Clustering of Web Access Patterns for Segmenting Web Users Using a Novel Fuzzy Based Cluster Estimation Method

Binu Thomas

Abstract **In this paper we are presenting a method for segmenting the web users based on their web access patterns. The history of web pages visited by users, which includes the access sequence and number of visits of web pages, reveals their interest in particular pages. The web user's access patterns can be segmented to group the users with similar interests. Such segmentation can be used for target marketing, page recommendation and prediction. Segmenting web users with similar interests has got immense possibilities in e-commerce. Here we introduce a simple frequency based technique for preprocessing web access data to convert it into a database with fixed number of attributes. This database is segmented into clusters to group the web users with similar web access patterns. Determining the number of clusters and initial cluster seeds become the most challenging part of Data Clustering. A novel approach proposed by the authors for unsupervised clustering is extended to identify the number of web user groups on the basis on their access patterns. This method starts with the assumption that all the data points are initial clusters and pairs of similar clusters are then merged based on a modified fuzzy membership values. The cluster count obtained with this approach is compared with Kohonen's unsupervised clustering algorithm. The tools available with IBM SPSS Modeler 14.1 are used to benchmark the quality of cluster estimation.**

Keywords Fuzzy Clustering, Web access pattern, Fuzzy Logic, k-means clustering, c-means clustering

1 Introduction

There is a vast amount of information on the World Wide Web (WWW) and more is becoming available daily. With explosive growth of data available with the web and the so called social networking sites, discovery and analysis of useful information from such data becomes a necessity [20]. Extracting useful information from the web data is termed as Web mining. World Wide Web data can be classified in to four groups namely, content, structure, usage and user profile data [3,16]. Based on the primary kinds of data used in the mining process, web mining tasks can be categorized into three types: web structure mining that mines web's linkage structure, web content mining that uses multimedia data on the web and web usage mining or web log mining that mine usage details of web surfers[4,17]. Servers register a Web log entry for every single access they get, in which important pieces of information about accessing are recorded, including the URL requested, the IP address from which the request originated, and a time-stamp. Applying Data Mining techniques on this web log data can reveal many interesting knowledge about the web users[18].

In Data Mining, Cluster analysis is a technique for breaking data down into related components in such a way that patterns and order becomes visible. Clusters are natural groupings of data items based on similarity metrics or probability density models. Clustering algorithms map a new data item into one of several known clusters [1,5]. Membership of a data item in a cluster can be determined by measuring its distance from each cluster center. In crisp clustering, the data item is added to a cluster for which this distance is minimal. In fuzzy clustering techniques, a data item is given partial memberships in all the clusters within a range of membership values from zero to one. A cluster has a center of gravity which is basically the weighted average of the cluster [2]. In this paper we model the access sequence in web logs as frequency values in a fixed database table structure. Also, we use a fuzzy based algorithm for segmenting web users based on their interests and access patterns[15]. We use the server log of individual browsing records of thousands of users at msnbc.com [21].

The remaining of the paper is organized as follows; section 2 introduces the related works in this field. Section 3 is about web access data and section 4 explains k-means clustering. Sections 5 and 6 are about fuzzy clustering and fuzzy c means algorithm. In Section 6 we explain the cluster estimation algorithm and section 8 is a brief introduction to Kohonen's unsupervised clustering algorithm. Section 9 is about the experiments done on Web access data. Finally section 10 concludes the paper.

2 Related Work

I. cadez et.al suggested in their work a new method for visualizing navigation pattern on a web site [6]. The authors presented a simple approach for clustering and visualizing user behaviour on a web site, and implemented the method in a visualization tool called WebCANVAS. They first formed clusters of site users with similar navigation paths using a mixture of first-order Markov models. Then they display the behaviour of a random sample of users in each cluster along with the size of each cluster. An important feature of the model-based clustering used in the proposed work is that learning time scales linearly with sample size. The limitation of the method is its inability to model page visits at the URL level when the number of different page categories that can be requested by a user is small.

