VOL. 10, NO. 7, APRIL 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

WEB PAGE ACCESS PREDICTION USING FUZZY CLUSTERING BY LOCAL APPROXIMATION MEMBERSHIPS (FLAME) ALGORITHM

P. Sampath and Prabhavathy M.
Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, India
E-Mail: samkala@gmail.com

ABSTRACT

Web page prediction is a technique of web usage mining used to predict the next set of web pages that a user may visit based on the knowledge of previously visited web pages. The World Wide Web (WWW) is a popular and interactive medium for publishing the information. While browsing the web, users are visiting many unwanted pages instead of targeted page. The web usage mining techniques are used to solve that problem by analyzing the web usage patterns for a web site. Clustering is a data mining technique used to identify similar access patterns. If mining is done on those patterns, recommendation accuracy will be improved rather than mining dissimilar access patterns. The discovered patterns can be used for better web page access prediction. Here, two different clustering techniques, namely Fuzzy C-Means (FCM) clustering and FLAME clustering algorithms has been investigated to predict the webpage that will be accessed in the future based on the previous action of browsers behavior. The Performance of FLAME clustering algorithm was found to be better than that of fuzzy C-means, fuzzy K-means algorithms and fuzzy self-organizing maps (SOM). It also improves the user browsing time without compromising prediction accuracy.

Keywords: fuzzy clustering, local approximation, cluster supporting objects.

1. INTRODUCTION

The World Wide Web is one of the main data sources for millions of people in the world in order to access the information from the huge amount of available data. While searching for the particular information on the web, it is important for the user to retrieve the information in less time. The web recommendation models provide access friendliness for users while browsing a website. The prediction models play a vital role in e-commerce, to give advertisement at specific pages of commercial website. Web access prediction is useful in personalization to send personalized web content to specific type of users.

Web mining is a kind of data mining techniques to automatically discover and extract information through the analysis of Web contents, Web structure and Web usages. Predicting the normal users' browsing behavior is one of the web usage mining techniques. Web Usage Mining is the process of extracting useful patterns from the web log files. There are different techniques namely SVM, clustering, Page ranking, Markov model, Modified Markov model, Association rule mining, Markov model with clustering etc. has been used for web page prediction.

Clustering is a process of grouping set of data items into number of clusters, in such a way that maximizes the similarity within clusters and minimizes the similarity between two different clusters. FCM is a data clustering technique in which a dataset is grouped into n clusters with each data point in the dataset belonging to every cluster to a certain degree. FLAME algorithm defines the neighborhood of each object and identifies cluster supporting objects. Fuzzy membership vector for each object was assigned by approximating the memberships of its neighboring objects through an iterative converging process.

The rest of this paper is organized as follows. Section 2 presents the related work on web page prediction domain Section 3 presents detailed description of Fuzzy C-Means clustering. Section 4 presents details about the FLAME Clustering technique. In Section 5 evaluation method for proposed system are stated. Finally in Section 6 we conclude our work.

2. RELATED WORKS

Dilpreet Kaur *et al.* [10] proposed Web Usage Mining technique to predict the browsing behavior of user using fuzzy Clustering methods such as Fuzzy C-Means and Kernalised Fuzzy C-Means. In this, web log file data is collected and then preprocessing step is performed to clean irrelevant data and required attributes are chosen from log file. After that fuzzy clustering methods are implemented and user future requests are predicted.

Neha Sharma *et al.* [9] proposed a framework to support the pre-fetching criteria's on web servers. In this framework, there are basically five steps to guess and pre-fetch the user's request such as Data extraction from web log, data preprocessing, data clustering, Prediction by partial matching and at last pre-fetching according to the Prediction by Partial Matching results obtained.

Khalil *et al.* [2] proposed an improved method by combining Markov models and association rules in order to provide better web page prediction accuracy with high coverage. Low order Markov models are used here to predict multiple pages to be accessed by a user and then association rules are used to predict the page to be accessed by the user next based on the long history.

