Abstract—In this paper we present clustering method is very sensitive to the initial center values, requirements on the data set too high, and cannot handle noisy data. The proposal method is using information entropy to initialize the cluster centers and introduce weighting parameters to adjust the location of cluster centers and noise problems. The navigation datasets which are sequential in nature. Clustering web data is finding the groups which share common interests and behavior by analyzing the data collected in the web servers, this improves clustering on web data efficiently using improved fuzzy c-means (FCM) clustering. Web usage mining is the application of data mining techniques to web log data repositories. It is used in finding the user access patterns from web access log. Web data clusters are formed using on MSNBC web navigation dataset.

Keywords: Datasets, improved FCM clustering, web usage mining.

I. INTRODUCTION

The World Wide Web has huge amount information [1,2] and large datasets are available in databases. Information retrieving on websites is one of possible ways how to extract information from these datasets is to find homogeneous clusters of similar units for the description of the data vector descriptions are usually used each its component corresponds to a variable which can be measured in different scales (nominal, ordinal, or numeric) most of the well known clustering methods are implanted only for numeric data (k-means method) or are too complex for clustering large datasets (such as hierarchical methods based on dissimilarity matrices). Fuzzy clustering relevant for information retrieval. as a document might be relevant to multiple queries, this document should be given in the corresponding response sets, otherwise the user would not be aware of it. Fuzzy clustering seems a natural technique for document categorization there are two basic methods of fuzzy clustering[4], one which is based on fuzzy c-partitions, is called a fuzzy c-means clustering method and the other, based on the fuzzy equivalence relations, is called a fuzzy equivalence clustering method.
given dataset and secondly, given two sets of clusters, computing relative measure of goodness between them. For both these purposes, a criterion function or a validation function is usually applied. The simplest and most widely used cluster optimization function is the sum of squared error \[5\]. Studies on the sum of squared error clustering were focused on the well-known k-Means algorithm \[6\] and its variants. In conventional clustering objects that are similar are allocated to the same cluster while objects that differ are put in different clusters. These clusters are hard clusters. In soft clustering an object may be in more than two or more clusters.

Information gain of feature is calculated as follows. Assume we have \( K \), the set of class label (a binary set in our case: document pairs belonging to the same cluster or not) and \( M_i \), the set of feature values for feature \( i \). With this information, we can calculate the database information entropy, the probabilities are estimated from the relative frequencies in the training set.

\[
H(k) = \sum_{k \in K} P(k) \log_2 P(k)
\]

The information gain of feature \( i \) is then measured by calculating the difference in entropy between the situations with and without the information about the values of the feature.

\[
W_i = H(k) - \sum_{m \in M_i} P(m) \times H(K/m)
\]  

\( W_i \) is a normalized version of information gain. It is information gain divided by \( \text{splitinfo } li(i) \), the entropy of the feature values. This I just the entropy of the database restricted to a single feature.

\[
W_i = \frac{H(k) - \sum_{m \in M_i} P(m) \times H(K/m)}{li(i)}
\]

\( li(i) = - \sum_{m \in M_i} P(m) \log_2 P(m) \)

III. IMPROVED FCM CLUSTERING ALGORITHM.

As the FCM algorithm is very sensitive to the number of cluster centers, cluster centers initialization often artificially get significant errors, and even get the actual opposite results. FCM algorithm[8] is hard on data sets too, so the data sets must be quite regular, in order to solve problems, first of all we use information entropy to initialize the cluster centers to determine the number of cluster centers. It can be reduce some errors, and also can improve the algorithm introductions an weighting parameters. After that, combine with the merger of ideas, and divide the large chumps into small clusters. Then merge various small clusters according to the merger of the conditions, so that you can solve the irregular datasets clustering, document similarity measures.

The algorithm as follows

Initialize number of clusters
Initialize \( C_j \) (cluster centers)
Initialize \( \alpha \) (threshold value)
Repeat
For \( i=1 \) to \( n \); update \( \mu(X_i) \)
  For \( k=1 \) to \( p \):
    \( \text{Sum} = 0 \)
    \( \text{Count} = 0 \)
    For \( i=1 \) to \( n \):
      If \( \mu(X_i) \) is maximum in \( C_k \) then
        If \( \mu(X_i) >= \alpha \)
        \( \text{Sum} = \text{Sum} + X_i \)
        \( \text{Count} = \text{Count} + 1 \)
      \( C_k = \text{sum/count} \)
    Until \( C_j \) estimate stabilize.

