Clustering Algorithm in Data Mining Based on Web Log

Liu Jing

Qilu University Of Technology, Jinan, Shandong,250353,China, 183626231@qq.com

***Abstract* - The advantages of FCM algorithm are that it is mainly applied in point data cluster and cannot directly process relational data, for which the paper proposes a clustering algorithm in data mining based on web log. Firstly, the paper improves FCM algorithm which makes it process relational data, and makes robustness improvement on the algorithm. Then, the traditional FCM algorithm needs to determine in advance on the basis of no prior knowledge, for which the paper introduces competition agglomerative algorithm and makes it combine with FCM algorithm, which generates CA-FCM algorithm making it automatically determine category number of the best classification. The experiments show that mining results of CA-FCM algorithm is close to the mining results of FCM algorithm, and the performance of CA-FCM algorithm is better than that of FCM algorithm when the number of users access to session is not large.**

Keywords - Web log, data mining, clustering algorithm, FCM algorithm

I. INTRODUCTION

As a rich source of information, Web has gradually entered all aspects of people's study, work and life. With the increasing complexity of Web structures and the increasing hugeness of information, it has become more and more difficult for users to

In the Web log mining，the first task is the Web log data preparation. The data preparation is the data mining earlier period essential work. It is the prerequisite to data mining algorithm that could provide effectively input and gain the valuable mining result. The traditional Web log data preparation can't eliminate the influence of frame page，and remove the logs of picture's link simply. So it can't recurrence the user access scene better and cause mining result interest to reduce. Therefore，this paper proposes an approach to restructure the Web site structure and proposes a data preparation which based on the restructured Web site structure. The experiment indicates，uses the approach can even better recurring user access scene and provides effective data for Web log mining.

The mining algorithm includes cluster analysis, association rule, classification and prediction and so on. This paper's target is obtaining user access clusters. So it adopts cluster technology.

As the main channel for information release on the internet, Web not only has shown huge commercial value and application potential, but also the popular and important means of people acquiring information [1]. But the change of it is huge, diversified and dynamic. With the increase of the scale and complexity of Web sits, design and maintenance of sits has become more and more difficult [2]. Website designers try their best to optimize their own website to attract and retain more users, but it must depend on full mastery of website information [3]. Websites operators not only need good aides design, but also need to adjust web page structure dynamically according to the users' access interest, access frequency and access time, and should improve the service for better meet the needs of visitors [4]. And the visitors hope to use the simplest way to get the most accurate information and expect to personalization service. However, a useful tool to slove these needs is Web data mining, that is, the ideas and methods of data mining are used to mine useful information on Web [5].

The paper analyzes and summarizes the experience and achievements of relevant researchers. And uses fuzzy $c-$Means clustering algorithm to mine Web log to realize the access to page clustering for the users. But fuzzy $c-$Means not only has the initial value which is difficult to determine, and can't directly process relational data, but also is very sensitive to isolated points. For the above disadvantages of fuzzy $c-$Means clustering algorithm, we propose a improved FCM clustering algorithm based on Web log mining to find similar customers and page clustering, which provides basis for adjusting website structure and personalized service.

II. IMPROVEMENT OF FCM ALGORITHM

A. Improvement on Robustness of Algorithm

In the set of user session and user visiting pages, there may be some cases as follows: user browsing behavior represented by them is objectless navigation on internet. As for other user sessions with the purpose of access, it can be regarded as noise data.

FCM algorithm demands that membership matrix

$U=[u_{ik}]_{i=1,2,\cdots,c,k=1,2,\cdots n}$ must meet conditions $\sum_{i=1}^{c}u_{ik}=1$, $k=1,2,\cdots,n$, and the sum of membership of each cluster attaching to noise data must be 1, which will obviously affects the accuracy of cluster, the reason for which is that the sum of membership of noise data in each cluster should be very small.

Noise clustering proposed by Dave in 1991 can be used to process noise data. And Dave proposes that noise data is included in a single noise class which makes noise data separate from other data and can't cause to reduce the quality of clustering analysis. Dave defined noise prototype as the representative of noise class, the distance from noise center to all data objects is equal, which is called noise distance). Although Dave pointed out that the noise distance can takes different values later, it is regulated as constant δ here.

c is used to represent the number of good cluster which is relative to noise data, and a cluster is added, that is, the class $(c+1)$ is used to represent noise clustering, and the distance from data object x_k to noise center is $d_{c+ik}=\delta$, the membership attaching to noise clustering is represented as $u_{c+ik}=1-\sum_{i=1}^{c}u_{ik}$, $k=1,2,\cdots,n$. And the constraint condition of membership for good cluster $\sum_{i=1}^{c}u_{ik}=1$, $k=1,2,\cdots,n$ can be changed into $\sum_{i=1}^{c}u_{ik}\le 1$, $k=1,2,\cdots,n$, which makes the sum of membership for noise data in good cluster is arbitrarily small. And the objective functions of FCM algorithm is changed into:

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(d_{ik})^2+\sum_{k=1}^{n}(u_{c+ik})^m\delta^2 \quad (1)$$

$(u_{c+ik})^m=\left(1-\sum_{i=1}^{c}u_{ik}\right)^m$, δ^2 is a constant set by the user which represents the squared distance of each data object to noise center.

