A Reasonable Rough Approximation for Clustering Web Users

Duoqian Miao, Min Chen, Zhihua Wei, and Qiguo Duan

[1] Department of Computer Science and Technology,
Tongji University, Shanghai, 201804, China
[2] The Key Laboratory of Embedded System and Service Computing,
Ministry of Education, China
miaoduoqian@163.com

Abstract. Due to the uncertainty in accessing Web pages, analysis of Web logs faces some challenges. Several rough k-means cluster algorithms have been proposed and successfully applied to Web usage mining. However, they did not explain why rough approximations of these cluster algorithms were introduced. This paper analyzes the characteristics of the data in the boundary areas of clusters, and then a rough k-means cluster algorithm based on a reasonable rough approximation (RKMrra) is proposed. Finally RKMrra is applied to Web access logs. In the experiments RKMrra compares to Lingras and West algorithm and Peters algorithm with respect to five characteristics. The results show that RKMrra discovers meaningful clusters of Web users and its rough approximation is more reasonable.

1 Introduction

Web usage mining [1] can be viewed as the application of data mining techniques to any collection of Web access logs. It is a promising research field because Web user information needs are acquired by mining Web access logs. In recent years, it has also become a subtopic of Web Intelligence (WI) [2, 3]. Clustering as an important data mining technique is generally used in Web usage mining. User profiles can be established by clustering Web access logs based on some sort of similarity measures. Clustering is done so that Web users within the same cluster behave more similarly than those in different clusters. Therefore, it is very useful for Web applications, such as personalized recommendation [4], business intelligence [13], and other Web based applications [1].

However, clustering faces some challenges in Web usage mining compared to traditional data mining. Due to the uncertainty in accessing Web pages and the ease of movement from one Web page to another, the clusters tend to have vague or imprecise boundaries. Rough set theory [5, 14] as a kind of tool dealing with imprecision and incomplete knowledge is widely used in clustering Web logs. Several rough k-means cluster algorithms have been proposed, for example the rough k-means cluster algorithm introduced by Lingras and West [6] and the

N. Zhong et al. (Eds.): WImBI 2006, LNAI 4845, pp. 428–442, 2007.

refined one by Peters [7]. Some other algorithms for clustering Web users have also been proposed in the literature [8, 9, 10, 11].

The concept of the rough approximation in clusters was presented by Lingras and West. In rough clustering each cluster has two approximations, namely the lower and upper approximations. Then Peters analyzed Lingras and West cluster algorithm and pointed out some refinements. However, both of them did not illustrate why these rough approximations in clusters were introduced.

The objective of this paper is to analyze the data objects in the boundary areas of clusters. Based on the analysis, a reasonable rough approximation will be suggested. Then the rationality of the rough approximation will also be explained.

The structure of the paper is as follows. In Section 2 we introduce two rough k-means algorithms, which are Lingras and West algorithm and Peters algorithm. Then theses algorithms are analyzed in Section 3. In Section 4 we suggest a more reasonable rough approximation. Based on the rough approximation, a k-means cluster algorithm is proposed. To evaluate the performance of the algorithm, experiments are presented in Section 5. Finally, the paper concludes with a summary in Section 6.

2 Review of Existing Rough *k*-Means Cluster Algorithms

2.1 Rough Properties of the Cluster Algorithms

Rough set is a kind of mathematical tool for dealing with uncertainty. All the previous rough k-means cluster algorithms use this characteristic of rough set theory. A cluster is represented by a rough set based on a lower approximation and an upper approximation. Although the rough k-means algorithms do not verify all the properties of rough set theory, they have some basic properties as follows:

- Property 1: A data object $\mathbf{X}$ belongs to one lower approximation at most.
- Property 2: For a cluster (set) C and a data object $\mathbf{X}$, if $\mathbf{X}$ belongs to the lower approximation of C, then it also belongs to the upper approximation of C.
- Property 3: If a data object $\mathbf{X}$ does not belong to any lower approximation, then $\mathbf{X}$ belongs to two or more upper approximations. That means $\mathbf{X}$ lies in two or more boundary areas of clusters.

