IJCSI International Journal of Computer Science Issues, Volume 15, Issue 4, July 2018
ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784
www.IJCSI.org https://doi.org/10.5281/zenodo.1346025

Parallel Distributed Web Access Patterns Two-layer Clustering Based on the Fuzzy and Rough

Xiaoli Jia[1], Rui Wu[2] and Siying Wu[3]

[1] Institute of Mathematics and Computer, Shanxi Normal University, Shanxi Linfen,China

[2] Institute of Mathematics and Computer, Shanxi Normal University, Shanxi Linfen,China

[3] Institute of Mathematics and Computer, Shanxi Normal University, Shanxi Linfen,China

Abstract

Web log mining analyzes user's browsing time to gain users' level of interest. In this paper,we use the fuzzy vector to represent the user access patterns, which reflects whether the user has visited the page and the time of browsing. Then, the users' access sequences are analyzed by different clustering methods. In addition, a two-layer clustering technique is proposed based on the fuzzy rough k-means and Angle cosine,which can reduce the sensitivity to the initial clustering center. And the feasibility of the clustering method is demonstrated by a series of experiments. The results of different clustering methods are verified by using the Davies-Bouldin index.When the data sets are too large, the algorithm is inefficient. Therefore, we use MapReduce to realize the parallelism of two-layer clustering , improving the efficiency of clustering.

Keywords: *Web mining, Fuzzy rough clustering, Web access patterns, Angle cosine, parallel*

1. Introduction

With the arrival of big data and cloud computing, a large amount of data will appear on the Internet every day. How to mine valuable information in a huge data set is a research hotspot of data mining.Web usage mining analyzes users' interest preferences and behavior patterns by mining information in user access patterns, and improves the recommendation quality of the recommendation system.By analyzing user's browsing behavior, user clustering can classify users who have similar access patterns into one category and provide personalized services to users more precisely.

Due to the uncertainty of web data, it is necessary to make data mining based on fuzzy and rough. Chien-liang Liu and others [1] processed document clustering by a fuzzy semi-supervised clustering algorithm called fuzzy semi-kmeans, which provided the flexibility to employ different fuzzy membership functions to measure the distance between data. A fuzzy clustering framework based on MDF [2] clusters user sessions ,which use fuzzy c-means (FCM) and fuzzy c-medoids (FCMed) algorithms. Indrajit Saha [3] developed simulated annealing algorithm based on fuzzy rough k-modes and genetic algorithm based on fuzzy rough k-modes, by cosidering clustering as a basic optimization problems to better clustering. Van Nha Pham [4] proposed the clustering approach combining fuzzy clustering and interval-valued fuzzy sets. The determination method of fuzzy c-means clustering based on extended rough sets model [5] can effectively avoid the impact of the initial clustering number.

The fuzzy rough k-means algorithm [6] can not only handle numerical data, but also deal with non-numeric data such as text and image data. However, it has limitations on the initial clustering center. When the data volume is too large, this algorithm will spends a lot of time. The redundant information caused by the dynamic change of clustering will also affect the algorithm. In this paper ,a new algorithm in a fuzzy rough environment is proposed. Firstly,data sets use rough k-means clustering.Then,when the cluster centers' change is less than the threshold, the cluster centers are outputted.At last, the similarity between data sets and the cluster centers are calculated by the angle cosine,then the data sets will be re-cluster.The algorithm can reduce the number of iterations, also does not have much restrictions on the initial clustering center.

When processing massive data, the algorithm has a large demand for storage space, and the serial web clustering algorithm has disadvantages such as low efficiency, long running time and so on. Therefore, many new clustering technologies are proposed in the environment of big data and cloud computing. In the cloud computing environment, a new solutions which was called DAMIS [7] optimized data classification clustering and dimensionality reduction.

The personalized knowledge service method based on users' behaviors [8] calculates the users' personalized knowledge demand degree through the analysis of user behavior in a cloud manufacturing environment. Zhou Runwu et al. [9] proposed the k-means algorithm (OSCK)

based on optimization sampling clustering in big data environment to solve the problem of clustering accuracy and convergence.

In this paper, a parallel web two-layer clustering based on fuzzy and rough is proposed in the big data environment, and the Mapreduce is used to parallelize the clustering algorithm based on the fuzzy rough k-means and Angle cosine.The parallel clustering algorithm is used to cluster the user access patterns.