In [7], Ajith Abraham presents a novel Hybrid web usage mining method named intelligent miner. The paper proposes a hybrid framework that optimizes a fuzzy clustering algorithm. Fuzzy C-means algorithm is used to identify the number of clusters from the cleaned and pre-processed log files. The clusters are then fed to a Takagi-Sugeno fuzzy inference system to analyze the trend patterns. The if-then rule structures are learned using an iterative learning procedure by an evolutionary algorithm and a back propagation algorithm is used to fine tune the rule parameters. The proposed frame work performs better than the earlier methods for daily trends but for hourly trends its performance is low [7]. The computational complexity of the algorithm is the important disadvantage of i-Miner.

D. Cosic and S. Loncaric presented an unsupervised algorithm for cluster estimation which is a combination of fuzzy k-means algorithm and the fuzzy maximum likely hood estimation [8]. In this work maximum likely hood estimation is used to decide whether to introduce a new cluster center. The authors propose three different methods of cluster estimation for fuzzy c-means algorithm. But these methods are applicable only for segmenting CT scan images.

X. Xiong and K.L. Tan proposed a similarity driven cluster merging method for unsupervised fuzzy clustering. This method starts with over specified number of clusters and then pairs of similar clusters are merged on the basis of similarity driven cluster merging criteria [9]. The cluster merging process in the work is based on a fuzzy similarity metric. This involves the calculation of a merging threshold value each time, which is computationally expensive.

3 Web Access Data

Web usage mining, also known as Web log mining, is process of discovering interesting access patterns of web pages from web access logs. It uses secondary data derived from interactions of users with web: web server logs, proxy server logs, user profiles, user queries, cookies [10].A web server usually registers a web log entry for every access of Web page. It includes URL requested, IP address of the origin of request, and a time stamp. Web log database provide rich information about web dynamics [19].Clustering analysis in web usage mining intends to find the clusters of user, page, or sessions from web log file, where each cluster represents a group of objects with common interest or characteristic. User clustering is designed to find user groups that have common interests based on their behaviors, and it is critical for user community construction [8].

User	Sequence				
1	frontpage	news	travel	travel	
2	news	news	news	news	news
3	frontpage	news	frontpage	news	frontpage
4	news	news			
5	frontpage	news	news	travel	travel
6	news	weather	weather	weather	weather
7	news	health	health	business	business
8	frontpage	sports	sports	sports	weather
9	weather				

Fig.1. A sample access sequence

We model the pieces of Web logs as sequences of events to segment the users on the basis of sequential patterns over a period of time. Each sequence is represented as an ordered list of discrete symbols and each symbol represents one of several possible categories of web pages requested by the user. Let E be a set of events. A Web log piece or (Web) access sequence S = e1,e2,….,en (ei ∈ E) for (1≤ i ≤ n) is a sequence of events, while n is called the length of the access sequence. An access sequence with length n is also called an n-sequence. In an access sequence S, repetition is allowed. Duplicate references to a page in a web access sequence imply back traversals, refreshes or reloads For example, 1,1,2 and 1,2 are two different access sequences, in which 1 and 2 are two events[11].

Figure.1. shows a sample of such sequence. The server logs from msnbc.com for a twenty four hour period typically produce roughly one million such sequences [21]. The focus of the work described in this paper is the problem of exploring and segmenting this type of data set to group the web users based on their interests. Alpha numeric characters are often used to represent the web access sequences for simplicity and mathematical modeling (figure 2)

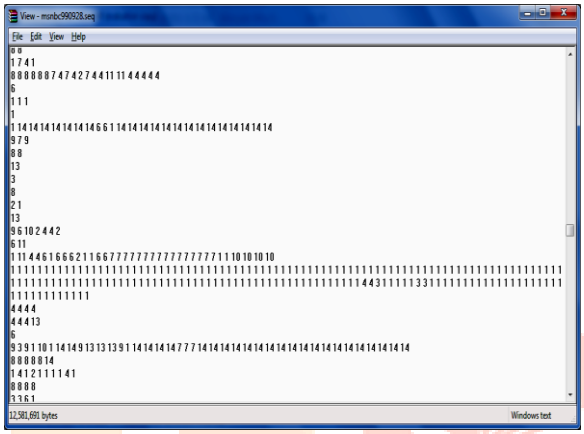

Fig.2. Numeric symbols are used to represent web pages

4 The k-means Clustering Algorithm

K-means algorithm and its different variations are the most well-known and commonly used partitioning methods. The value ‘k’ stands for the number of cluster seeds initially provided for the algorithm. The technique works by computing the distance between a data point and the cluster center to add an item into one of the clusters so that intra-cluster similarity is high but inter-cluster similarity is low. A common method to find the distance is Euclidian distance [1,3]. Finding the optimum value of k is the biggest challenge in this method