2.2 Existing Rough *k*-Means Cluster Algorithms

Let $\mathbf{X_n}$ represents the nth data object which is a multidimensional vector. C_k is the kth cluster (set), and its upper and lower approximation are $\overline{C_k}$ and $\underline{C_k}$ respectively. $C_k^B = \overline{C_k} - \underline{C_k}$ is the boundary area of the cluster. $\mathbf{m_k}$ represents the centroid of cluster C_k .

K-means clustering is a process of finding centroids for all clusters, and assigns objects to each cluster based on their distance from the centroids. This process

is done iteratively until stable centroid values are found. Rough k-means cluster algorithms incorporate rough sets into k-means clustering, which requires the addition of the concept of lower and upper bounds, such as Lingras and West algorithm and Peters algorithm.

Lingras and West algorithm use Eq. (1) to calculate the centroids of clusters that is modified to include the effects of lower as well as upper bounds.

$$\boldsymbol{m_k} = \begin{cases} \omega_l \sum\limits_{\mathbf{X_n} \in \underline{C_k}} \frac{\mathbf{X_n}}{|\underline{C_k}|} + \omega_b \sum\limits_{\mathbf{X_n} \in C_k^B} \frac{\mathbf{X_n}}{|C_k^B|} & \text{for} \quad C_k^B \neq \phi \\ \omega_l \sum\limits_{\mathbf{X_n} \in \underline{C_k}} \frac{\mathbf{X_n}}{|\underline{C_k}|} & \text{otherwise} \end{cases} \tag{1}$$

where ω_l is the lower weight and ω_b is the boundary weight.

The next step in Lingras and West algorithm is to design criteria to determine whether an object belongs to the upper or lower bound of a cluster. When assigning the data object $\mathbf{X_n}$ to the lower or upper approximation, we look for the centroid $\mathbf{m_s}$ closest to $\mathbf{X_n}$ firstly, and then the following set T must be determined first(see Eq. (2)).

$$T = \{t : d(\mathbf{X_n}, \mathbf{m_k}) - d(\mathbf{X_n}, \mathbf{m_s}) \leq \varepsilon \wedge k \neq s\} \tag{2}$$

- If $T \neq \phi$, then $\mathbf{X_n} \in \overline{C_t}$, $\forall t \in T$.
- Else $\mathbf{X_n} \in \underline{C_s}$.

where ε is the threshold.

Lingras and West algorithm, described above, depends on three parameters ω_l, ω_b and ε. Experimentation with various values of the parameters is able to develop a reasonable rough set clustering and it also delivers meaningful results. However, there exist some problems in the algorithm as presented by Lingas and West, such as its numerical instability and its instability in computing the number of clusters. Therefore, Peters made some improvement for the rough cluster algorithm to resolve these problems.

The rough cluster algorithm proposed by Peters use Eq. (3) to calculate the centroids of clusters.

$$\boldsymbol{m_k} = \omega_l \sum_{\mathbf{X_n} \in \underline{C_k}} \frac{\mathbf{X_n}}{|\underline{C_k}|} + \omega_u \sum_{\mathbf{X_n} \in \overline{C_k}} \frac{\mathbf{X_n}}{|\overline{C_k}|} \qquad \text{with} \qquad \omega_l + \omega_u = 1 \tag{3}$$

where ω_l is the lower weight and ω_u is the upper weight.

The next step is to forces a data object as a lower approximation for each cluster (see Eq. (4)). Then, in order to assign any one of other data objects $\mathbf{X_n}$, except the data objects satisfying Eq. (4), to the lower or upper approximation, looking for the centroid $\mathbf{m_s}$ closest to $\mathbf{X_n}$, so the set T' is determined (see Eq. (5)).

$$d(\mathbf{X_l}, \mathbf{m_s}) = \min_{n,k} d(\mathbf{X_n}, \mathbf{m_k}) \Rightarrow \mathbf{X_l} \in \underline{C_s} \wedge \mathbf{X_l} \in \overline{C_s} \tag{4}$$

$$T' = \left\{ t : \frac{d(\mathbf{X_n}, \mathbf{m_k})}{d(\mathbf{X_n}, \mathbf{m_s})} \leq \zeta \wedge k \neq s \right\} \tag{5}$$

– If $T' \neq \phi$, then $\mathbf{X_n} \in \overline{C_t}$, $\forall t \in T'$.
– Else $\mathbf{X_n} \in \underline{C_s}$.

where ζ is the threshold.