2.Data preprocessing

It can reflect the users' different interest about the web page ,when browsing the same web page by different users at different times . Each user access pattern is transformed into an equivalent length of the fuzzy vector form, each element in the vector is either 0 or a fuzzy language variable that represents the browsing time.

The user browsing modes are noted as $P=\{p_1,p_2,\cdots,p_n\}$, n represents the number of users, p_i represents the specific browsing behavior of the i-th user.The visited pages are marked as $V=\{u_1,u_2,\cdots,u_m\}$, the number of pages visited by all users and the time spent on the page is recorded as $U=\{(u_{i1},t_{i1}),(u_{i2},t_{i2}),\cdots,(u_{ik},t_{ik})\}$, u_{ik} is the k-th page visited by the i-th user , t_{ik} indicates the time of visit the u_{ik} .

The order of visiting the web page is not considered firstly, p_i can be expressed as the following vector form. Therefore each mode can be converted into a real vector of length m .

$$Q_i=<q_{i1}^t,q_{i2}^t,\cdots,q_{im}^t>$$

$$q_{ik}^t=\begin{cases}t_{ik},(u_{ik},t_{ik})\in s_i\\0,\ (u_{ik},t_{ik})\notin s_i\end{cases}\qquad(1\le k\le m)$$

(1)

The dwell time when visiting web pages is divided into some fuzzy intervals. Each fuzzy region is described by a fuzzy linguistic variable, and the membership function of each fuzzy linguistic variable is given according to simulation methods or the expert system. According to the membership function, the relationship between q_{ik}^t and the fuzzy linguistic variable $\beta_{ik}(1\le i\le n)(1\le k\le m)$ is represented as shown below:

$$F(\alpha_i)=\beta_{ik}$$

(2)

Among them α_i is fuzzy linguistic variables,F is a mapping function from α_i to β_{ik}.

According to formula (1) and (2),each user browsing mode can be converted to the following form:

$$f_{qi}=<\beta_{i1},\beta_{i2},\cdots,\beta_{im}>$$
$$(\beta_{ik}\in\{0,\alpha_1,\alpha_2,\cdots,\alpha_y\})$$

(3)

3.Improvement of Fuzzy Vector Clustering Based on Rough k-Means

3.1 Fuzzy vector clustering based on rough k-means

Assuming there are n users browsing web modes,which are labeled as $P=\{p_1,p_2,\cdots,p_n\}$. For any two of the patterns p_i and p_j, they can be converted into the form of the following fuzzy vector:

$$f_{qi}=<\beta_{i1},\beta_{i2},\cdots\beta_{im}>(\beta_{ik}\in\{0,\alpha_1,\alpha_2,\cdots,\alpha_y\})$$
$$f_{qj}=<\beta_{j1},\beta_{j2},\cdots\beta_{jm}>(\beta_{jk}\in\{0,\alpha_1,\alpha_2,\cdots,\alpha_y\}(1\le k\le m))$$

The sum of these two fuzzy vectors is defined as:

$$sum(f_{qi},f_{qj})=<E[\beta_{i1}+\beta_{j1}],E[\beta_{i2}+\beta_{j2}],\cdots,E[\beta_{im}+\beta_{jm}]>$$

(4)

($E[\beta_{ik}+\beta_{jk}](1\le k\le m)$ is obtained by the fuzzy simulation algorithm).

The dissimilarity (distance) of any two user browsing modes p_i and p_j is defined as:

$$d(p_i,p_j)\cong d(f_{qi},f_{qj})=\sqrt{\frac{\sum_{k=1}^{m}(E[\beta_{ik}-\beta_{jk}])^2}{m}}$$

(5)

(In the same way, $(E[\beta_{ik}-\beta_{jk}])^2(1\le k\le m)$ is solved by using the fuzzy simulation algorithm).

Since the boundaries of each class can be ambiguous, each class is defined as a rough variable $\delta_i(1\le i\le k)$ in a measurable rough space [6]. Then the definition of the center point of each class is as follows:

$$m_i=\begin{cases}w_{low}\dfrac{\sum_{f_{qj}\in\underline{\delta_i}}f_{qj}}{|\underline{\delta_i}|}+w_{up}\dfrac{\sum_{f_{qj}\in(\overline{\delta_i}-\underline{\delta_i})}f_{qj}}{|\overline{\delta_i}-\underline{\delta_i}|}\\w_{low}\dfrac{\sum_{f_{qj}\in\underline{\delta_i}}f_{qj}}{|\underline{\delta_i}|}\end{cases}$$

(6)

Among them, the parameter w_{low} determines the lower approximate weight at the center point, and w_{up} determines the upper approximate weight. $0.5<w_{low}<1$, $w_{up}=1-w_{low}$.