The clustering framing as follows

A set clusters \( C = \{C_1, C_2, C_3, \ldots, C_k\} \)

Maximum precision values:

\[
\text{Purity} = \sum_{j=1}^{k} \left( \frac{|C_j|}{n} \right) \text{Max}^n \text{Precision}(C_i, L_j)
\]

\[
\text{Precision}(C_i, L_j) = \frac{|C_i \cap L_j|}{|C_i|}
\]

Inverse Purity:

\[
\text{Inverse Purity} = \sum_{j=1}^{m} \left( \frac{|L_j|}{n} \right) \text{Max}^m \text{Recall}(C_i, L_j)
\]

Recall:

\[
\text{Recall}(C_j, L_i) = \text{Precision}(L_i, C_j)
\]

To calculate their harmonic mean, the F-means

\[
\text{Purity-F} = \sum_{j=1}^{k} \left( \frac{|L_j|}{n} \right) \text{Max}^k \left\{ \text{F}(C_j, L_i) \right\}
\]

Where the maximum is taken over all cluster \( F(C_j, L_i) \) is defined as

\[
F(C_j, L_i) = \frac{2 \times \text{Recall}(C_j, L_i) \times \text{Precision}(C_j, L_i)}{\text{Recall}(C_j, L_i) + \text{Precision}(C_j, L_i)}
\]

IV. WEB USAGE DATA SET

In this section we describe the dataset used and the description of the dataset used for the experimental results.

Information about the dataset

| Description | Value |
|-------------|-------|
| Number of users | 989818 |
| Average number of visits per user | 5.7 |
| Number of URLs per category | 0 to 5000 |
Data format: % Sequence

sequence Order of user page visits
1 1
2 2
3 3 2 2 2 3 3
4 5
5 1
6 6
7 1 1
8 6
9 6 7 7 7 6 6 8 8 8 8
10 6 9 10 3 10 5 10 4 4 10
11 1 1 1 1 1 1
12 1 2
13 1

Description of the Dataset
We collected the data from the UCI dataset repository that consists of sever logs from msnbc.com for the month of September 1998. Each sequence corresponds to page views of a user during that 24 hour period. Each event in the sequence corresponds to a user request for a page. There are 17 page categories "FrontPage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports".

Each category is associated—in order—with an integer starting with "1". For example, "FrontPage" is associated with 1, "news" with 2, and "tech" with 3. Each row below "% Sequences:" describes the hits in order of a single user. For example, the first user hits "FrontPage" twice, and the second user hits "news" once.

The length of the user sessions ranges from 1 to 500 and the average length of session is 5.7.

Similarity matrix with p=0.5
Table 1: Clustering of web usage data

The first similarity upper approximation at threshold value 0.5 is given by

R(T1)={T1,T5,T7,T11,T13}
R(T2)={T2}
R(T3)={T3}
R(T4)={T4}
R(T5)={T1,T5,T7,T11,T13}
R(T6)={T6,T8}
R(T7)={T1,T7,T11,T13}
R(T8)={T6,T8}
R(T9)={T9}

Graphical representation of the clusters found after first upper approximations

This denotes that user visiting the hyper links in t1 may visit the hyperlinks in t5 ad then t7, t11 and hyper links in t13.

One who visits the hyper links in t6 may also visit the hyper links in t8.

S1={T1,T5,T7,T11,T13}
S2={T2}, S3={T3}, S4={T4}, S5={T1,T5,T7,T11,T13}, S6={T6,T8}, S7={T1,T5,T7,T11,T13}, S8={T6,T8}, S9={T9}, S10={T10}, S11={T1,T5,T7,T11,T13}, S12={T12}, S13={T1,T5,T7, T11,T13}.
V. CONCLUSIONS:

In this paper we presented a clustering web usage data from msnbc.com which is useful in finding the user access patterns and the order of visits of the hyperlinks of the each user. The suggested approach was used for efficacy contained a hard clustering of the msnbc.com data set and as the analysis indicated each of the clusters seems to contain observations with specific common characteristics and improve the algorithm efficiency. Experiments prove the improved algorithm has been able to identify the initial cluster centers.

REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, PN. Tan, Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations, Vol. 1, No. 2, 2000, pp.12–23.

[2] M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim Data mining and the web: past, present and future, In Proc. of the second international workshop on webinformation and data management, ACM, 1999.

[3] A. K. Jain and R. C. Dubes, “Data clustering: A review,” ACM Computing Surveys, vol. 31, 1999.

[4] U. Maulik and S. Bandyopadhyay, “Genetic algorithm based clustering technique,” Pattern Recognition, vol. 33, pp. 1455–1465, 2000.

[5] P. Zhang, X. Wang, and P. X. Song, “Clustering categorical data based on distance vectors,” The Journal of the American Statistical Association, vol. 101, no. 473, pp. 355–367, 2006.

[6] A. Vakali, J. Pokorný and T. Dalamagas, An Overview of Web Data Clustering Practices, EDBT Workshops, 2004, pp. 597-606.

[7] Lin Zhu, Fu-Lai Chung, Shitong Wang. Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions[J].IEEE Transactions on Systems,2009:39-3.

[8] Cheul Hwang, Frank Chung-Hoon Rhee. Uncertain Fuzzy Clustering: Interval Type-2 Fuzzy Approach to C-Means[J]. IEEE Transactions on Fuzzy Systems,2007:15-1.