Its significance is that if the squared distance of a data object to any good cluster is greater than δ^2, the data objects is seen noise data, and the degree attaching to noise clustering is the biggest.

FCM algorithm needs to change as follows:

In step (1), the value of noise clustering c and noise distance δ are determined, and the initialization membership matrix is $U=[u_{ik}]_{i=1,2,\cdots,c,k=1,2,\cdots,n}$.

In step 2, $d_{c+ik}^{(t+1)}=\delta$ and $k=1,2,\cdots,n$.

In step 3, $c+1$ replaces c:

$$u_{ik}^{(t)}=\frac{\left[\frac{1}{\left(d_{ik}^{(t+1)}\right)^2}\right]^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left[\frac{1}{\left(d_{jk}^{(t+1)}\right)^2}\right]^{\frac{1}{(m-1)}}+\left(\frac{1}{\delta^2}\right)^{\frac{1}{(m-1)}}} \quad (2)$$

And the membership of data objects attaching to noise clustering can be calculated:

$$u_{c+ik}^{(t)}=\frac{\left[\frac{1}{\delta^2}\right]^{\frac{1}{(m-1)}}}{\sum_{j=1}^{c}\left[\frac{1}{\left(d_{jk}^{(t+1)}\right)^2}\right]^{\frac{1}{(m-1)}}+\left[\frac{1}{\delta^2}\right]^{\frac{1}{(m-1)}}} \quad (3)$$

The selection of δ is a very complicated problem, which needs to estimate the size of each cluster. However, in the actual problem, noise distance can take the follow constant:

$$\delta^2=\lambda\left[\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(d_{ik})^2}{cn}\right] \quad (4)$$

λ is proportionality coefficient which needs to be selected according to the type of the data to be clustered, and it reduces gradually with the algorithm.

It is more direct for noise clustering metho to apply to objective data compared with relational data, the reason for which is that there is clustering center in relational data cluster in the strict sense, so there is no noise center. In relational cluster, Noise clustering is defined as the cluster with the same dissimilarity degree between any two data objects, that is, * is used to replace noise clustering, $(R_{jk})_*=\delta(1\le j\le n,1\le k\le n)$, in which δ is constant and is called dissimilarity degree of noise. According to the definition, the dissimilarity degree between any two data objects in noise clustering is δ.Therefore, the dissimilarity degree between any two data objects in a good cluster can't be greater than δ.

Obviously, if the dissimilarity degree between the data objects of each good cluster and some data object of the data concentration to be clustered is greater tha δ, the data object should be divided into noise clustering. On the other hand, if the dissimilarity degree between some data object and at least one data object in a good cluster is less than δ, the data object should belong to the good cluster and not belong to noise clustering.

Noise clustering (NC) is applied in FCM algorithm of relational data, and the objective function should be

changed into:

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(d_{ik})^2+\sum_{k=1}^{n}(u_{*k})^m(d_{*k})^2 \quad (5)$$

In the formula, * means noise clustering. Hathaway deduced $d_{*k}^2=\frac{1}{2}\delta$, so FCM algorithm of relational data needs to do the following changes:

In step (1), the value of noise clustering c and noise distance δ are determined, and the initialization membership matrix is $U=[u_{ik}]_{i=1,2,\cdots,c,k=1,2,\cdots,n}$.

In step 2, $d_{c+ik}^{(t+1)}=\delta$ and $k=1,2,\cdots,n$.

