ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

User Future Request Prediction Using KFCM in Web Usage Mining

Dilpreet Kaur[1], A.P. Sukhpreet Kaur[2]

Master of Technology in Computer Science & Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India[1]

Assistant Professor, Department Of Computer Science & Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India[2]

Abstract: Web usage mining is a type of web mining which deals with log files for extracting the information about user browsing behavior. User future request prediction is an approach of web usage mining to predict the next web page for user. In this paper, KFCM method of fuzzy clustering is proposed to predict the user future requests. In this firstly log file data is collected and then preprocessed. After that clustering algorithms FCM and KFCM are implemented to predict the user future requests. The experimental results defining the betterment of KFCM for prediction.

Keywords: Web usage mining, Fuzzy C-Mean, Kernelized Fuzzy C-Mean.

I. INTRODUCTION

Web mining is an application of data mining which uses data mining techniques to extract useful information from web documents. Web mining is further divided into three types Web Usage Mining, Web Content Mining and Web Structure Mining. Web usage mining is a process of mining useful information from server logs. When user use the internet and open different websites then browsing behavior of the user automatically save into log file. Web usage mining deals with these log files for extracting information about user browsing behavior on internet. This information is used in Personalization, Improving the website design, Business intelligence and predicting the user future requests. [2]

User future request prediction is a technique of web usage mining for predicting the next requests of user. For this purpose, web log files are analyzed and user's next requests are predicted according to the earlier related activities. The main use of prediction is for increasing the user browsing speed efficiently, Decreasing the user latency as well as possible, Reducing the loading of web server.[3] This paper uses Fuzzy clustering methods Fuzzy C-Mean and Kernelized Fuzzy C-Mean for clustering.

Main objective of proposed work 'User Future Request Prediction using KFCM in Web Usage Mining' is to predict the browsing behavior of user using fuzzy Clustering methods FCM and KFCM. In this first we collect web log file data and then preprocessing step is performed, by preprocessing irrelevant data is removed and required attributes are selected from log file. After that fuzzy clustering methods are implemented and user future requests are predicted.

The rest of the paper is organized as below section 2 represents the literature review, section 3 represents web log file introduction, section 4 represents introduction to fuzzy clustering, fcm and kfcm algorithm**,** section 5 represents proposed work, section 6 represents results and section 7 represents conclusion and future scope.

II. LITERATURE REVIEW

Yi-Hung Wu and Arbee L.P. Chen in year 2002 [1] present user behavior by sequences of consecutive web page access, derived from access log of a proxy server. The patterns are organized as index. Predictions made based on index. Siriporn Chimphlee, Naomie Salim, Mohd Salihin Bin Ngadiman, Witcha Chimphlee in year 2006 [4]Propose a method for constructing first-order and second-order Markov models of Web site access prediction based on past visitor behavior and association rule mining technique is used for prediction and they show comparison of these three techniques. Christos Makris, Yannis Panagis, Evangelos Theodoridis, and Athanasios Tsakalidis in year 2007 [5]. Proposed a technique for predicting web page usage patterns by modeling users' navigation history using string processing techniques, and validated experimentally the superiority of proposed technique. Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, Ali Mamat in year 2010 [6] Proposed a recommendation system called WebPUM, an online prediction using Web usage mining system for effectively provide online prediction and propose a novel approach for classifying user navigation patterns to predict users' future intentions. Chu-Hui Lee, Yu-lung Lo, Yu-Hsiang Fu in 2011 [7] predicted users browsing behavior and propose two level prediction model using a novel aspect of

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

natural hierarchical property from web log data. V. Sujatha, Punithavalli in 2012 [8] Propose the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data.

III. WEB LOG FILE

Web log file is a file that automatically created and manipulated by the web server. Every hit to the web site include each view of HTML document, image or other document is logged. The raw web log file format is initially one line of text for each hit to the website. This contains information about who was visiting the site, where they came from and exactly what they are doing on the web site. Different server log files have their different formats like extended log file format, Common log file format, Combined log file format etc. [2] log file contains the information like IP Address, User ID, Computer Name, Date, Time, Request Method, Uri Stem, Http Status Code, Size of requested file, Referred webpage, User Agent etc. this information varies from format to format, some log files contain few fields and some contains many fields. Web Log file is a data source of Web Usage Mining.