$|\underline{\delta_i}|$ is the number of modes in the i-th class lower approximation, $|\overline{\delta_i} - \underline{\delta_i}|$ denotes the difference between the number of lower and upper approximations.Center point m_i is a real-valued vector which is represented as $\mathrm{m}_i = < c_{i1}, c_{i2}, \cdots, c_{im} >$.

The distance from mode $p_i (1 \le i \le n)$ to center point $\mathrm{m}_j (1 \le j \le k)$ is defined as follows:

$$d(p_i, \mathrm{m}_j) = d(f_{qi}, \mathrm{m}_j) = \sqrt{\frac{\sum_{l=1}^{m} (E[\beta_{il} - c_{jl}])^2}{m}} \tag{7}$$

If the difference between $d(p_i, \mathrm{m}_{k1})$ and $d(p_i, \mathrm{m}_{k2})$ is small and is less than a given threshold,then pattern p_i $(1 \le i \le n)$ cannot be clearly classified into a class.It may belong to category $k1$ or category $k2$, so it belongs to the upper approximation of category $k1$ and category $k2$.Otherwise, if $d(p_i, \mathrm{m}_{k1})$ is minimum over the k clusters , the p_i mode must belong to class $k1$, so it belongs to the lower approximation of class $k1$. As you can see, there is overlap between classes and classes.

Supposing $P = \{p_1, p_2, \cdots, p_n\}$ is clustered into k categories, $\mathrm{C} = \{C_1, C_2, \cdots, C_k\}$ expresses classification framework,each class in them can be characterized as a rough variable $\delta_i (1 \le i \le k)$, which can be represented by its lower approximation and upper approximation of $(\underline{\delta_i}, \overline{\delta_i})$.

3.2 Similarity calculation

In collaborative filtering recommendation algorithm,the similarity calculation methods include: angle cosine, Jaccard similarity coefficient, correlation coefficient and so on.The user's preference for an item is regarded as a vector. For example, the score or preference for item 1 of n users is expressed as $A_1 = (X_{11}, X_{21}, X_{31}, ..., X_{n1})$, the score or preference of n users to item M is expressed as $A_M = (X_{1m}, X_{2m}, X_{3m}, ..., X_{nm})$, in which m is the item and n is the number of users.

(1) Angle cosine:

$$sim_{1m} = \frac{\sum_{k=1}^{n} x_{k1} x_{km}}{\sqrt{\sum_{k=1}^{n} x_{k1}^2} \sqrt{\sum_{k=1}^{n} x_{km}^2}} \tag{8}$$

The range of values is [-1,1],when the cosine value is close to +1 or -1, it indicates that the two vectors have fairly similarity. When the cosine value is 0, it means uncorrelated.

(2) Jaccard similarity coefficient:

$$J(A_1, A_M) = \frac{|A_1 \bigcap A_M|}{|A_1 \bigcup A_M|} \tag{9}$$

The denominator $A_1 \cup A_M$ indicates the total of users who like the item 1 or the item M, and the numerator $A_1 \cap A_M$ indicates the number of users who like the item 1 and the item M at the same time.

(3) Correlation coefficient:

$$sim_{1m} = \frac{\sum_{k=1}^{n} (x_{k1} - \overline{A_1})(x_{km} - \overline{A_m})}{\sqrt{\sum_{k=1}^{n} (x_{k1} - \overline{A_1})^2} \sqrt{\sum_{k=1}^{n} (x_{km} - \overline{A_m})^2}} \tag{10}$$

The correlation coefficient is between -1 and 1. The greater the absolute value of the correlation coefficient, the higher the degree of correlation .

3.3 Improvement of Fuzzy Vector Clustering Based on Rough k-Means

Fuzzy rough k-means clustering method is simple, easy to understand, but its operation will be limited by the initial clustering center, and each iteration needs to calculate all data objects, heavy calculation burden, more iterations. Therefore, a two-layer clustering method is presented which combines fuzzy rough k-means and Angle cosine.