In step 3, the following formula replaces, and the membership of data object attaching to good cluster is calculated as follows:

$$U_{ik}^{(t)}=\frac{\left[\frac{1}{\left(d_{ik}^{(t+1)}\right)^2}\right]^{\frac{1}{m-1}}}{\sum_{j=1}^{c}\left[\frac{1}{\left(d_{jk}^{(t+1)}\right)^2}\right]^{\frac{1}{(m-1)}}+\left(\frac{2}{\delta}\right)^{\frac{1}{(m-1)}}} \quad (6)$$

And the membership of data object attaching to noise clustering can be calculated:

$$U_{*k}^{(t)}=\frac{\left[\frac{2}{\delta}\right]^{\frac{1}{(m-1)}}}{\sum_{j=1}^{c}\left[\frac{1}{\left(d_{jk}^{(t+1)}\right)^2}\right]^{\frac{1}{(m-1)}}+\left[\frac{2}{\delta}\right]^{\frac{1}{(m-1)}}} \quad (7)$$

The above changes and the introduction of a new noise clustering can form FCM algorithm of strong relational data which can process the data set including noise. And the paper call it as strong FCM algorithm of relational data.

B. Realization of CA-FCM Algorithm

In the CA-FCM algorithm of robust relational data, the objective function is added a feature, which makes it as:

$$J=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m(d_{ik})^m+\sum_{k=1}^{n}(u_{*k})^m(d_{*k})^m-a\left[\sum_{i=1}^{c}\left(\sum_{k=1}^{n}u_{ik}\right)^m+\left(\sum_{k=1}^{n}u_{*k}\right)^m\right] \quad (8)$$

as a new objective function, the iterative method is used to evaluate the approximate value of the minimum. And the steps of algorithm are as follows:

(1) Initialization. $m=2$, and the objective function is:

$$J=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^2(d_{ik})^2+\sum_{k=1}^{n}(u_{*k})^2(d_{*k})^2-a\left[\sum_{i=1}^{c}\left(\sum_{k=1}^{n}u_{ik}\right)^2+\left(\sum_{k=1}^{n}u_{*k}\right)^2\right] \quad (9)$$

In the formula, * represents noise clustering, the iteration counter t is 0, and c value, η_0 value, τ value and λ value of the maximum clustering category number are determined, the membership matrix $P^{(t)}$ is initialized. The minimal threshold of class base is n_ε, iterative stop threshold value is ε. $\beta=0$, and R_β is the original dissimilarity matrix $R=[R_{jk}]_{n\times n}$. $P^{(t)}$ meets the following formulas:

$\sum_{i=1}^{c}u_{ik}+u_{*k}=1$, $k=1,2,\cdots,n$

$0\le u_{ik}\le 1$, $i=1,2\cdots,c$, $k=1,2,\cdots,n$

$0\le u_{*k}\le 1$, $k=1,2,\cdots,n$

Calculating the cardinality of the cluster i:

$$n_i^{(t+1)}=\sum_{k=1}^{n}u_{ik}^{(t+1)},\quad i=1,2\cdots,c \quad (9\text{-A, }36)$$

Calculating the cardinality of noise cluster:

$$n_*^{(t+1)}=\sum_{k=1}^{n}u_{*k}^{(t+1)}$$

Updating the distance, and calculating the new membership vector and the distance between data objects and cluster:

$$p_i^{(t+1)}=\frac{\left(\left(u_{i1}^{(t+1)}\right)^2,\cdots,\left(u_{in}^{(t+1)}\right)^2\right)^T}{\sum_{k=1}^{n}\left(u_{ik}^{(t+1)}\right)^2},\quad 1\le i\le c \quad (10)$$

$$\left(d_{ik}^{(t+1)}\right)^2=\left(R_\beta p_i^{(t+1)}\right)_k-\frac{1}{2}\left(p_i^{(t+1)}\right)^T R_\beta\left(p_i^{(t+1)}\right)$$

$1\le i\le c$, $1\le k\le n$ (38)

As for any i and k, if $\left(d_{ik}^{(t+1)}\right)^2<0$, calculating:

$$\Delta\beta=\max\left\{\frac{-2\left(d_{ik}^{(t+1)}\right)^2}{\left\|p_i^{(t+1)}-e_k\right\|^2}\right\} \quad (11)$$

$$\left(d_{ik}^{(t+1)}\right)^2=\left(d_{ik}^{(t+1)}\right)^2+\left(\frac{\Delta\beta}{2}\right)\left\|p_i^{(t+1)}-e_k\right\|^2 \quad (12)$$

$\beta=\beta+\Delta\beta$

$$\left(R_\beta\right)_{jk}=\begin{cases}R_{jk}+\beta, & j\ne k\\ 0, & j=k\end{cases} \quad (13)$$

e_k i n the above formula represents the vector of the k column of unit matrix in R^n.

