group;
        iii. The rest are assigned with equal memberships to all clusters and the outlier group;
        2. Then the fuzzy memberships of all type 3 objects are updated by a converging iterative procedure called Local/Neighbourhood Approximation of Fuzzy Memberships, in which the fuzzy membership of each object is updated by a linear combination of the fuzzy memberships of its nearest neighbours.

3. Cluster construction from fuzzy memberships in two possible ways:

1. One-to-one object-cluster assignment, to assign each object to the cluster in which it has the highest membership;
2. One-to-multiple object-clusters assignment, to assign each object to the cluster in which it has a membership higher than a threshold.

In this algorithm, weblog data is given as input and the structure information is extracted from the web log data. For this, neighbourhood graph is constructed in order to connect each object to its K-Nearest Neighbours (KNN). The similarities between every pair of objects are calculated, and the nearest neighbours are also identified. After constructing the neighbourhood graph, density for each object is computed based on its proximities to its KNN [16]. The distance/proximity between each object and its k-nearest neighbours is mainly used to determine the object density. Objects are classified into 3 types such as Cluster Supporting Object (CSO), Cluster Outliers and the rest. Object with density higher than all its neighbouring objects are known as CSO. Object with density lower than all its neighbouring objects, and lower than a predefined threshold are known as Cluster outliers.

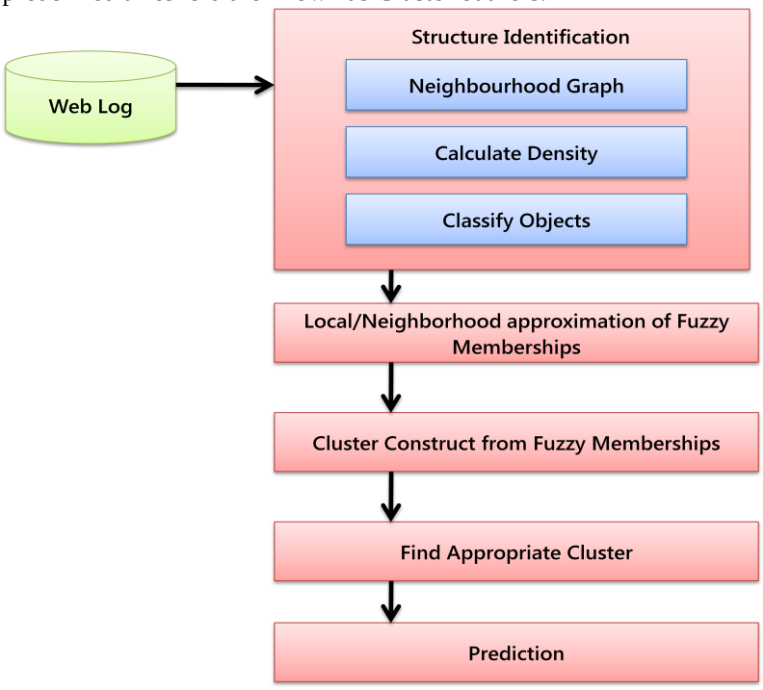

Fig.2 Flow of FCLAM clustering algorithm

The clusters are constructed from the fuzzy memberships in two ways:

- By assigning each object to the cluster in which it has the highest membership degree,
- Threshold value is applied on the memberships, and assigns each object to one or more clusters in which it has a membership degree higher than the threshold.

After clustering, the page which has more weight has more probability for opening that webpage in future by user.