First, the user access patterns is clustered according to the features data of user, thus forming some clusters. When the change of the clustering center is less than the threshold, the fuzzy rough k-means is terminated, recording the final cluster centers. Then, the Angle cosine is used to calculate the similarity between the user access patterns and each cluster center, and then clustering again. Finally, output the clustering result .

The two-layer clustering algorithm which based on fuzzy rough k-means and Angle cosine can reduce calculating amount, the number of iterations, and this algorithm is less affected by the initial clustering center.It also effectively reduces the empty clustering and the situation that clustering centers with two classes are similar .

The algorithm is described as follows:

Enter: User transaction $P = \{p_1, p_2, \cdots, p_n\}$, the membership function of the browsing time, the threshold of distance difference from the cluster center $\phi \in [0,1]$, the upper approximate weight: w_{up} , the lower approximate weight: w_{low}, the threshold of the cluster center variation $\theta \in [0,1]$, and the cosine difference from the cluster center threshold $\varepsilon \in [0.05, 0.1]$.

Output: Classification Framework C

(1)Start.

(2)According to section 2, each user mode p_j is converted to the corresponding fuzzy vector f_{qj}.

(3)Initialize the center point m_i ,and randomly select k user patterns as the initial cluster center.

(4)For each $f_{qj}(1 \le j \le n)$ do

For i=1 to k do

According to formula 7 compute $d(f_{qj}, \mathrm{m}_i)$

(5)For each $f_{ql}(1 \le l \le n)$ do

If $d(f_{ql}, \mathrm{m}_i) - d(f_{ql}, \mathrm{m}_j)(i \ne j) \le \phi$ ($\phi = 0.1$), then $p_l \in \overline{\delta_i}$, $p_l \in \overline{\delta_j}$;

Else if $d(f_{vl}, \mathrm{m}_i)$ is minimum over the k clusters, then $s_l \in \underline{\eta_i}$.

(6) $C_1 = \{\underline{\delta_1}, \overline{\delta_1}\}, \cdots, C_k = \{\underline{\delta_k}, \overline{\delta_k}\}$

(7)According to formula 6, recompute the center point m_i of each class.

(8)Repeating (4)-(7) to carry out iteration, and after each iteration, the absolute value of the difference between the two clustering centers is calculated. When the value is less than the threshold value θ, the calculation of k-means will be finished, and the clustering results and clustering center will be saved.

(9)According to formula 8, compute $\mathrm{sim}(f_{qj}, \mathrm{m}_i)$, clustering the data again.

(10)Output the clustering results.

(11)Algorithm was terminated.

The above algorithm is applicable to the clustering on small data sets. When the data volume is too large, on the one hand ,there is insufficient memory space , and on the other ,the algorithm is inefficient and the computation time is long. Therefore,the Mapreduce is used to realize the parallelization of the above algorithm. Mapreduce mainly performs distributed computing through Map and Reduce functions. Its operating model is shown in Figure 1.