5 Fuzzy Clustering

Fuzzy logic is capable of handling the challenges posed by massive collections of natural data. The central idea in fuzzy clustering is the non-unique partitioning of the data in a collection of clusters. The data points are assigned membership values for each of the clusters. In some cases the membership value may be zero indicating that the data point is not a member of the cluster under consideration [12,13].

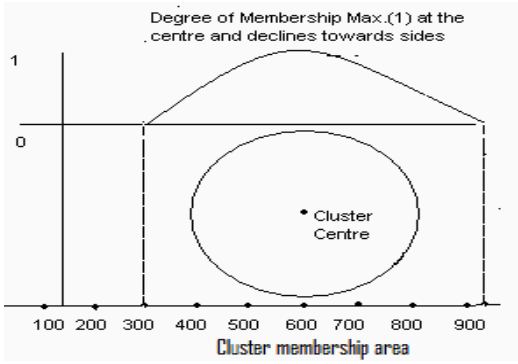

Fig.3. Fuzzy membership in a cluster

The non-zero membership values, with a maximum of one, show the degree to which the data point represents a cluster(figure 3). Thus fuzzy clustering provides a flexible and robust method for handling natural data with vagueness and uncertainty. In fuzzy clustering, each data point will have an associated degree of membership for each cluster [12].

6 C-means fuzzy clustering algorithm

Fuzzy c-means clustering involves two processes: the calculation of cluster centers and the assignment of points to these centers using fuzzy membership values. This process is repeated until the cluster centers stabilize. The algorithm is similar to k-means clustering in many ways but incorporates fuzzy set's concepts of partial membership and forms overlapping clusters to support it. The algorithm needs a fuzzification parameter 'm' in the range [1, n] which determines the degree of fuzziness in the clusters [12]. The algorithm calculates the membership value μ with the formula,

$$\mu_j(x_i) = \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{p}\left(\frac{1}{d_{ki}}\right)^{\frac{1}{m-1}}} \qquad \text{Expression (1)}$$

Where,

$\mu_j(x_i)$: is the membership of x_i in the j^{th} cluster
d_{ji} : is the distance of x_i in cluster c_j
m : is the fuzzification parameter
p : is the number of specified clusters
d_{ki} : is the distance of xi in cluster C_k

The new cluster centers are calculated with these membership values using the equation:

$$c_j = \frac{\sum_i \left[\mu_j(x_i)\right]^m x_i}{\sum_i \left[\mu_j(x_i)\right]^m} \qquad \text{Expression (2)}$$

Where,

C_j : is the center of the j^{th} cluster
x_i : is the i^{th} data point
μ_j : the function which returns the membership
m : is the fuzzification parameter

This is a special form of weighted average. We modify the degree of fuzziness in xi's current membership and multiply this by xi. The product obtained is divided by the sum of the fuzzified membership.

6.1 Limitations of the algorithm

The fuzzy c-means approach to clustering suffers from several constrains that affect the performance. The main drawback is from the restriction that the sum of membership values of a data point x_i in all the clusters must be 'one' as in expression 3, and this tends to give high membership values for the outlier points [14].

$$\sum_{j=1}^{p} \mu_j(x_i) = 1 \qquad \text{Expression (3)}$$

So the algorithm has difficulty in handling outlier points. Secondly, the membership of a data point in a cluster depends directly on its membership values in other cluster centers and this happens to produce unrealistic results[12]. In fuzzy c-means method a point will have partial membership in all the clusters. The third limitation of the algorithm is that due to the influence (partial membership) of all the data members, the cluster centers tend to move towards the center of all the data points[12].

In view of these limitations, a modified c-means algorithm[14] was proposed and this modified c-means method is used for the development of the fuzzy based unsupervised cluster estimation method[15]. The new method is considering all the data points as cluster centers initially, and later these clusters are merged on the basis of fuzzy membership values. Due to the limitation imposed by the expression 3, membership values generated by c-means algorithm is too low with large number of clusters. So we have to modify the algorithm to handle such situations.