3 Comments on Existing Rough *k*-Means Algorithms

Peters analyzed Lingras and West cluster algorithm from several aspects and then put forward some problems existing in the algorithm. Based on his analysis a refined rough k-means cluster algorithm was proposed. However, there still exist some improvements to be made.

1. Computation of Centroid

As can be seen from Eq. (2), the importance of the lower and upper approximations are defined by the weight ω_l and ω_u respectively. Moreover, Peters suggested a limitation $\omega_l + \omega_u = 1$. Obviously, the weights are determined by end users and not related to the data objects in the lower or boundary area.

2. Numerical stability

Lingras and West algorithm is numerical instable since there are data constellations where $|\underline{C}| = 0$. When $|\underline{C}| = 0$, the cluster C seems to have no sure representative according to the definition of the lower approximation in rough set theory. To avoid such kind of case, Peters suggested that each cluster has at least one lower member and was forced to have a lower member in the initial cluster assignment. Therefore, it is better for a cluster algorithm to assure $|\underline{C}| \neq 0$ whether it forces the lower member for each cluster or not.

3. Interpretation issues and objective function

Peters gave two (extreme) examples of data constellation to illustrate that the objective function of relative distance between data objects was better than that of absolute distance. However, if the objective function is taken into account from other aspects, the data objects in the lower and boundary areas may be explained more intuitively.

4 A Rough *k*-Means Algorithm Based on a Reasonable Rough Approximation

4.1 Analysis of the Data Objects in the Boundary Areas of Clusters

The data objects that the cluster algorithms deal with are usually multidimensional data sets. Suppose that the data objects are in the multidimensional space now. After a rough k-means cluster algorithm is performed, several cluster means (centroids) that are the representatives of clusters are generated. These centroids are also multidimensional vectors. Note that an arbitrary object A and two cluster centroids (C and B) in the multidimensional space form a triangle, which also decide a plane (see Fig. 1). Moreover, the data objects E, D and A are in the same plane.

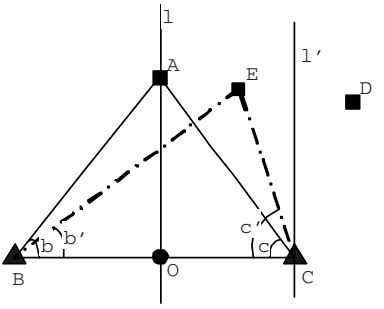

Fig. 1. The boundary area

Suppose

- A is an arbitrary data object. B and C are two cluster centroids. They are all in the multidimensional space. Moreover, C is the cluster centroid closet to A.
- O is the mid-point of the line segment from B to C.
- The straight line l is perpendicular to the line segment from B to C, so is the straight line l' .
- Angle b represents $\angle ABC$ and angle c represents $\angle ACB$. Similarly, Angle b' represents $\angle EBC$ and angle c' represents $\angle ECB$.

Given A is not an outlier. Intuitively, it is so hard to make clear whether the object A belongs to the cluster C or the cluster B when A lies in the straight line l , since the distance from A to C equals that from A to B. In this case, it is easy to find that c=b. The closer A gets to C, namely the further A is from B, the easier it is for A to be distinguished. Similarly, the larger the value of c-b is, the further A is from the boundary area (such as the data object E). Obviously, it is more intuitive to assign A to the lower approximation of the cluster C when A is in the right side of l' (such as the data object D). It can be explained that the boundary area between cluster C and cluster B gets to its largest area when c equals the maximum 90. Therefore, a more reasonable rough approximation is suggested in the following to replace the distance measure for determining the set T" :

- If $(c - b) \leq \varepsilon'$ and $c \leq \vartheta$, then $A \in \overline{C}$ and $A \in \overline{B}$
- Else $A \in \underline{C}$

Accordingly, the weights ω_l and ω_u are changed as follows:

$$\omega_l = \frac{360 - 2\vartheta}{360} \qquad \text{and} \qquad \omega_u = \frac{2\vartheta}{360}$$

where ε' and ϑ are two given thresholds. Note that $0 \leq \varepsilon', \vartheta \leq 90$. The threshold ε' defines the biggest difference between from A to C and from A to another cluster centriod B. The threshold ϑ determines the weights and reflects the maximum of the boundary area. Note that ϑ must be selected from 0 to 90.