IV. INTRODUCTION TO FUZZY CLUSTERING, FCM AND KFCM ALGORITHM

Fuzzy Clustering is also known as soft clustering. In this data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. One of the most widely used fuzzy clustering algorithms is the Fuzzy C-Mean. Even though it is better than the hard *K*-means algorithm at avoiding local minima, FCM can still converge to local minima of the squared error criterion. The design of membership functions is the most important problem in fuzzy clustering; different choices include those based on similarity decomposition and centroids of clusters. [9] Fuzzy Clustering methods are Fuzzy C-Means and Kernelized Fuzzy C-Means.

A. Fuzzy C-Means (FCM) Algorithm

Fuzzy C-Mean clustering algorithm is one of the most widely used fuzzy clustering algorithms. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one. After each iteration membership and cluster centers are updated. [10]

B. Kernelized Fuzzy C-Means (KFCM) Algorithm

KFCM is an algorithm which is generated from FCM by modifying the objective function using Kernel induced distance matrix instead of Euclidean distance in FCM. And thus the corresponding algorithm is derived and called as the kernelized fuzzy c-means (KFCM) algorithm, which is more robust than FCM. The main motives of using the kernel methods consist in: (1) inducing a class of robust non- Euclidean distance measures for the original data space to derive new objective functions and thus clustering the non- Euclidean structures in data; (2) enhancing robustness of the original clustering algorithms to noise and outliers, and (3) still retaining computational simplicity. [10]

V. PROPOSED WORK

Proposed work done on the Web Log file. Firstly log file is collected, Then Preprocessing step is implemented which consist of three phases namely Data Cleaning, User Identification and Session Identification. In Data cleaning phase unwanted entries from the log file are deleted and file is arranged into organized structure. In User Identification phase users are identified based on the IP Address. In Session identification phase sessions are identified by taking threshold value of time. After preprocessing fuzzy clustering algorithms are implemented for prediction and results are analyzed. At last user future request is predicted.

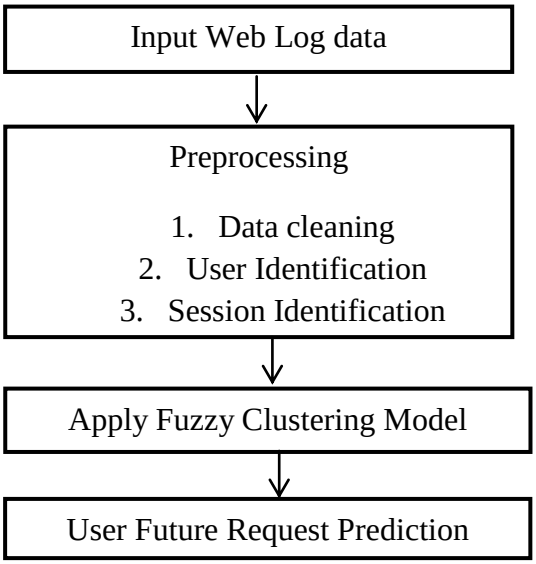

Fig. 1 Flow Diagram of Proposed Work

A. Proposed Algorithm

Step1: Read web log file.

Step2: Preprocessing

(i) Select required attribute from log file like IP Address, User Requests, Request Method, Date, Time, and Status Code.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

(ii) Remove irrelevant entries like all log entries with file name .jpg, .gif, .jpeg, robots files, error code, Request method HEAD, POST.
(iii) Cleaned log file obtained. From cleaned log file identify unique users according to IP Address and unique webpages.
(iv) Session identification step is performed after user identification. In this step sessions are identified for all users by taking 30 minute time threshold value. Pages visited by user less than or equal to 30 minute time put into one session and another pages which are visited after 30 minute put into another session.
(v) Assigning the unique session id to all sessions.

Step3: Clustering

(i) Put the whole data of user session ids and Webpage visited by each user in an array to make clusters.
(ii) Divide the data into clusters using Fuzzy C-Means and Kernelized Fuzzy C-Means algorithms.
(iii) Find the webpages with highest grade of membership in each cluster.