IV. EXPERIMENTAL EVALUATION

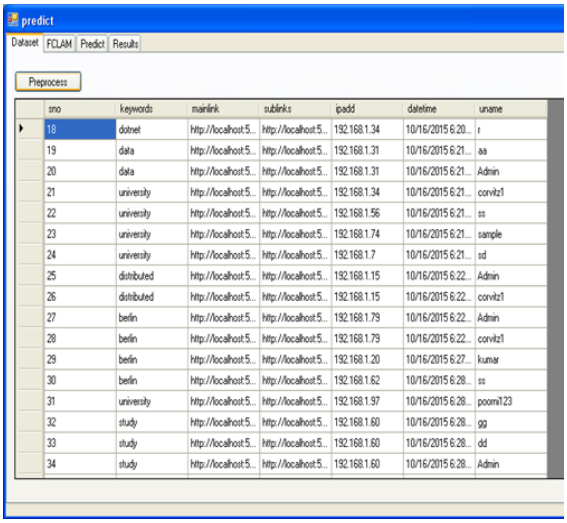

Fig. 3 Preprocessed Data

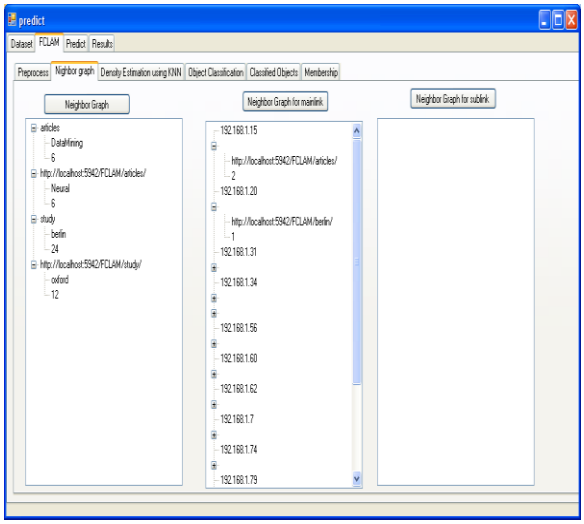

Fig.4 FCLAM Process1: Neighbour Graph Analysis

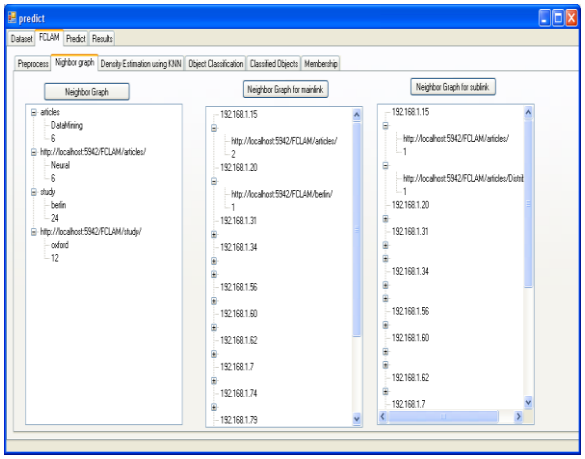

Fig.5 FCLAM Process2: Neighbour Graph for Main Link

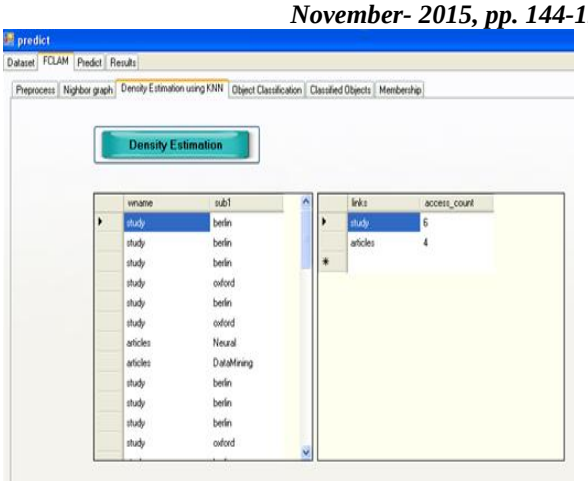

Fig.6 FCLAM Proces3: Density Estimation and Sub Link

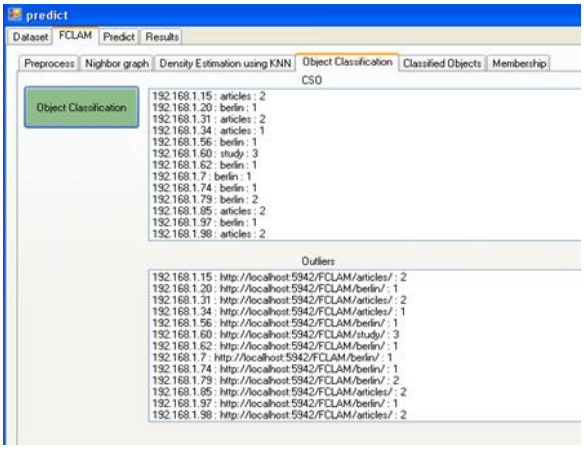

Fig.7 FCLAM Process4: CSO Object Detection in FCLAM

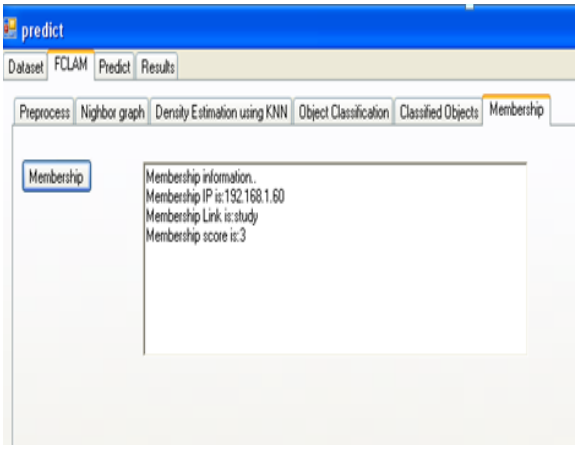

Fig.8 FCLAMProcess5: Membership View

The experimental result for the FCLAM Clustering is given in above figs. [3-8]. Therefore, Fuzzy based FCLAM algorithm can improve the accuracy of Web page prediction with less prediction time. Web page access prediction can be helpful in many applications. The web advertisement area can be changed by improving the accuracy of web page access prediction. Using web page access prediction, the right advertisement will be placed in the website according to the users' browsing patterns. Also, web page access prediction helps the web administrators to restructure the Web site. By predicting the Web page, we can improve the browsing speed and navigation paths.

V. EXPERIMENTAL RESULTS