Kim *et al.* [4] presented a method that combines the association rules, Markov models, sequential association rules and clustering for web page prediction. This paper presented the use of four Web personalization models in order to improve the performance of the system mainly when it comes to precision and recall. Both association rules and sequential association rules methods can use All-Kth order model to increase coverage but this method produces less precision.

Anitha [7] has proposed a new web usage mining approach in order to predict next page access. Pair-wise Nearest Neighbor (PNN) based clustering is proposed to recognize similar access patterns from web server log and then these patterns are used to find out next page access by using sequential pattern mining technique. The clusters are formed by setting similarity threshold. In traditional models, clustering by non-sequential data decreases recommendation accuracy. Mining the web access log of users of similar interest provides good recommendation accuracy.

Meenu Brala *et al.* [19] proposed an intelligent web page prediction approach based on the history of web page visit. The proposed approach is three level approaches in which markov model along with association mining and clustered approach are combined to make efficient web page access prediction.

3. FUZZY CLUSTERING

Fuzzy clustering is a process of categorizing elements such as usage clicks or usage sessions into groups, where each element can belong to several groups with different degrees of membership. Fuzzy clustering is also known as soft clustering. In fuzzy clustering, the data points can belong to more than one cluster, and associated with each of the data points are membership grades which indicate the degree to which the data points belong to the different clusters.

a) Data preprocessing

Pre-processing the web log data is an important and requirement phase in Web mining. The data pre-processing is used to select required features, clean data from irrelevant records and finally convert data into sessions. The steps involved in Data Preprocessing are

Step-1: Read the web log file.

Step-2: Select required parameters from web log file like IP Address/ URL, Date and Time, Request Type of User request, Protocol, Port Number and Page Number & remove other parameters if present.

Step-3: Remove irrelevant entries like robot request.

Step-4: From cleaned log file identify unique users according to IP Address and unique web pages.

Step-5: After user identification, the pages accessed by particular user must be divided into individual session, which is referred to as session identification. Session identification is performed by setting threshold value of 30 minutes.

b) FCM algorithm

Fuzzy C-Means is an algorithm of clustering, which uses the idea of categorizing the data in two or more clusters that belongs to the same group as generated in Fuzzy Logics. The FCM algorithm assigns a membership grade to each data point by calculating distance between the cluster center and the data point. It attempts to divide a finite set of elements X={x_1, x_2, ... , x_n} into a collection of c fuzzy clusters with respect to some specified criterion.

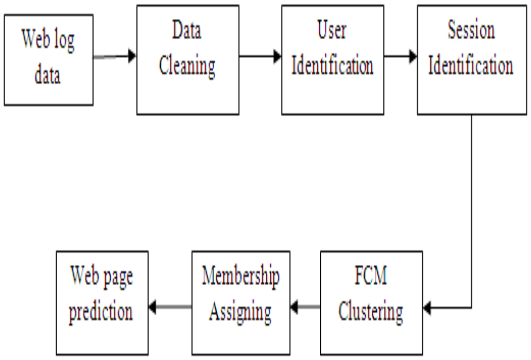

Figure-1. Flow chart of FCM clustering algorithm.

Weblog is preprocessed in order to remove the redundant entries. After cleaning the weblog dataset, user IDs and sessions are identified. The whole datasets of user session IDs and Webpage visited by each user is put in an array to create clusters. Data is divided into clusters using Fuzzy C-Means algorithms. After clustering, identify the Web Pages with the highest grade of membership in each cluster and weight is assigned to each webpage according to the grade of membership. The page with highest weightage has a higher membership grade and page with low weightage has low membership. The page which has more weight has the more probability to access that webpage in future by the user.

FCM is mainly used to analyze the access patterns of the organization. FCM based clustering algorithm provides best results for overlapped data sets comparatively better than that of k-means algorithm. One data point may belong to more than one cluster center.