Step4: Prediction

(i) Assign weightage to each webpage according to grade of membership, page with highest weightage has higher membership and page with low weightage has low membership.
(ii) Predicting user future webpage using Fuzzy C-Means and Kernelized Fuzzy C-Means algorithms according to each user in particular session. The webpage which has more weight has more probability for opening in future by user.

VI. RESULTS

In this research first we take log file, then we apply preprocessing step on it and removing the unwanted data to clean the file. Our raw log file has 5991 web requests and after cleaning we obtain 1839 web requests. Our log file has 145 webpages and 270 unique users. After that we apply FCM & KFCM algorithms on it for predicting the users future requests. The results of preprocessing, clustering and prediction are given below, in these results we are taking no. of clusters 40:-
Fig. 2 and 4 shows total no. of clusters and center point of each cluster using fcm and kfcm respectively. Fig. 3 and 5 shows membership of each data point in a cluster. Fig. 6 shows visited webpage according to each user. Fig. 7 shows
Clusters of fcm and kfcm, red clusters are fcm clusters and green clusters are of kfcm clusters. Fig. 8 and 9 shows pie chart of fcm and kfcm clusters. Fig. 10 shows no. of users in clusters and betterment of kfcm, red line is of kfcm which makes equal size clusters as compare to fcm. Fig. 11 and 12 shows the prediction results of fcm and kfcm and result shows that kfcm pick more pages which has highest weightage and highest probability for opening in future by user.

TABLE1 Preprocessing Results

	Before Preprocessing	After Preprocessing
Total no. of Hits	5991	1839
Memory Used	1.25 MB	396 KB