The performance study of the proposed method PCM-FCLAM is compared with existing techniques namely K-Means clustering. In this study for conducting experiment the experimental phase used three different data sets; they are dataset 1, dataset 2, and Dataset 3. The dataset 1 is collected from the proposed website and the second dataset2 has been used with synthetic dataset.

Performance Analysis In Terms Of Accuracy

The Table 1 and Fig. 9 shows that the proposed PCM-FCLAM performs better for all the three datasets with the highest accuracy of 92.74%, 94.81, and 95.96% than the K-MEANS techniques.

Table 1 Performance analysis in terms of accuracy

Techniques	Data set 1	Data set 2	Data set 3
K-Means	86.03	89.75	91.26
PCM-FCLAM	92.74	94.81	95.96

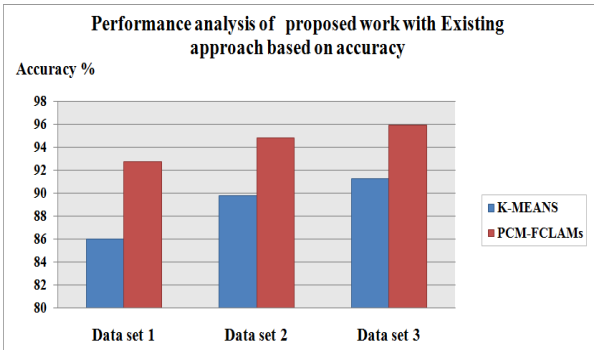

Fig.9 Performance analysis in terms of accuracy

Table 2 Performance analysis in terms of time taken

Techniques	Data set 1	Data set 2	Data set 3
K-Means	20	18	45
PCM-FCLAM	19	18	40

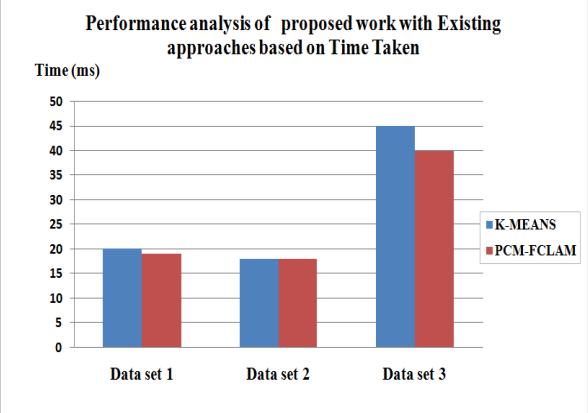

Fig.10 Performance analysis in terms of time taken

From the Table 2 and Figure 10, it is shown that the time taken by the proposed work is very less time taken while comparing the other existing technique.