Determining the distance δ from data object to noise clustering:

$$\delta^2 = \lambda \left[\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}\left(d_{ik}^{(t+1)}\right)^2}{c \times n} \right] \quad (14)$$

$$\left(d_{*k}^{(t+1)}\right)^2 = \frac{1}{2}\delta \quad (15)$$

Updating $a(t+1)$ and membership matrix:

$$\eta(t+1) = \eta_0 e^{-(t+1)/\tau} \quad (16)$$

$$a(t+1) = \eta(t+1)\frac{\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^2(d_{ik})^2 + \sum_{k=1}^{n}(u_{*k})^2(d_{*k})^2}{\sum_{i=1}^{c}\left(\sum_{k=1}^{n}u_{ik}\right)^2 + \left[\sum_{k=1}^{n}u_{*k}\right]^2} \quad (17)$$

$k = 1,2,\cdots,n$ defines the following set:

$$I_k = \left\{ i \middle| 1 \le i \le c; \left(d_{ik}^{(t+1)}\right)^2 = 0 \right\} \quad (18)$$

$$\overline{I_k} = \{1,2,\cdots,c\} - I_k \quad (19)$$

If $u_{ik}^{(t)} > 1 \Rightarrow u_{ik}^{(t)} = 1$ and $u_{ik}^{(t)} < 0 \Rightarrow u_{ik}^{(t)} = 0$,

$u_{*k}^{(t)} = \frac{\frac{2}{\delta}}{\sum_{j=1}^{c}\frac{1}{\left(d_{jk}^{(t+1)}\right)^2} + \frac{2}{\delta}} + \frac{a(t+1)}{\frac{\delta}{2}}\left(n_*^{(t+1)} - \overline{n^{(t+1)}}\right)$。

If $u_{*k}^{(t)} > 1 \Rightarrow u_{*k}^{(t)} = 1$,and $u_{*k}^{(t)} < 0 \Rightarrow u_{*k}^{(t)} = 0$,

If I_k is not null set $\Rightarrow \begin{cases} u_{ik}^{(t)} = 0, \forall i \in \overline{I_k} \\ \sum_{i \in I_k} u_{ik}^{(t)} = 1 \end{cases}$

Updating clustering number and calculating clustering cardinality:

$$n_i^{(t)} = \sum_{k=1}^{n} u_{ik}^{(t)}, \quad i = 1,2,\cdots,c \quad (20)$$

Calculating the cardinality of noise clustering:

$$n_*^{(t)} = \sum_{k=1}^{n} u_{*k}^{(t)} \quad (21)$$

If $n_i^{(t)} < n_\varepsilon$ and $i = 1,2,\cdots,c$, the cluster i is discarded and the clustering category number c is updated.

Checking if it is convergence. If the clustering category number c doesn't change and $\left\| P^{(t)} - P^{(t+1)} \right\| < \varepsilon$, the calculation stops, or $t = t+1$ and returning to the second step.

III. SIMULATION EXPERIMENT OF ALGORITHM

The paper uses Visual C++ 6.0 to realize the above-mentioned similarity algorithm of user access, and uses some data of data table session_20071215 as experimental data to get the similarity of user sessions. The results are represented by similarity matrix to be as the input of user session clustering algorithm.

Because of the limitation of the space, Table 1 only gives partial similarity matrix. S1,S2 ,… ,S8 in the table are the number of user sessions. Because the similarity matrix is symmetric matrix, the table only gives the upper triangular matrix.

TABLE 1 SIMILARITY MATRIX OF USERS VISITING SESSIONS

Session	S1	S2	S3	S4	S5	S6	S7	S8
S1	1.0000	0.1482	0.3162	0.2225	0.2236	0.1933	0.1491	0.2000
S2		1.0000	0.0000	0.2747	0.0000	0.2387	0.1159	0.0000
S3			1.0000	0.0000	0.3536	0.0000	0.2357	0.3162
S4				1.0000	0.0000	0.3584	0.1741	0.0000
S5					1.0000	0.0000	0.1667	0.2236
S6						1.0000	0.1512	0.0000
S7							1.0000	0.1491
S8								1.0000

According to the membership matrix of the fourth iteration, each user visiting session is assigned to the cluster with the maximum membership, and 7 clusters are achieved, as shown in Table 2.

TABLE 2 SUMMARY OF CLUSTERING RESULTS

Cluster	Sessions belonging to the cluster	Number of sessions
C1	S4, S6, S11, S12, S26, S35	6
C2	S1, S5,S7,S14 S16, S17, S25, S36, S39,S43,S44,S46,S49	13
C3	S1, S9 , S10, S21, S27,S31, S33,S41,S50	9
C4	S3, S15, S38	3
C5	S1,S5, S6, S7, S13, S18, S19, S30, S34,S40,S43,S45,S47,S51,S52	15
C6	S5, S22	2
C7	S2, S8, S20, S23,S24,S28,S29,S32,S37,S42	10

In the above table, most weights of URL in the cluster C4 is less than 0.2, which can be judged as noise cluster.