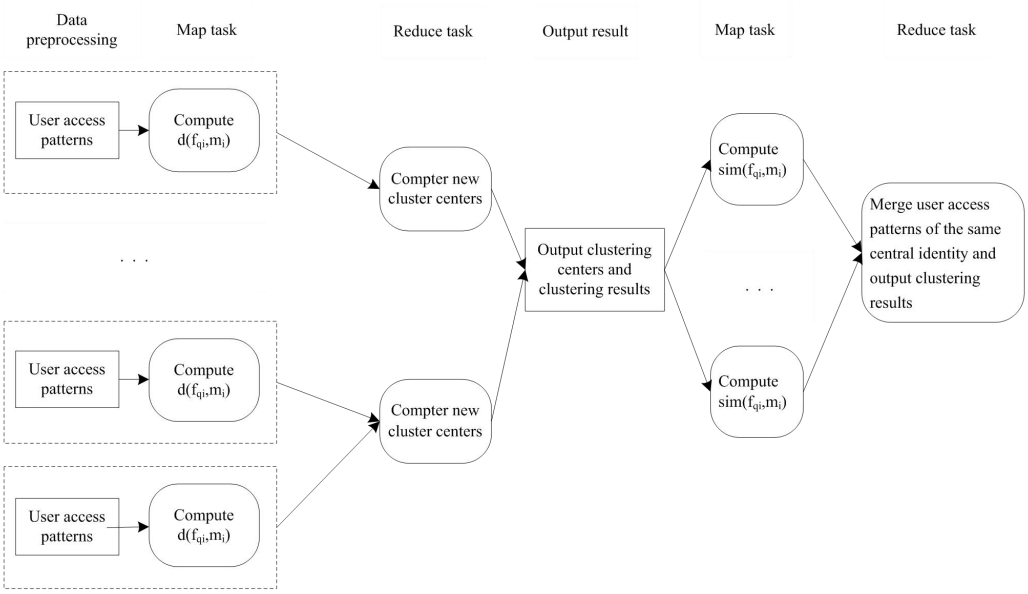

Fig.1 Operation mode of two-layer clustering based on Mapreduce

4.Example analysis

This algorithm is described through an example below . Assume that the user access modes are preprocessed as shown in Table 1.

According to the membership function given by the expert system, the time that users visit web pages is characterized by three trapezoidal fuzzy linguistic variables, short(5,5,30 ,60), middle(30,60,90,120) and long(90,120,150,150). The relation between the browsing time q_{ik}^t and the fuzzy linguistic variable $\beta_{ik}(1 \le i \le n)(1 \le k \le m)$ is as follows.

$$\beta_{ik} = \begin{cases} 0, & q_{ik}^t = 0 \\ short, & 5 \le q_{ik}^t \le 45 \\ middle, & 45 < q_{ik}^t \le 105 \\ long, & 105 < q_{ik}^t \le 150 \end{cases} \quad (11)$$

Table 1: User browsing mode set

Client id	user browse sequence	Client id	user browse sequence
1	(A,25),(B,30),(D,120),(E,80)	2	(A,65),(B,77),(F,115)
3	(A,50),(B,80),(D,30),(G,90),(H,130)	4	(A,75),(C,89),(G,95),(H,100)
5	(A,20) ,(B,35),(D,135)	6	(A,98),(B,88),(G,70),(H,125)
7	(B,85),(D,60) ,(G,40)	8	(B,68),(D,115),(F,30)
9	(A,70), (B,90),(G,35)	10	(B,60),(D,83),(G,110)
11	(A,79),(B,90),(E,15),(G,128)	12	(B,69) ,(D,80),(E,113),(G,96)
13	(B,86),(D,60),(E,35)	14	(A,88),(B,30) ,(E,108), (G,135)
15	(B,33),(C,55),(E,80)		

Making P=$\{ p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}, p_{13}, p_{14}, p_{15} \}$, $U = \bigcup_{i=1}^{15} p_i$, $V = \{A, B, C, D, E, F, G, H\}$,then each pattern can be expressed as the following fuzzy vector form according to section 3.1.

p_1 =< *short* , *short* , 0 , *long* , *middle* , 0 , 0 , 0 >
p_2 =< *middle* , *middle* , 0 , 0 , 0 , *long* , 0 , 0 >
p_3 =< *middle* , *middle* , 0 , *short* , 0 , 0 , *middle* , *long* >
p_4 =< *middle* , 0 , *middle* , 0 , 0 , 0 , *middle* , *middle* >
p_5 =< *short* , *short* , 0 , *long* , 0 , 0 , 0 , 0 >
p_6 =< *middle* , *middle* , 0 , 0 , 0 , 0 , *middle* , *long* >
p_7 =< 0 , *middle* , 0 , *middle* , 0 , 0 , *short* , 0 >
p_8 =< 0 , *middle* , 0 , *long* , 0 , *short* , 0 , 0 >
p_9 =< *middle* , *middle* , 0 , 0 , 0 , 0 , *short* , 0 >
p_{10} =< 0 , *middle* , 0 , *middle* , 0 , 0 , *long* , 0 >
p_{11} =< *middle* , *middle* , 0 , 0 , *short* , 0 , *long* , 0 >
p_{12} =< 0 , *middle* , 0 , *middle* , *long* , 0 , *middle* , 0 >
p_{13} =< 0 , *middle* , 0 , *middle* , *short* , 0 , 0 , 0 >
p_{14} =< *middle* , *short* , 0 , 0 , *long* , 0 , *long* , 0 >
p_{15} =< 0 , *short* , *middle* , 0 , *middle* , 0 , 0 , 0 >

Assuming that these 15 user access patterns are clustered into 3 classes by using fuzzy rough k-means clustering and another clustering by the two-layer clustering.