In c-means, the membership of a data point in a cluster depends directly on the sum of distances of the point from other cluster centers (expression. 1). Instead, if we consider the sum of distances of data members in a cluster for the calculation of memberships in that cluster, it is improving the performance of the algorithm[14]. This was leading to the modification of the algorithm. The new modified membership function for i^{th} data point in j^{th} cluster is given below,

$$\mu_j(x_i) = n * \frac{\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}}{\sum_{i=1}^{n}\left(\frac{1}{d_{ji}}\right)^{\frac{1}{m-1}}} \qquad \text{Expression(4)}$$

With the new membership method the sum of memberships of a data point in all the clusters now becomes n (number of points). So this expression can handle large number of cluster centers.

7 The new cluster estimation algorithm

The algorithm presented here, works in a single step and initially it assumes that all the data points as cluster centers. This fuzzy based unsupervised clustering algorithm converges to the optimum number of clusters in a single step and it requires only one threshold value [15]. The algorithm does not require the user to provide the number of clusters as an initial parameter and it doesn't also require the user to initialize the cluster seeds based on the general distribution of data. The method then uses one threshold value and it is the cluster center membership threshold (β). It is used to delete (merge) a cluster if it has a membership value greater β than any of the other existing clusters. If a cluster center has a membership value greater than β with any other existing cluster center then it means that it is strongly associated with another cluster and one of the cluster centers can be deleted(both the clusters can be merged). The experiments showed that initializing β to .5 (half of maximum fuzzy membership) produces desired outputs with natural data. This is because, with the new liberalized constrain for the fuzzy membership , we expect the maximum fuzzy membership value (*one*) for each data member in a cluster. A point becomes the member of a cluster if it has at least half of the expected membership in that cluster. The pseudo code for the proposed algorithm is given in table 2. The variables and other data structures used in the algorithm and flow chart are explained in table 1

Table 1 Variables and Data Structures Used in the Algorithm

Sl No	Variable	Purpose
1	m	Fuzzyfication parameter
2	n	Number of data points
3	C_count	Cluster count
4	memb[n][n]	Membership values of points in clusters
5	Sum[n]	Sum of distances in all the clusters
6	Dist(p,i)	Distance of the i^{th} point from the p^{th} cluster
7	W[n]	The data points

In the algorithm, we assume that there are n data points and these points are stored in the array *W[n]*. The algorithm starts with n initial centroids these centroids are stored in the array *C[n]*. The fuzzy memberships of all the points in all the clusters are found using *expression 4*. These membership values are stored in the array *memb[n][n]*. The sums of the distance to all the points from the cluster centers are also calculated. These values are stored in the array *Sum[n]*. It is found that, a centroid situated at the center of a group of points will have minimum sum of distance to other data points. A centroid which is away from a group of points will have maximum sum of distance. The centroids are selected in the descending order of sum of distances for deletion. So the new centroids which are away from the groups of points (clusters) are considered for deletion first. Such centroids are deleted if it has a fuzzy membership of at least β in any other existing clusters. The point will be deleted from the array C[n] and the corresponding values are updated in the array *memb[n][n]*. When we continue cluster deletion process like this, only the centroids situated at the middle of clusters and the extreme outlier points will be remaining. The algorithm ends by finding the natural cluster centers and extreme outliers in the dataset.

When the algorithm terminates the variable *C-count* will have the optimum number of clusters. The array *C[n]* contains the final cluster centers that remain after the merging operation. So the algorithm can finally locate the natural cluster centers and it doesn't require the number of clusters as an initial parameter. The clustering algorithm can be effectively used in the cluster count estimation phase of many conventional clustering algorithms for improving their efficiency.