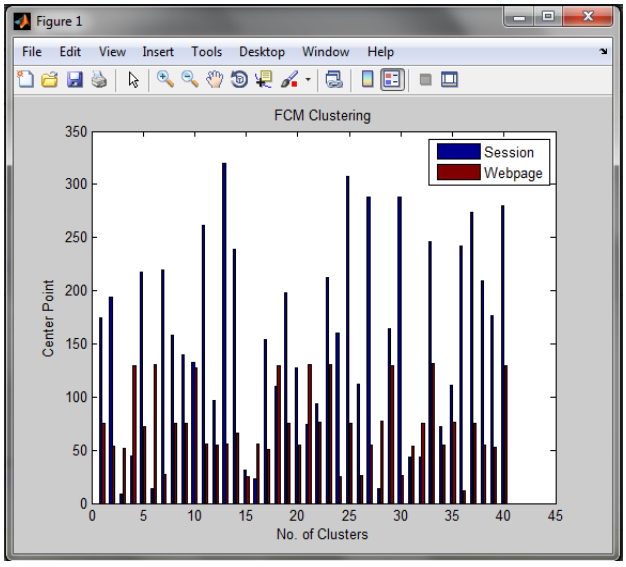

Fig. 2 FCM Clusters Center Point

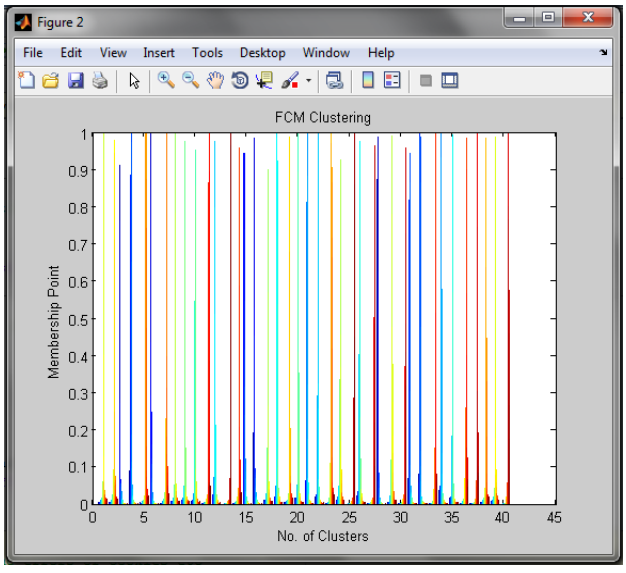

Fig. 3 FCM Membership Point

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

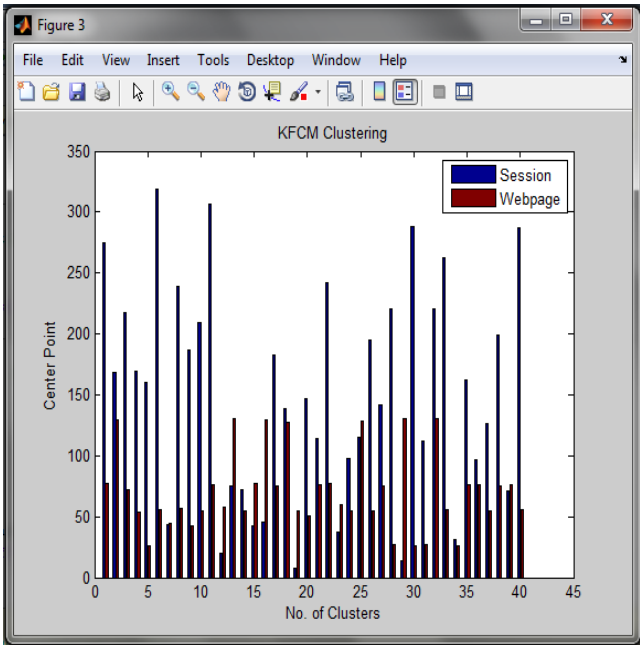

Fig. 4 KFCM clusters Center Point

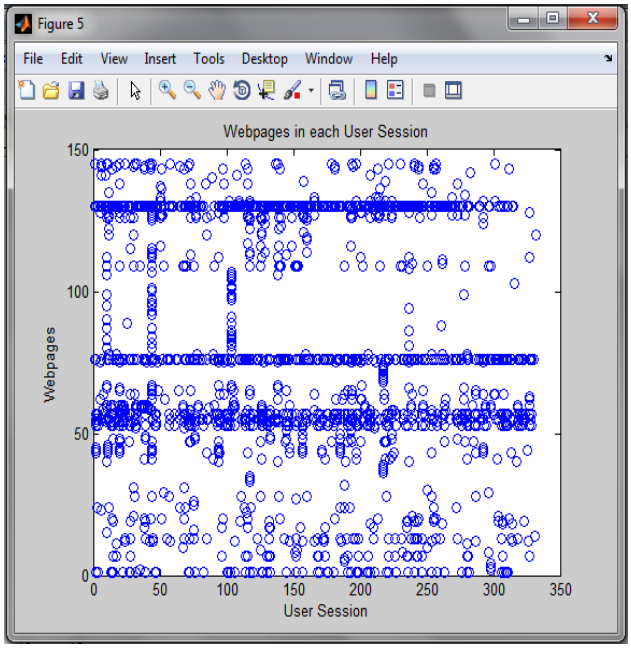

Fig. 6 Webpages according to User Sessions

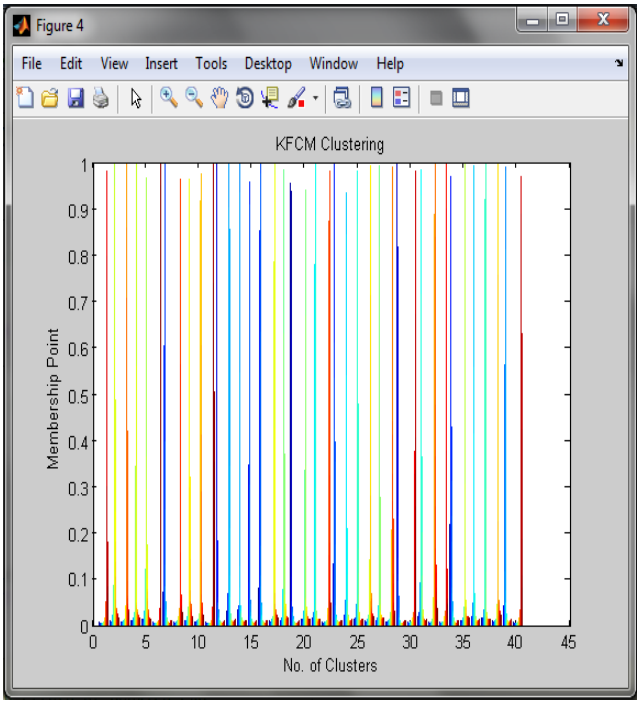

Fig. 5 KFCM Membership Point

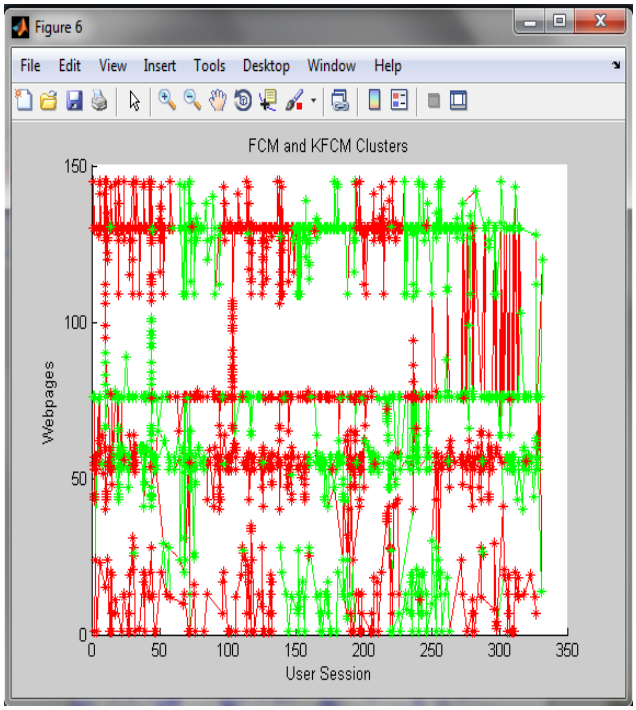

Fig. 7 FCM and KFCM Clusters

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

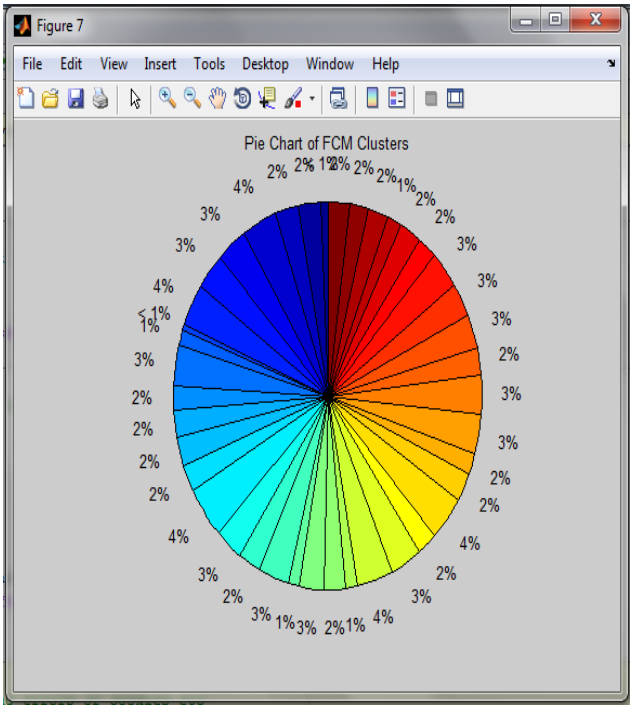

Fig. 8 Pie chart of FC M Clusters

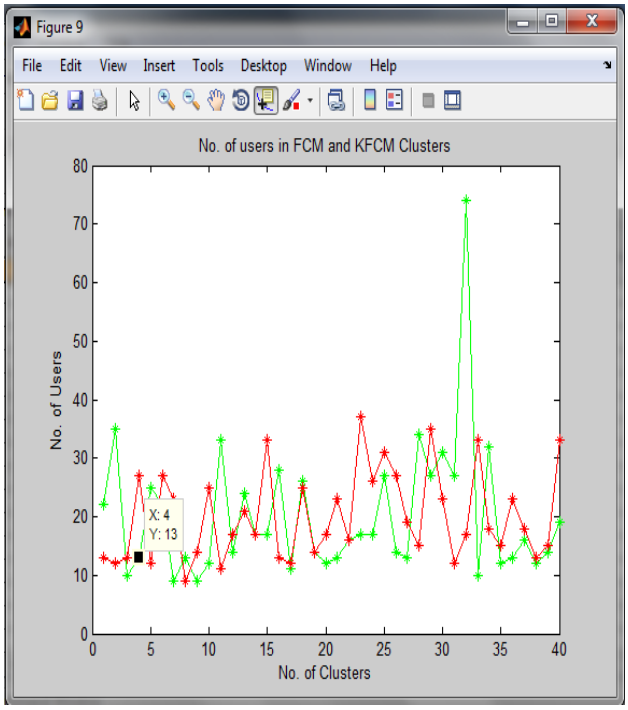

Fig. 10 Number of users in FCM and KFCM clusters

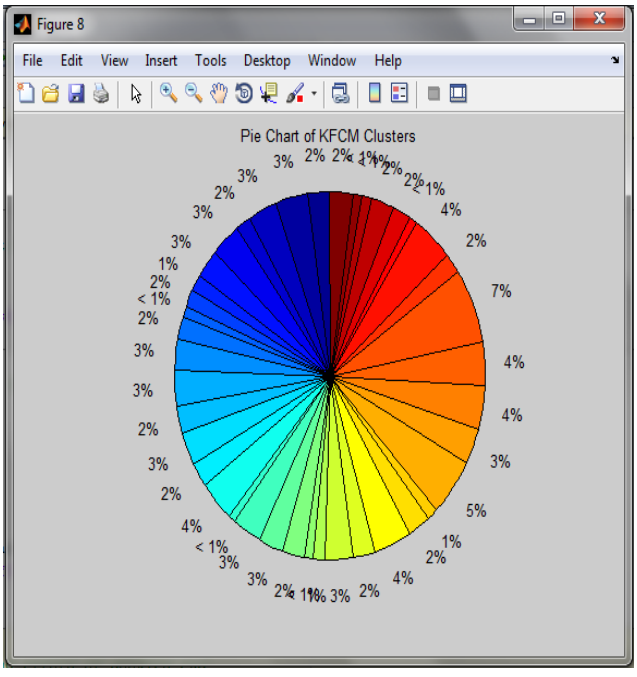

Fig. 9 Pie chart of KFCM Clusters

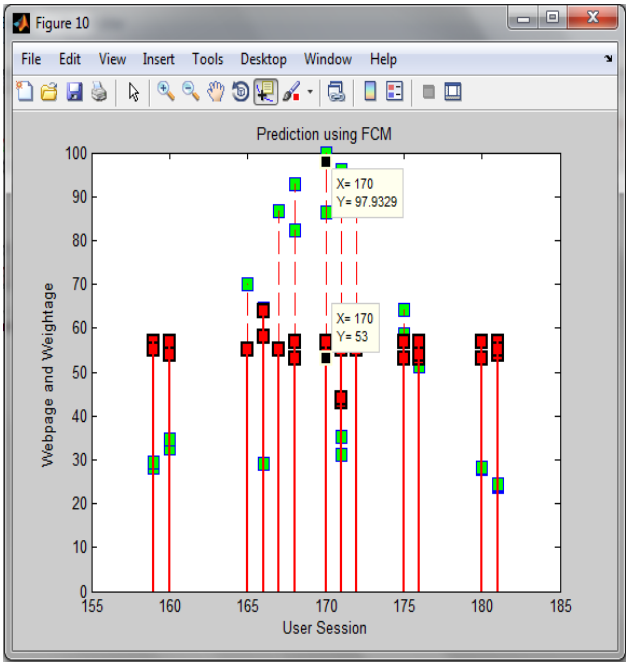

Fig. 11 Prediction of user future webpage using FCM

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 8, August 2013