Fuzzy vector clustering based on rough k-means:Calculate $d(f_{ql}, \mathrm{m}_i)(1 \le l \le 15)(1 \le i \le 3)$,if $d(f_{ql}, \mathrm{m}_i) - d(f_{ql}, \mathrm{m}_j)(i \ne j) \le \phi$, then $p_l \in \overline{\delta_i}$ $p_l \in \overline{\delta_j}$; Otherwise, if $d(f_{ql}, \mathrm{m}_i)$ is minimum over the three classes, then $p_l \in \underline{\delta_i}$.

Fuzzy vector two-layer clustering based on rough k-means and Angle cosine: Firstly, fuzzy rough k-means clustering is used, calculate the absolute value θ of the difference between the cluster centers' value before and after each iteration. When $\theta \le 1$, fuzzy rough k-means clustering is terminated ,the final cluster centers is recorded. Then the Angle cosine is used to calculate the similarity between each user access pattern and the final clustering center. If $\mathrm{sim}(f_{ql}, \mathrm{m}_i) - \mathrm{sim}(f_{ql}, \mathrm{m}_j)(i \ne j) \le \varepsilon$,so $p_l \in \overline{\delta_i}$, $p_l \in \overline{\delta_j}$. Otherwise, if $\mathrm{sim}(f_{ql}, \mathrm{m}_i)$ is maximum over the three similarity value, then $p_l \in \underline{\delta_i}$.

Table 2: Comparison of clustering results

	fuzzy vector clustering based on rough k-means	two-layer clustering
iterations	12	5
clustering results	$p_{14} \in \underline{\delta_1}$, $p_{10}, p_{11}, p_{12} \in \overline{\delta_1}$ $p_{15} \in \underline{\delta_2}$, $p_1, p_2, p_4, p_5, p_7, p_8, p_{12}, p_{13} \in \overline{\delta_2}$ $p_3, p_6, p_9 \in \underline{\delta_3}$, $p_1, p_2, p_4, p_5, p_7, p_8, p_{10}, p_{11}, p_{13} \in \overline{\delta_3}$	$p_1, p_5, p_7, p_8, p_{13} \in \underline{\delta_1}$ $p_{12}, p_{14}, p_{15} \in \underline{\delta_2}$ $p_2, p_3, p_4, p_6, p_9, p_{10}, p_{11} \in \underline{\delta_3}$
clustering analysis	The five patterns can be classified exactly, and the overlap between classes and classes is numerous, and the final clustering result is poor.	15 patterns can be classified accurately, and the clustering effect is better.

Table 3: Comparison of clustering results

	fuzzy vector clustering based on rough k-means	two-layer clustering
iterations	5（the clustering centers of two classes are the same）	2
clustering results	$p_2 \in \underline{\delta_1}, p_3, p_4, p_6, p_8, p_9 \in \overline{\delta_1}$ $p_1, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12},$ $p_{13}, p_{14}, p_{15} \in \overline{\delta_2}$ $p_1, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11},$ $p_{12}, p_{13}, p_{14}, p_{15} \in \overline{\delta_3}$	$p_2, p_5, p_8 \in \underline{\delta_1}, p_7, p_9, p_{13} \in \overline{\delta_1}$ $p_1, p_{12}, p_{14}, p_{15} \in \underline{\delta_2}, p_{13} \in \overline{\delta_2}$ $p_3, p_4, p_6, p_{10}, p_{11} \in \underline{\delta_3}$ $p_7, p_9 \in \overline{\delta_3}$
clustering analysis	Only one model can be classified exactly, and the overlap between classes and classes is numerous, and the final clustering result is poor.	All 12 patterns can be classified accurately, and the clustering effect is better.

1. Select p_1, p_2, p_3 as the clustering initial center points, $m_1 = p_1$, $m_2 = p_2$, $m_3 = p_3$, $w_{up} = 0.3$, $w_{low} = 0.7$, threshold value $\phi = 0.1$, $\varepsilon = 0.5$. User access patterns are respectively clustered by using fuzzy rough k-means clustering and another clustering by the two-layer clustering namely combining fuzzy rough k-means and Angle cosine. Finally, the clustering results are shown in table 2.