Table 2. The new clustering algorithm

```
n=The number of data points,
C-count=n
Sum[n]=0,
memb[n][n]=0
C[n]=W[n]  //(initialize all the points as centroids)
initialize β=.5

For p=1 to n

 {

    Update μp(xi) for each data points   applying (Eq. 6.2)

    Find sum[p]      in Cp

 }

Sort C[n] in the descending order of  sum[n]

 For  each cluster center Ci in the descending order of  sum of distances

 {

   If μ(Ci)>= β in any of the remaining cluster centers then

   Delete Ci

   For j=1 to n

   Update memb[i][j]=0

    p=p-1

}

C-count = p

For i=1 to C-count

 Print C[i]

}
```

8 Kohonen's unsupervised clustering algorithm

Self-Organizing Map it as a synonym of Kohonen's Self Organizing Map (SOM). This is also known as Kohonen Neural Networks (Kohonen 1982; Kohonen 2001). The self-organizing map (SOM) network was originally designed for solving problems that involve tasks such as clustering, visualization, and abstraction. Kohonen's SOM networks have been successfully applied as a classification tool to various problem domains. The self-organizing map (SOM) network is a special type of neural network that can learn from complex, multi-dimensional data and transform them into visually

separated clusters. Unlike other neural network approaches, the SOM network performs unsupervised training; that is, during the learning process the processing units in the network adjust their weights primarily based on the lateral feedback connections. The more common approach to neural networks require supervised training of the network (i.e., the network is fed with a set of training cases and the generated output is compared with the known correct output). Deviations from the correct output result in adjustment of the processing units' weights. On the other hand, unsupervised learning does not require the knowledge of target values. The nodes in the network converge to form clusters to represent groups of entities with similar properties. The number and composition of clusters can be visually determined based on the output distribution generated by the training process. The performance of the unsupervised clustering algorithm presented here is compared with the SOM algorithm.

9 The experiment and Evaluation.

The Data we used for the experiment comes from Internet Information server (IIS) logs for msbc.com and news related portions of msn.com for one entire day[16]. Each sequence in the data set corresponds to page views of a user during that day. We selected 1500 random samples. Each event in the sequence corresponds to a request for a page. Requests are recorded only at the level of page category. There are 16 categories of pages and these categories are given numeric codes from 1 to 16.

Table2. A portion of the database created based on frequency count

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	1	2	0	0	0	0	0	0	0	0	0	0	5	7	0
3	0	1	5	4	0	1	4	0	0	0	0	0	2	4	0
0	1	0	0	0	0	0	0	0	3	2	1	2	0	0	0
2	2	2	1	0	0	2	0	0	4	4	0	0	0	0	1
4	2	0	0	0	0	3	2	2	2	0	0	0	0	0	0
3	3	0	0	0	0	0	1	1	2	2	0	0	1	1	2
1	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0
1	0	1	1	2	12	14	0	0	0	0	0	0	0	0	0
4	1	3	2	2	2	0	0	0	0	0	0	0	0	0	0
13	2	1	2	2	3	3	3	3	2	12	0	0	0	0	0
1	1	0	0	0	0	0	2	2	4	0	0	0	2	2	1
0	3	2	2	3	4	0	0	0	0	0	2	2	4	0	0
1	3	2	0	0	0	0	0	0	0	0	0	2	0	0	0
4	3	2	12	0	0	0	0	0	0	0	0	0	0	0	0
1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0

The pages are included into one of these categories based on their content. These categories are front page(1), news(2), technology(3), local(4), opinion(5), on-air(6), miscellaneous(7), weather(8), health(9), living(10), business(11), sports(12), summary(13), bbs(14), and travel(15), msn-news(16)[6]. Although many other information pertaining to the web access are available, we model only the categories of page requests.

To preprocess the access patterns and to convert it into a database table we found the frequency of each web category in every sequence. We created a database table with 16 attributes with one record created for each user session. Then for each user session we calculated the frequency of all web categories and the frequency counts are entered into corresponding column positions. Now each record of this table represents the number of occurrences of different categories of web pages in each user session(Table 3). All the access patterns are brought in to this fixed table structure with frequency count values so that we can apply clustering algorithms to segment users based on access patterns. In the next step we applied the novel cluster estimation algorithm presented here with the newly formed database. The algorithm has converged to *nine* clusters. It shows there are *nine* user groups with similarities in their access patterns. We then used the popular data mining tool *IBM SPSS modeler 14.1* to form and visualize these *nine* clusters.