Respectively calculating the intra-class similarity of each cluster S_{intra} and similarity between classes S_{inter}, as shown in Table 3.

TABLE 3 INTRA-CLASS SIMILARITY AND SIMILARITY BETWEEN CLASSES FOR EACH CLUSTER

Cluster	Intra-class similarity	Similarity between classes
C1	0.427	0.079
C2	0.213	0.061
C3	0.335	0.032
C4	0.056	0.042
C5	0.204	0.093
C6	0.372	0.041
C7	0.162	0.084
Mean	0.253	0.072

From the calculation results of intra-class similarity in the above table, the intra-class similarity of cluster C4 is evidently lower than that of other clusters, so it is noise cluster, which is the same as the result of Table 2. The mean value of intra-class similarity and similarity between classes are the mean value after removing the noise cluster. Although the absolute quantity of the mean value (0.253) of intra-class similarity is not larger, but it is larger compared with the average similarity of all sessions between two users. The mean value of intra-class similarity for each clustering is more close to the average similarity of sessions between two users. And we can know that the clustering result is better.

IV. SUMMARY AND CONCLUSIONS

Based on detailed analysis on the advantages and disadvantages of the existing clustering algorithms, the paper studies and improves FCM algorithm in view of the characteristics of Web log data. FCM algorithm has the disadvantages that it is mainly applied to point data clustering and can't directly process relational data, for which, the paper firstly improves FCM algorithm to make it process relational data, and makes robust improvement on the algorithm. Then, as traditional FCM algorithm needs to be on the basis without priori knowledge, the paper predetermines the disadvantages of the number of clustering categories, introduces CA algorithm which is combined with FCM algorithm, which forms CA-FCM algorithm to automatically determine the best number of categories. And a weight is added to the membership to reduce the influence of the outlier data on clustering center. Finally, the paper proposes the weighting CA-WFCM algorithm based on competitive agglomeration aiming at the characteristics of Web log data.

The paper improves FCM algorithm based on the characteristics of Web log data, which makes it can directly cluster the relational data. And for the disadvantage of FCM algorithm that it is difficult to determine in advance the clustering category number c, the paper uses the mechanism based on competition condensation algorithm combined with FCM algorithm, which forms CA-FCM algorithm and makes robust improvement on algorithm. At last, the paper makes detailed experiment and analysis on the algorithm, and compares it with the results of FCM algorithm, which proves the feasibility and correctness of the algorithm.

REFERENCE

[1] Wang Jicheng,Huang Yuan,Wu Gangshan and Zhang Fuyan.Web Mining: Knowledge Discovery on the Web[J]. IEEE,PP,137-141,1999.

[2] Chen Xinzhong, Liyan, Yang Bingru. Development on mining technology of Web log[J]System Engineering and Electronic Technology, PP.492-495,2003, 25(4).

[3] Guo Yan, Bai Shuo, Yang Zhifeng. Analysis on network log scale and mining on user interest [J]. Chinese Journal of Computers, PP.1483-1496,2005, 28(9).

[4] Guo Yan, Bai Shuo, Yu Manquan. Summary on using information mining [J]. Computer Science, PP.1-7,2005, 32(1).

[5] Wang Shi, Gao Wen, Li Jintao. Knowledge discovery of path cluster in Web sites[J]. Computer Research and Development, PP.482-486,2001, 38(4).

[6] Song Qinbao, Shen Junyi. High-efficiency mining algorithm of Web log [J].Computer Research and Development, PP.328-333,2001, 38(3):.

[7] Xing Dongshan, Shen Junyi, Song Qinbao. Mining preference path of users browsing [J]. Chinese Journal of Computers, PP. 1518-1523,2003, 26(11).

[8] Dong Yihong, Zhuang Yueting. Web log mining based on new competitive neural network [J]. Computer Research and Development, PP. 661-667,2003, 40(5).

[9] M, Kilger. A shadow handler in a video-based real-time traffic monitoring system.Proceedings of IEEE Workshop on Applications

of Computer Vision,PP.11-18.1 992, 30,Nov-2, Dec1992.
[10] Chaba,Yogesh,Singh,Yudhvir,Aneja,Preeti. Performance Analysis of Disable IP Broadcast Technique for Prevention of Flooding-Based Algorithm for Image Edge Detection[J]. Journal of Networks,pp.56,2011,6(6).
DDoS Attack in MANET[J]. Journal of Networks,pp.45,2009,4(3).
[11] Shaosheng,Fan,Hainan,Wang. Multi-direction Fuzzy Morphology