2. Select p_2, p_4, and p_6 as the clustering initial center points, $m_1 = p_2$, $m_2 = p_4$, $m_3 = p_6$, $w_{up} = 0.3$, $w_{low} = 0.7$, threshold value $\phi = 0.1$, $\varepsilon = 0.5$. User access patterns are respectively clustered by using fuzzy rough k-means clustering and another clustering by the two-layer clustering. Finally, the clustering results are shown in table 3.

From the above mentioned different initial cluster centers, it can be seen that the initial clustering center has a greater influence on the fuzzy rough k-means clustering. Different initial cluster centers have different clustering effects. In addition, fuzzy vector clustering based on rough k-means may give rise to "empty clustering" or close to the same clustering phenomenon . However, if using the two-layer clustering, the above problems can be effectively avoided and the clustering effect is better.

5.Experimental analysis

In order to further prove the effectiveness of the proposed two-layer clustering method by combining fuzzy rough k-means and angle cosine , we performed the following three experiments.

5.1 Comparison between fuzzy rough k-means and the two-layer Clustering

We take 50 pre-processed user access patterns in the synthetic database for experiments.Each pattern can be expressed as a fuzzy vector according to section 3.1.

Suppose user browsing patterns are clustered into 3 categories, each of which is a measurable rough variable $\delta_i (1 \le i \le 3)$ in rough space ,with $w_{up} = 0.3$, $w_{low} = 0.7$, $\phi = 0.1, \theta = 1, \varepsilon = 0.05$.

The Davies-Bouldin Index (DBI) is used to evaluate the clustering effect. The smaller DB means the smaller the distance in the class, and means the greater the distance between the classes. Its definition is as follows:

$$DB = \frac{1}{c}\sum_{k=1}^{c} \max_{l \ne k} \left\{ \frac{P(U_k) + P(U_l)}{d(U_k, U_l)} \right\} \tag{12}$$

Table 4: Comparison of clustering results

initial clustering centers	compare items	fuzzy vector clustering based on rough k-means	two-layer clustering
p_1, p_2, p_3	iterations	14(generate empty clusters)	5
	DB value	1.109	0.693
p_1, p_3, p_5	iterations	32(the clustering centers of two classes are the same)	5
	DB value	1.033	0.738

Among them, $P(U_k\}$ represents the intra-class distance of the k-th class and $d(U_k,U_l)$ represents the inter-class distance.
Select different initial clustering centers, respectively using fuzzy rough k-means clustering and another clustering by two-layer clustering. Finally, the clustering results are shown in Table 4 .
The results show that the initial cluster center has less influence on fuzzy vector clustering which based on rough k-means and Angle cosine, and effectively avoids the situation of empty clustering and the approximation of two classes' clustering centers, and the clustering effect is also ideal.In addition, a large number of user access patterns are selected to cluster by using Mapreduce to parallelize the clustering algorithm, which has short running time, high efficiency and good clustering effect.

5.2 The effect of clustering number on clustering results

Select the data set in Section 5.1. Respectively select 3, 5, 10, 15, and 20 as the number of clusters. The experiment is performed by using two-layer clustering and then the DB value is calculated. The comparison results are shown in Table 5.

Table 5: The effect of clustering number on clustering results

clustering number	DB value
3	0.693
5	0.645
10	0.712
15	0.723
20	0.728

It can be seen from table 5 that if the number of clusters is different, the clustering effect will also be different. Therefore, it is very important to select the appropriate number of clusters for clustering.

5.3 The two-layer clustering based on Mapreduce

When the data set is small, the algorithm of two-layer clustering is relatively efficient. However, when the amount of data is large, the two-layer clustering has long running time and low efficiency. Therefore, we use Mapreduce to achieve parallelization of two-layer cluster-ing.Taking a log file of the Shanxi Normal University official website as a sample data set, the total number of pages is 1020. After the data cleaning, there are 72683 access patterns, the number of users is 5280.Then the user access patterns are grouped into 20 categories. The fuzzy rough k-means and the two-layer clustering which are based on Mapreduce are applied respectively to cluster analysis. As the number of computing nodes increases, the acceleration of the two clustering algorithms is shown in the following figure:

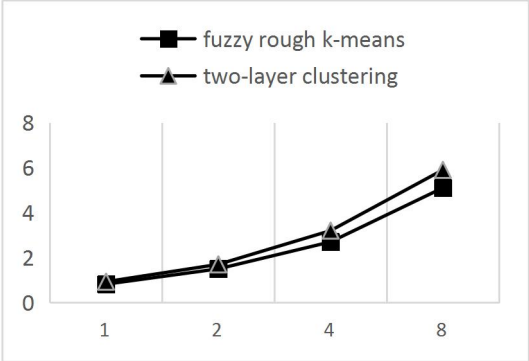

Fig.2 The comparison of the speedup of the algorithms

It can be seen from Figure 2 that the speedup ratio of the two-layer clustering algorithm based on Mapreduce increases steadily with the increase of computing nodes, and the two-layer clustering is superior to fuzzy-rough k-means.