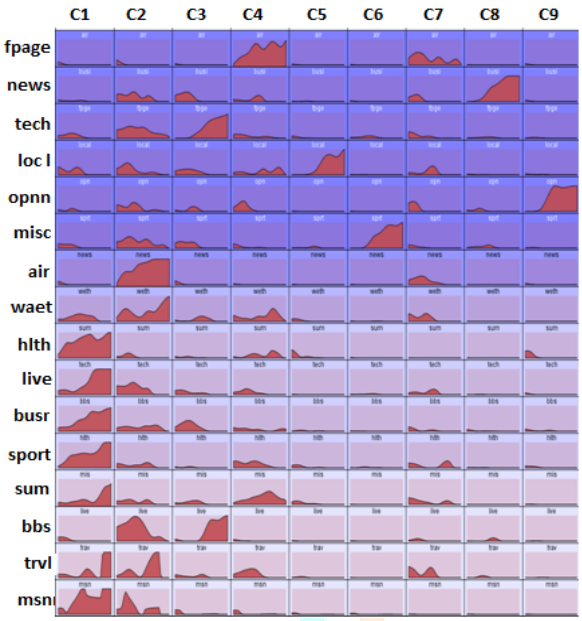

Fig.4. The nine clusters formed using IBM SPSS modeler 14.1

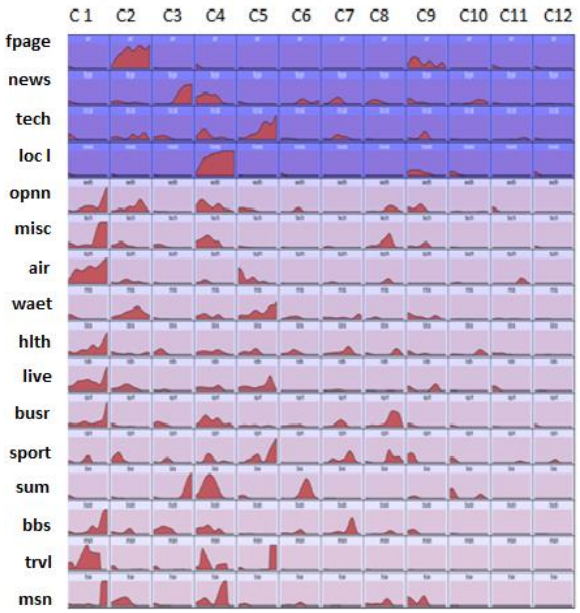

Fig 5. The twelve clusters formed using IBM SPSS modeler 14.1

To compare the quality of the cluster estimation using the new method, we applied Kohonen's algorithm available with *IBM SPSS Modeler 14.1* and the algorithm converged to 12 clusters. This is an unsupervised clustering algorithm which can automatically detect the number of clusters from the data set.

From the above graphical representations (Figures 4 & 5) it can be found that only first *nine* clusters are relevant. The remaining *three* clusters do not have any significant contributions from any of the web pages. The cluster quality of these two segmentations is analyzed using Silhouette measure of cohesion and dispersion. The following graphs show the cluster's quality with *nine* clusters and with *twelve* clusters. These are generated with IBM SPSS Modeler 14.1(figure 6 & 7).

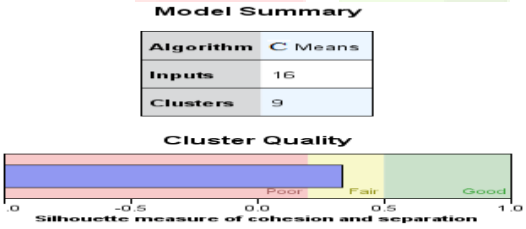

Fig.6. Cluster quality with 9 clusters

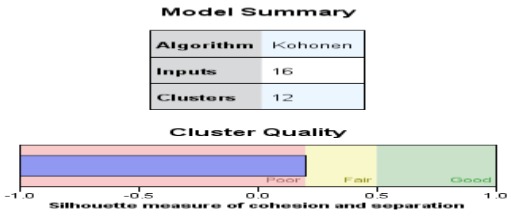

Fig.7. Cluster quality with 12 clusters

So it is evident from the study that the cluster count estimation done by the proposed algorithm is optimum. The algorithm identified *nine* clusters and the cluster quality of the data segmentation with *nine* clusters is much better than that with *twelve* clusters. So the algorithm is capable of identifying natural groups from web access patterns.