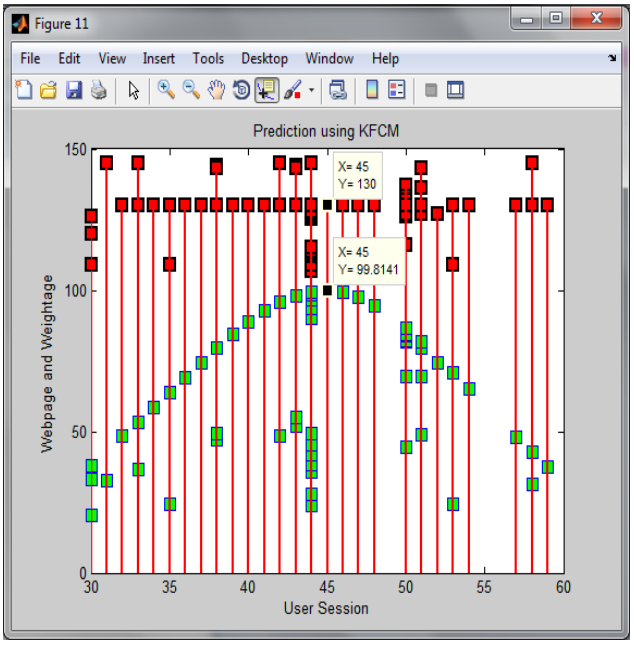

Fig. 12 Prediction of user future webpage using KFCM

VII. CONCLUSION

In Web Usage Mining, user future request prediction has been and still is a significant area of research due to growing popularity of World Wide Web. As internet become popular day by day, there is a heavy traffic on internet and result of heavy traffic is delay in response. To overcome this difficulty User future request prediction is used. In this research work FCM and KFCM algorithms are used for user future request prediction. KFCM is a new approach to user future request prediction. KFCM is better than FCM because KFCM use kernel induced function instead of Euclidean distance function. The results show that KFCM pick maximum data that has highest probability and it makes center point at that place where the data points are more. Thus the clusters of KFCM are better than FCM clusters and prediction is also