A Kind of Improved Data Clustering Algorithm in Web Log Mining

Jin Guo[1, a], Shengbing Zhang[2, b], Zheng Qiu[3, c]

[1]Department of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China

[2]Department of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China

[3]The Fifteenth Laboratory, 631st Research Institute of Aviation Industry Corp, Xi'an,710068, China

[a]email: guojin1019@163.com, [b]email: zhangsb@nwpu.edu.cn, [c]email: qiuzheng1213@163.com

Keywords: Web Log; Clustering; K-means; Fuzzy Matrix;

Abstract. Aiming at the user clustering and page clustering in Web log mining and based on the analysis of K-means clustering algorithm and matrix clustering algorithm, the paper presented an improved clustering algorithm that combining fuzzy matrix algorithm with K-means algorithm. Extract compressed sub-matrix from relational matrix of user and page, establishing user interval, and then divide all users into large intervals and separate the noise data, obtain initial value and classified number for K-means algorithm, effectively solve the defect in the K-means algorithm that always suppose or make a try to definite the classified number and the initial value, also include the lacking to exclude the noise data obstruction.

Introduction

Clustering analysis can not only act as a stand-alone tool to get data distribution and to focus on further analysis on some clusters by observing characteristics of each cluster, but also can be used as pre-processing steps of other algorithms such as association rules, Category, and then extraction or classification can be executed based on this to improve the accuracy and efficiency of mining [1, 2, 3]. Web access log data includes the user's IP address, access time, access methods (such as CET, POST), the requested file URL, Hypertext Transfer Protocol (HTTP) version number, return code (the request status, success or error code), the number of bytes transferred properties, which can be seen as nominal data in view of user and page access relationship. It has characteristics of large amount and sparse data distribution, so clustering of Web log has some difference with clustering of common sense [4, 5]. Aiming at the user clustering and page clustering in Web log mining and based on the analysis of K-means clustering algorithm and matrix clustering algorithm, the paper presented an improved clustering algorithm that combining fuzzy matrix algorithm with K-means algorithm. Extract compressed sub-matrix from relational matrix of user and page, establishing user interval, and then divide all users into large intervals and separate the noise data, obtain initial value and classified number for K-means algorithm, effectively solve the defect in the K-means algorithm that always suppose or make a try to definite the classified number and the initial value, also include the lacking to exclude the noise data obstruction. The specific arrangement of the paper is as follows: Section 2 introduces Web log clustering method; Section 3 analyzes on K-means clustering algorithm and matrix clustering algorithm and points out shortcomings of these two algorithms; Section 4 presents an improved data mining algorithm in Web log mining by combining these above two kinds of algorithms; Section 5 concludes our work.

Web Log Clustering Method

Clustering of Web log includes user clustering and page clustering:

User clustering. User clustering is the analysis on user session. It will look for users with similar behavior pattern according to users' browsing behavior, so as to generally meet characteristics of interests of users' similar access the page with more overlap. Then these users can be divided as a group, so the users in the group can share this page collection that has high user access frequency.

Page clustering. Page clustering tries to look for pages being accessed by same users through analysis on access situations, and then classify them as a group. So pages in the group have same characteristics, and they are accessed by a same group of users.

In the log mining clustering, membership $U = R^{c \times n}$ matrix can be obtained in the end of session clustering, where *c* is the final number of clusters and *n* is the number of session that belongs to cluster. After the implementation of clustering, each session is divided into the substituted type, that is:

$$c_i = \{s_k \in S \mid u_{ik} > u_{jk}, \forall j \neq i\}, 1 \leq i \leq c \tag{1}$$

Where c_i *i*s the clustering result *i*, s_k is the user session *k*, *S* is the total sessions, U_{ik} is the user session *k* belongs to type *i*. Then the clustering *i* can be expressed as the following form:

$$c_i = \{(p_{i1}, url_{i1}), \cdots, (p_{im}, url_{im})\} \tag{2}$$

Where *m* is the total number of page URL, url_{ij} is the page *j* in the cluster *i*. When necessary, we can also compute the important degree of page *j* in the cluster *i*, namely weight that is set as p_{ij} as the following:

$$p_{ij} = \frac{|c_{ij}|}{|c_i|} \tag{3}$$

Where $|c_{ij}|$ is the weight quality of page *j* in the cluster *i*, $|c_i|$ is the total quality of user sessions in cluster *i*.