6 Conclusion

With the development of the need for web personalized services and intelligent recommendation, it is very important for us to gain users' interest through web mining. Based on the fuzzy rough k-means ,the two-layer clustering algorithm is proposed .This algorithm solves some problems, such as the limitations of the initial clustering center, heavy calculation burden and the problem of empty clustering. At the same time, the parallelization of the above algorithms is realized by using Mapreduce in the big data environment.This clustering method can more effectively explore the rules of user access patterns and provide users with more accurate personalized service.Subsequently, the problem of parallel improvement of Web clustering based on fuzzy and rough will be further studied, so as to better cluster the user access patterns.

Acknowledgments

This work was supported by National Natural Science Foundation of China Grant No.71572122 and Shanxi Province Natural Foundation No. 2016041017-3.

References

[1] Chien-Liang Liu,Tao-Hsing Chang, and Hsuan-Hsun Li,"Clustering documents with labeled and unlabeled documents using fuzzy semi-Kmeans," in Fuzzy Sets and Systems,vol. 221 pp.48-64,2013.
[2] Zahid Ansari, Syed Abdul Sattar , A. Vinaya Babu , and M. Fazle Azeem,"Mountain density based fuzzy approach for

discovering web usage clusters from web log data," in Fuzzy Sets and Systems,vol. 279,pp.40-63,2015.
[3] Indrajit Saha,Jnanendra Prasad Sarkar,and Ujjwal Maulik," Integrated Rough Fuzzy Clustering for Categorical data Analysis," in Fuzzy Sets and Systems,2018.
[4] Van Nha Pham, Long Thanh Ngo , and Witold Pedrycz,"Interval valued fuzzy set approach to fuzzy co-clustering for data classification," in Knowledge-Based Systems,vol. 107,pp.1-13, 2016.
[5] Shi Wenfeng, and Shang Lin,"Determing clustering number of FCM algorithm based on DTRS," in Computer Science,vol.44,no.9,pp.45-48,2017.
[6] Wu Rui, and Ning Yu-fu,"Clustering User Access Patterns based on Fuzzy Rough k-Means,"in Systems Engineering-Theory & Practice,vol. 7,pp.116-121,2007.
[7] Viktor Medvedev, Olga Kurasova, Jolita Bernatavicien, Povilas Treigys,Virginijus Marcinkevi cius,and Gintautas Dzemyda,"A new web based solution for modelling data mining processes,"in Simulation Modelling Practice and Theory,vol. 76,pp.34-46,2017.
[8] Li Yingxin,Jing Shikai,Li Xiangqian,Zhan Hongfei,Zhou Jingtao,Yang Haicheng,Zhang Hong, and Mou Xiangfeng,"Personalized knowledge service approach for cloud manufacturing based on user behaviors," in Computer Integrated Manufacturing Systems,vol. 21,no.3,pp.848-858,2015.
[9] Zhou Runwu,Li Zhiyong,Chen Shaomiao,Chen Jing, and Li Renfa," Parallel optimization sampling clustering K-means algorithm for big data processing," in Journal of Computer Applications,vol.36, no.2,pp.311-315,2016.

Xiaoli Jia is a postgraduate student of the Department of Mathematics and Computer at Shanxi Normal University, China.Her research interests include knowledge discovery and acquisition.
Rui Wu is a professor of the Department of Mathematics and Computer at Shanxi Normal University, China. She received the MS degrees in computer science from Shanxi University, in 1999, and the PhD degree in management science and engineering from the Tianjin University, China, in 2006, respectively. Her research interests include information management in the uncertain environment, knowledge discovery and acquisition.
Siying Wu is a undergraduate student of the Department of Mathematics and Computer at Shanxi Normal University, China.