better. Our prediction is session oriented and page oriented and we make prediction for all the webpages. The result of our proposed work shows that performance of this work is useful for predicting user next page. Our proposed work is useful in prediction we can apply it on web log file which has large data. In future proposed work can apply on different kinds of websites to evaluate its performance and effectiveness and in future we apply it on large data sets.

REFERENCES

[1] Yi-Hung Wu and Arbee L.P. Chen, "Prediction of Web Page Accesses by Proxy Server Log", World Wide Web: Internet and Web Information Systems, 5, 67–88, 2002.
[2] S.K. Pani, et al., "A Survey on Pattern Extraction from Web Logs" IJICA, Volume 1, Issue 1, 2011.
[3] Ujwala Patil and Sachin Pardeshi, "A Survey on User Future Request Prediction: Web Usage Mining" ISSN 2250-2459, Volume 2, Issue 3, 2012.
[4] Siriporn Chimphlee, et al., "Using Association Rules and Markov Model for Predict Next Access on Web Usage Mining", © 2006 Springer.
[5] Christos Makris, et al., "AWeb-Page Usage Prediction Scheme Using Weighted Suffix Trees", © Springer-Verlag Berlin Heidelberg 2007.
[6] Mehrdad Jalali, et al., "WebPUM: A Web-based recommendation system to predict user future movements" Expert Systems with Applications 37, 2010.
[7] Chu-Hui Lee, et al., "A novel prediction model based on hierarchical characteristic of web site", Expert Systems with Applications 38, 2011.
[8] V. Sujatha and Punithavalli, "Improved User Navigation Pattern Prediction Technique from Web Log Data", Procedia Engineering 30, 2012.
[9] http://en.wikipedia.org/wiki/Fuzzy_clustering
[10] Mofreh A. Hogo, "Evaluation of E-learners Behaviour Using Different Fuzzy Clustering Models: A Comparative Study" IJCSIS Vol. 7, No. 2, 2010.
[11] Jaideep Srivastva, et al., "Discovery and Applications of Usage Patterns from Web Data" ACM SIGKDD, Volume 1, Issue 2, 2000.