In classical clustering algorithm, the evaluation of clustering results has two classical indexes of intra-cluster distance and inter-cluster distance. As to the log clustering based on Web, the session distance can be seen as the similarity of two sessions, The average value S_{intra} of similarity among pairs of sessions in a cluster is the intra-cluster distance.

$$S_{\text{int}ra} = \frac{\sum_{S_k \in C_i, S_l \in C_i, k \neq l} sim_{kl}}{|C_i|(|C_i - 1|)} \tag{4}$$

Where sim_{kl} is the distance from user session *k* to session *l* in the cluster *i*; $|c_i|$ is the total number of user session in cluster *i*. Larger S_{intra} indicates more member similarity in the cluster and better clustering effect. The value 1 means all members in a cluster are totally same.

Similarly, we can also obtain the similarity of user sessions between that in one cluster and from other session S_{inter} as the inter-cluster distance, which is the average value of distance between user session in one cluster and that in other clusters:

$$S_{\text{int}er} = \frac{\sum_{S_k \in C_i, S_l \in C_j} sim_{kl}}{|C_i||C_j|} \tag{5}$$

Where sim_{kl} is the distance from user session *k* in the cluster *i* to the user session *l* in the cluster *j*; $|c_i|$ is the total number of user session in cluster *i* and $|c_i|$ is the total number of remaining all user session apart from cluster *j*. Smaller S_{inter} means more loosely cluster and better clustering effect.

Analysis on K-means Clustering Algorithm and Matrix Clustering Algorithm

A. K-means Clustering Algorithm

The processing flow of K-means clustering algorithm is as follows: firstly randomly select K data objects as the initial clustering centers from N data objects. For the remaining objects, assign them to the most similar cluster according to the distance between them and clustering centers. And re-calculate the average value of each cluster. The process should be constantly repeating until the standard measure function begins to converge.

In the Web log mining, substitute the distance of user clustering with similarity to arrive at the K-means clustering algorithm in log mining as shown in Fig. 1.

The biggest drawback of K-means algorithm is that it needs to pre-determine the number of clusters, if the user has no sense about it in advance, the algorithm is not suitable. Furthermore, K-means algorithm is very sensitive to noise data and edge data, so even small noise or edge data will greatly shift the clustering center and impact the clustering effect. Therefore, in the application of K-means algorithm for clustering users, the data preprocessing effect of Web log data will directly affect clustering effect.

```
Input: Number of clusters k and user session containing k initial centers.
Output: k values that minimize squared error criterion function.
repeat:
1. For each user session, compute session similarity between it and central user session.
2. For each user session, select cluster center user session with maximize similarity and assign it to corresponding cluster.
3. Re-calculate the centroid of each session cluster and re-determine central use session value.
4. Determine change of centroid according to error criterion function until the change less than a certain range.
5. Output final session cluster division.
end
```

Fig.1. K-means clustering algorithm in log mining

```
Input: $A_{m\times n}$, user confidence, page confidence, user threshold and interface threshold.
Output: User and page relationship matrix.
Initalization;
Delete all rows that less than user confidence
Delete all columns that less than page confidence
repeat
{
Find the beginning of rows that contain non-zero items
Delivered to the column
Clear row marks
repeat
{
Column is passed to the row
Statistics received activation number of row
Row cutting according to received activation number such as small user threshold
Row is passed to the column
Statistics received activation number of column
Column cutting according to received activation number such as small interface threshold
} until there is no change in row and column
Determine whether the row number of matrix is bigger than or equal to user confidence, column number of matrix is bigger than or equal to page confidence. If it is, output.
Select next initial row.
} until all rows have been selected.
```

Fig.2. The matrix clustering algorithm

B. Fuzzy Matrix Clustering Algorithm

To problem to be solved by fuzzy matrix clustering can be described as activating transmission of matrix clustering. In each time of activation transmission, select user and page whose activation number is bigger than user threshold and page threshold, and then to find the problem that row number bigger than or equal to user confidence and column number bigger than or equal to page confidence. In fact, the abstracted sub-matrix represents a kind of user cluster, which is called interval division, but this kind of division is rough and limited.