VI. CONCLUSION

To use the web usage mining efficiently, it is important to use the pre-processing steps. Steps of pre-processing are analyzed and tested successfully with sample web server log files. This paper delivers the steps of pre-processing consist of data cleaning, user identification and session identification and Clustering. User interest analysis is the primary focus of this project and this will learn more about the different stages involved in this mining process and conclude this research report with the results and analysis of the experiment carried out on the web access logs. The experimental result shows that FCLAM clustering performs better than fuzzy clustering in terms of validity accuracy and execution time. FCLAM algorithm is used for predicting the next web page to be accessed in future.

REFERENCES

[1] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. "Web Usage Mining:Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations", Vol.1 (2), pp. 1-12, 2000.

[2] A. J. Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques", *Journal of Theoretical and applied information technology*, 2005.

[3] Baoyao Zhou, Siu Cheung Hui, and Alvis C.M.Fong, "An Effective Approach for Periodic Web Personalization", *Proceedings of the IEEE/ACM International Conference on Web Intelligence. IEEE*, 2006.

[4] G. T. Raju, and P. S. Satyanarayana, "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", *IJCSNS International Journal of Computer Science and Network Security*, Vol.8(1), January 2008.

[5] K. R. Suneetha, and Dr. R. Krishnamoorthi, "Identifying User Behavior by Analyzing Web Server Access Log File", *IJCSNS International Journal of Computer Science and Network Security*, Vol. 9(4), April 2009.

[6] V. Chitraa, and Dr. Antony Selvadoss Thanamani, "A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing", *International Journal of Computer Applications*, Vol. 34(9), 2011.

[7] Albanese, A. Picariello, C. Sansone, and L. Sansone, "A Web Personalization System based on Web Usage Mining Techniques", *in Proc. of WWW2004*, 2004.

[8] P. Makkar, P. Gulati, and Dr. A.K. Sharma, "A Novel Approach for Predicting User Behavior for Improving. 2010.

[9] Mr. Sanjay Bapu Thakare, and Prof. Sangram. Z. Gawali, "A Effective and Complete Preprocessing for Web Usage Mining , *(IJCSE) International Journal on Computer Science and Engineering*, Vol. 2(3), pp. 848-851, 2010.

[10] M. Spilipoulou, and B. Mobasher, B. Berendt, "A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis", *INFORMS Journal on Computing Spring*, 2003

[11] Robert F. Dell, Pablo E.Roman, and Juan D.Velasquez, "Web User Session Reconstruction Using Integer Programming," *IEEE/ACM International Conference on Web Intelligence and Intelligent Agent*, 2008.

[12] V. V. R. Maheswara Rao, and Dr. V. Valli Kumari, "An Analysis and Knowledge Representation System to attain the genuine web user usage Behavior", *International Journal on Computer Science and Engineering (IJCSE)*, Vol. 5(5), pp. 302-316, 2013.

[13] A. Anitha. 2010. "A New Web Usage Mining Approach for Next Page Access Prediction", *International Journal of Computer Applications (0975-8887)*, Vol. 8, no. 11, pp. 4-12.

[14] N. Labroche, "Fast ant-inspired clustering algorithm for web usage mining". *Proceedings of the Information Processing and Management of Uncertainty Conference*. Paris, France, pp. 2668-2675, 2006.

[15] Y. Fu, K. Sandhu, and M. Shih. "A Generalization-Based Approach to Clustering of Web Usage Sessions". *In Proceedings of the 1999 KDD Workshop on Web Mining,San Diego, CA*, vol. 1836 of LNAI,. Springer, pp. 21-38, 2000

[16] Neha Sharma, and Sanjay Kumar Dubey, "Fuzzy C-Means Clustering based Pre-fetching to Reduce Web Traffic", International Journal of Advances in Engineering & Technology, Vol. 6, pp. 426-435, 2013.