The main drawback of the FCM algorithm is its sensitivity to noises. It implements the clustering task for a particular data set by reducing an objective-function subject to the probabilistic limit that the summation of all the membership degrees of every data point to all clusters must be one. This restriction results in the problem of this membership assignment.

4. FLAME CLUSTERING ALGORITHM

FLAME is a kind of fuzzy clustering algorithm used to define clusters in the dense parts of a dataset and performs cluster assignment specially based on the neighborhood relationships between the objects. The main feature of this algorithm is that the neighborhood relationships between neighboring objects in the feature space are used to constrain the memberships of neighboring objects in the fuzzy membership space.

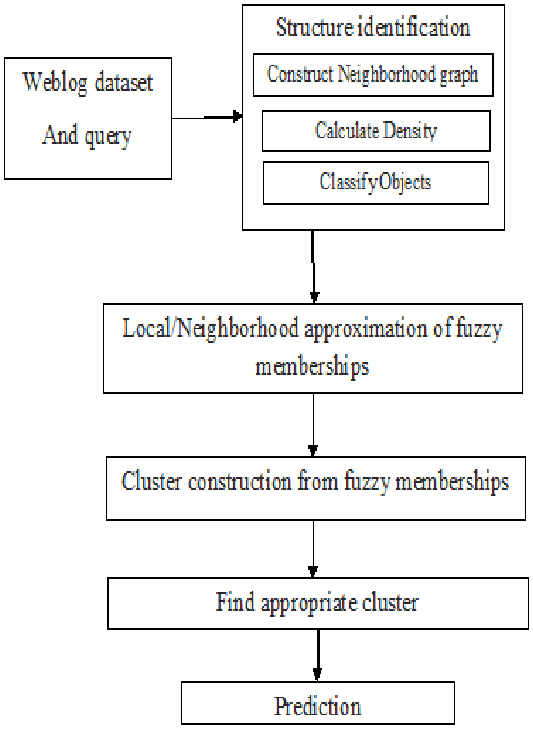
Weblog dataset
And query
Structure identification
Construct Neighborhood graph
Calculate Density
Classify Objects
Local/Neighborhood approximation of fuzzy memberships
Cluster construction from fuzzy memberships
Find appropriate cluster
Prediction

Figure-2. Flowchart of FLAME clustering algorithm.

In this algorithm, weblog data is given as input and the structure information is extracted from the web log data. For this, neighborhood graph is constructed in order to connect each object to its K-Nearest Neighbors (KNN). The similarities between every pair of objects are calculated, and the nearest neighbors are also identified. After constructing the neighborhood graph, density for each object is computed based on its proximities to its KNN. The distance/proximity between each object and its k-nearest neighbors 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 neighboring objects are known as CSO. Object with density lower than all its neighboring objects, and lower than a predefined threshold are known as Cluster outliers.

In fuzzy clustering, each object x is associated with a membership vector $v(x)$, in which each element $v_i(x)$ indicates the membership degree of x in cluster i:

$$x : v(x) = (v_1(x), v_2(x),..., v_M(x)),$$

Where,

$$0 \leq v_i(x) \leq 1; \sum_{i=1}^{M} v_i(x) = 1;$$

$$M = |X_{CSO}| + 1$$

Each element of membership vector takes value between 0 and 1, representing how much percentage an object belonging to a cluster, or being an outlier. By using iterative process of local approximation, membership vector is assigned to each object. The neighborhood relationships are calculated for all objects, and are used to limit the fuzzy memberships.

The clusters are constructed from the fuzzy memberships in two ways: (i) by assigning each object to the cluster in which it has the highest membership degree, or (ii) threshold value is applied on the memberships, and assign 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.

5. EXPERIMENTAL EVALUATION

Web users are facing the problems of information overload due to the significant and rapid growth in the amount of information and the number of users. Weblog files are collected and then preprocessing operation is performed in order to removes the unwanted entries in the weblog files. The weblog file has 3945 web requests and after cleaning we obtain 1233 web requests. Files such as *.jpg, *.gif, *.jpeg are filtered and only requested web pages are recorded. When identifying user sessions, session timeout is set to 30 minutes, with a minimum of 10 page views per session. After filtering out the web session data by preprocessing, the training data set contained 487 records and 154 sessions. After that, FCM & FLAME Clustering algorithms are evaluated in order to predict the webpage that is to be accessed in near future.