The algorithm of matrix clustering is as Fig. 2.

Improved Clustering Algorithm and Example Analysis

A. Improved Clustering Algorithm

Through the analysis on above clustering algorithms, we can know that the two algorithms have their advantages and disadvantages. As to the fuzzy clustering algorithm, it can better distinguish general interval clusters but cannot execute more accurate clustering analysis. The K-means clustering algorithm needs to determine the number of cluster centers and the initial value. It is important to determine these values, but it is difficult to find ideal values in most cases. Therefore, we can combine the advantages of these algorithms, so as to find number of central area through fuzzy matrix clustering and select initial value from it, then it can be used as input for K-means algorithm.

Assume that the user interval division result is {a, b, c, e}, then the initial clustering point of K-means clustering can select the point close to central point in this user interval. The number of clusters of matrix clustering can be used as that of K-means. For the cluster with less data sets, it can be identified as edge points or edge data set, which can be removed or be specially treated, so as to greatly reduce impact of noise on clustering.

B. Clustering Algorithm Experiment

The test was implemented on log data after preprocessing. As the data amount is large, here just extract sessions whose page amount between 4 and 10 for experimental test. Fuzzy matrix clustering algorithm was used for large interval division, and then to determine the initial value and cluster amount according to interval. Finally, the improved clustering result was compared with that of K-means clustering. The average error of K-means clustering was set as 0.1. The K-means clustering result is shown in Table 1.

Table.1.K-means clustering result

Cluster	Session number	Session id belongs to the cluster
Cluster 1	1	1
Cluster 2	524	2, 11, 12, 15, 17, 18 ,19, 20, 24, 27…
Cluster 3	13	3,77,85,94,114,115,131,188,386,647…
Cluster 4	1	4
Cluster 5	24	5,141,159,171,178,261,262,265,269,271…
Cluster 6	148	10,14,16,22,23,26,28,34,41,48…
Cluster 7	14	7,25,47,74,83,187,270,275,406,507…
Cluster 8	44	6,8,13,79,104,113,130,150,179,185…
Cluster 9	4	9,153,521,526

Table.2. Fuzzy clustering interval result

Cluster	Session number	Session id belongs to the cluster
Interval 1	63	4,158,170,472,474,557,140,177,261,264…
Interval 2	53	5,7,8,12,129,149,152,178,228,229…
Interval 3	13	6,24,46,73,82,269,274,405,506,508…
Interval 4	30	15,47,60,64,65,66,91,97,151,207…
Interval 5	44	16,45,98,106,126,135,137,150,154,180…
Interval 6	11	44,55,96,271,296,300,355,477,491,595,119…
Interval 7	12	62,105,111,122,153,185,316,333,402,504…
Interval 8	13	85,87,132,611,173,311,358,370,409,527…
Interval 9	11	140,158,170,260,261,264,268,270,406, 557…

Firstly the fuzzy clustering is as follows: obtain the division classification as follows and select session number larger than 10 as an interval, where the page threshold is 3 and user threshold is 3. The interval division result is shown in Table 2.

Select appropriate initial value in the determined space, the best choice is the point at the area center. Here {502, 5, 6, 15, 16, 44, 62, 85, 261} is selected as initial value. The clustering after improvement is shown in Table 3.