10 Conclusion

In web mining, clustering the user sessions should be tackled by exploiting inter-session similarities of web users. The patterns identified from such clustering process can be used for web personalization and community construction. Estimating the number of clusters for modeling natural data is the biggest challenge faced by the traditional supervised clustering algorithms. In this paper, we have applied a novel fuzzy based unsupervised cluster estimation method in segmenting the web access data. It is found that it can detect the natural groupings of web users. These days, devising such clustering techniques is important to find user groups that have common interests based on their behaviors, and it is critical for user community construction. Such knowledge is also useful for inferring user demographics to provide personalized web content to the users.

References

[1] W.H. Inmon, "The Data Warehouse and Data Mining", *ACM Commn* , 1996, 39:.49-50.

[2] P. Berkhin, "Survey of Clustering Data Mining Techniques", Extracted from: *http://citeseer.ist.psu.edu/berkhin02survey.html*.

[3] J Han, M Kamber, *Data Mining Concepts and Techniques*, Elsevier, 2003.

[4] R. Cooley, B. Mobasher, J. Srivastava: Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, 1999, 1(1).

[5] G. Raju, A. Singh, Th. Shanta Kumar, Binu Thomas, Integration of Fuzzy Logic in Data Mining: A comparative Case Study, Proc. of International Conf. on Mathematics and Computer Science, Loyola College, Chennai, 2008, pp.128-136,

[6] I.V. Cadez, C. Meek, Visualization of Navigation Patterns on a Web Site Using Model Based Clustering, *citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.163.1042*.

[7] Ajith Abraham, Business Intelligence from Web Usage Mining, *Journal of Information & Knowledge Management*, 2003, 2(4).

[8] D. Cosic , S. Loncaric, New Methods for Cluster Selection in Unsupervised Fuzzy Clustering, *citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.2642*.

[9] X. Xiong, K.L.Chan, K.L. Tan, Similarity-driven cluster merging method for unsupervised fuzzy clustering, Proceedings of the 20th conference on Uncertainty in artificial intelligence AUAI Press Arlington, Virginia, United States, 2004.

[10] J.Srivastava, R.Cooley, M. Deshpande, and P.-N. Tan. Web Usage Mining: Discover and Applications of Usage Patterns from Web Data. In *ACM SIGKDD Explorations*, 2000, 1(2) , pp 12-23.

[11] J. Pei, J. Han, B. Mortazavi, H. Zhu: Mining Access Patterns Efficiently from Web Logs. Proceedings of the 4th PAKDD, Kyoto, Japan, 2000, pp.396-407.

[12] E. Cox, *Fuzzy Modeling and Genetic Algorithms for Data Mining And Exploration*, Elsevier, 2005.

[13] Sankar K. Pal, P. Mitra, Data Mining in Soft Computing Framework: A Survey, *IEEE transactions on neural networks*, 2002, 13(1).

[14] Binu Thomas, Raju G., and Sonam Wangmo, A Modified Fuzzy C-Means Algorithm for Natural Data Exploration, *www.waset.org/journals/waset/v49/v49-88.pdf*.

[15] B. Thomas and G. Raju, A Fuzzy Threshold Based Modified Clustering Algorithm for Natural Data Exploration, *Lecture Notes in Computer Science*, 2010, 6122, pp. 167-172.

[16] J. Pei, J. Han, B. Mortazavi-asl, and H. Zhu. Mining Access Pattern Efficiently from Web Logs in Proc. 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD'00), Kyoto, Japan, April 2000

[17] M. Spiliopoulou, L. C. Faulstich, K. Winkler, A Data Miner Analyzing the Navigational Behaviour of Web Users in Proc. of the Workshop on Machine Learning in User Modelling of the ACAI'99 Int. Conf., Creta, Greece, July 1999

[18] R.Cooley, B. Mobasher and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web in Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.

[19] B. Mobasher, R.Cooley, and J. Srivastava. Automatic Personalization Based on Web Usage Mining in Communication of ACM, August, 2000 (Volume 43 , Issue 8)

[20] T. Joachims, D. Freitag, T. Mitchell WebWatcher: A Tour Guide for the World Wide Web in IJCAI97 -- Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, pages 770--775, Nagoya, Japan.
[21] *http://kdd.ics.uci.edu/databases/msnbc/ msnbc.html*