Table-1. Prediction accuracy.

Datasets	FCM algorithm	FLAME clustering algorithm
100	0.512	0.639
200	0.501	0.686
300	0.583	0.752
400	0.617	0.822

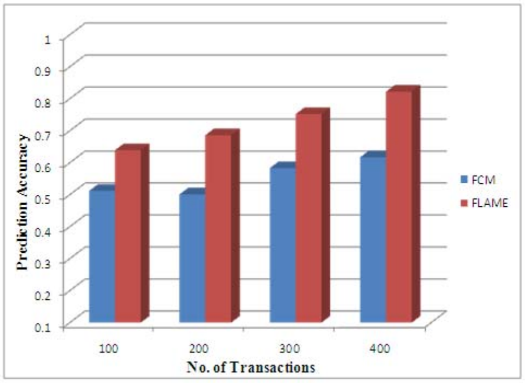

Figure-3. Prediction accuracy.

The experimental result for the FCM and FlAME Clustering are given in Table 1. Therefore, Fuzzy based FLAME 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.

6. CONCLUSIONS

FLAME clustering is helpful to mine complex and multi-dimensional data sets. Among the various developed techniques, FLAME algorithm is used for predicting the next web page to be accessed in future. Its performance was found to be better than that of fuzzy C-means, fuzzy K-means algorithms and fuzzy self-organizing maps (SOM). Prediction made by FLAME algorithm is better than FCM algorithm.

REFERENCES

[1] P.Sampath, Dr. Amitabh Wahi and D. Ramya. 2014. “A Comparative Analysis of Markov Model with Clustering and Association Rule Mining for Better Web Page Prediction”, Journal of Theoretical and Applied Information Technology, Vol.3, pp. 1-5.

[2] Khalil, F., J. Li and H. Wang. 2006. “A framework of combining Markov model with association rules for predicting web page accesses”, Proceedings of the Conference on Data Mining and Analytics, pp. 17-28.

[3] Khalil F., J. Li and H. Wang 2007. “Integrating Markov model with clustering for predicting web page accesses”, Proceedings of the WWW Conference, pp. 1-26.

[4] D. Kim, L Lm, N Adam, V Atluri, M Bieber and Y. Yesha. 2004. “A Clickstream-Based Collaborative Filtering Personalization Model: Towards A Better Performance”, ‘WIDM 04 Conference’, pp. 12-13.

[5] B. D. Gunel and P. Senkul. 2011. “Investigating the Effect of Duration, Page Size and Frequency on Next Page Recommendation with Page Rank Algorithm”, ACM, pp. 122 – 145.

[6] M. Eirinaki and M. Vazirgiannis. 2005. “Usage-based page rank for web personalization”, In Data Mining, IEEE International Conference on, pp. 1-8.

[7] 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.

[8] T.Vijaya Kumar1 and Dr. H. S. Guruprasad. 2014. “Clustering Of Web Usage Data Using Chameleon Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 6.

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

[10] Dilpreet Kaur and A. P. Sukhpreet Kaur. 2013. “User Future Request Prediction Using KFCM in Web Usage Mining”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, pp.121 -139.

[11] M. Awad, L. Khan and B. Thuraisingham. 2008. “Predicting WWW surfing using multiple evidence combination”, VLDB Journal, Vol.17, pp.401–417.

[12] Y. Z. Guo, K. Ramamohanarao and L. Park. 2007. “Personalized page rank for web page prediction based on access time-length and frequency”, In Web Intelligence, IEEE/WIC/ACM International Conference, pp. 687-690.

[13] P. Makkar1, P. Gulati and Dr. A.K. Sharma. 2010. “A Novel Approach for Predicting User Behavior for Improving.