From the above two tables we can see that the effect of K-means clustering is not better. It produced 2 isolated points and a cluster has only 4 data. The fuzzy matrix clustering for interval division can not only effectively determine classification number of clustering, but also determine the initial value. From the comparison of these two groups of data, it is not difficult to find that the improved clustering algorithm has not produced isolated points and has better clustering effect. If necessary, the data that has not been assigned into large interval can be processed specially or removed to avoid noise interference. The intra-cluster distance and inter-cluster distance of these clustering can be obtained with formula 4 and 5, which are shown in Table 4.

Table.3.Clustering result after improvement

Cluster	Session number	Session id belongs to the cluster
Cluster 1	13	211,214,215,232,256,264,265,380,502,525…
Cluster 2	12	5,159,178,329,350,473,475,558,586,667…
Cluster 3	164	1,6,8,9,10,13,14,21,22,23…
Cluster 4	468	4,7,11,12,15,18,19,20,24,25…
Cluster 5	22	16,48,61,65,66,67,152,208,282,353…
Cluster 6	60	2,17,46,53,55,64,99,107,108,127…
Cluster 7	10	58,62,146,320,332,400,551,587,608,759
Cluster 8	12	3,77,85,94,114,115,131,188,386,701…
Cluster 9	12	141,171,261,262, 269,271,407,474,610,627…

From the comparison of Table 4 and Table 5 we can see that the average intra-cluster distance of improved algorithm is 0.336, which is larger than that of K-means algorithm 0.298. The average inter-cluster distance 0.0032 is just half of that of K-means clustering algorithm, which indicates that the clustering effect is better. Note that the distance is actually obtained from similarity, because larger intra-distance means larger similarity and vice versa.

In summary, the paper applied fuzzy matrix clustering algorithm in K-means clustering improvement can not only solve problems of classification number and initial value setting in K-means clustering, but also effectively distinguish effect of isolated points, which reflects a better advantage.

Table.4. Intra-cluster distance and inter-cluster distance of K-means clustering algorithm

Cluster	Intra-cluster distance	Inter-cluster distance
1（Isolated point）	\	\
2	0.121	0.0001
3	0.477	0.0043
4（Isolated point）	\	\
5	0.329	0.0026
6	0.067	0.0004
7	0.696	0.0057
8	0.240	0.0014
9	0.155	0.0206
Average value	0.298	0.0050

Table.5. Intra-cluster distance and inter-cluster distance of improved clustering algorithm

Cluster	Intra-cluster distance	Inter-cluster distance
1	0.216	0.0019
2	0.471	0.0058
3	0.066	0.0004
4	0.134	0.0001
5	0.248	0.0025
6	0.586	0.0009
7	0.567	0.0067
8	0.492	0.0044
9	0.245	0.0057
Average value	0.336	0.0032

Conclusion

Clustering analysis plays important role in the log mining based on Web. Aiming at the user clustering and page clustering in Web log mining and based on the analysis of K-means clustering algorithm and matrix clustering algorithm, the paper presented an improved clustering algorithm that combining fuzzy matrix algorithm with K-means algorithm. Extract compressed sub-matrix from relational matrix of user and page, establishing user interval, and then divide all users into large intervals and separate the noise data, obtain initial value and classified number for K-means algorithm, effectively solve the defect in the K-means algorithm that always suppose or make a try to definite the classified number and the initial value, also include the lacking to exclude the noise data obstruction.

References

[1] B. Mobasher, H. Dai, T. Luo. Effective Personalization Based on Association Rule Discovery from Web Usage Data, Proceedings of the ACM Workshop on Web Information and Data Management, 2001, pp. 176-184.

[2] Ester M., Kriegel H.P. and Sander J. A. Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. Of KDD99, 1999.

[3] R. Agrawal, Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules, Proc. 20th Int. Conference on Very Large Data Bases, VLDB94, 1994.

[4] Buchner A. G, Mulvenna M. D. Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining, ACM SIGMOD, vol. 27, 1998, pp. 54-61.

[5] Chen, M.S, Park, J.S.& Yu, P.S. Data Mining for Traversal Patterns in a Web Environment, Proc. 16th Int. Conference on Distributed Computing Systems, pp. 385-392, 1996.