4.2 The Proposed Algorithm

The outline of the rough k-means cluster algorithm based on a reasonable rough approximation (RKMrra) can be stated as follows:

Step 1. Initialization. Randomly assign each data object to exactly one lower approximation. By definition (Property 2, Section 2.1) the data objects in the same cluster belong to both the lower and the upper approximations of the cluster.

Step 2. Calculation of the new cluster centroids according to Eq. (2).

Step 3. Assign the data objects to the lower and upper approximations.

(i) For a given data object $\mathbf{X_n}$ determine its closest centroid $\mathbf{m_s}$:

$$d_{n,s}^{min} = d(\mathbf{X_n}, \mathbf{m_s}) = \min_{k=1,\ldots,K} d(\mathbf{X_n}, \mathbf{m_k}) \tag{6}$$

Assign $\mathbf{X_n}$ to the upper approximation of the cluster s: $\mathbf{X_n} \in \overline{C_s}$.

(ii) Determine whether $\mathbf{X_n}$ belongs to other approximations:

- Calculation of the set T":

  Step 3.1 Initialization. The set T" is set to ϕ. The set L is set to {1,2, ..., K}.

  Step 3.2 $L = L - \{s\}$. For a centroid $\mathbf{m_j}$ ($j \in L$) calculate two angles as follows:

$$\theta_s = \arccos \frac{d(\mathbf{X_n}, \mathbf{m_s})^2 + d(\mathbf{m_s}, \mathbf{m_j})^2 - d(\mathbf{X_n}, \mathbf{m_j})^2}{2d(\mathbf{X_n}, \mathbf{m_s})d(\mathbf{m_s}, \mathbf{m_j})}$$

$$\theta_j = \arccos \frac{d(\mathbf{X_n}, \mathbf{m_j})^2 + d(\mathbf{m_s}, \mathbf{m_j})^2 - d(\mathbf{X_n}, \mathbf{m_s})^2}{2d(\mathbf{X_n}, \mathbf{m_j})d(\mathbf{m_s}, \mathbf{m_j})}$$

  Step 3.3 If $(\theta_s - \theta_j) \leq \varepsilon'$ and $\theta_s \leq \vartheta$ ($0 \leq \vartheta \leq 90$), then $T" = T" \cup \{j\}$. Where ε' and ϑ are two given thresholds.

  Step 3.4 $L = L - \{j\}$. If $L \neq \phi$, continue with Step 3.2.

- If $T" \neq \phi$, then $\mathbf{X_n} \in \overline{C_t}$, $\forall t \in T"$
- Else $\mathbf{X_n} \in \underline{C_s}$

(iii) Update the weights ω_l and ω_u according to the following equations:

$$\omega_l = \frac{360 - 2\vartheta}{360} \qquad \text{and} \qquad \omega_u = \frac{2\vartheta}{360} \tag{7}$$

Step 4. Check convergence of the algorithm.

- If the algorithm has not converged, continue with Step 2.
- Else STOP.

4.3 Analysis of the Rationality of the Proposed Rough Approximation

We analyze the rationality of the proposed rough approximation from the following three aspects:

1. Computation of Centroid

As can be seen from Step 3.3, the threshold ϑ defines the width of the boundary area. Furthermore, it also decides the weights ω_l and ω_u (see Eq. (7)). Therefore the weights are closely related to the boundary area. This leads to the ease of decision made by end users or experts for the parameters ω_l and ω_u.

2. Numerical stability

Unlike the algorithms proposed by Lingras et al. or by Peters, the algorithm proposed above is numerical stable since there doesn't exist that $|\underline{C}| = 0$. Therefore, no data object need to be forced as lower members of clusters. Moreover, each cluster has definite representatives.

3. Interpretation issues and objective function

The objective function (see Step 3.3) is taken into account from the angle aspect instead of from the distance aspect. Moreover, the data objects in the lower and boundary areas are explained more intuitively.

5 Experiments and Discussion

Experiments were conducted on the Web access logs of the introductory first year course in computing science at Saint Mary's University. Lingras and West showed that the visits from students attending these courses could fall into one of the following three categories (for more details see [6]):

1. Studious: These students always download the current set of notes regularly.
2. Crammers: These students download a large set of notes just before the exam for a pre-test cramming.
3. Workers: These group of students are more interested in doing class and lab assignments than downloading the notes.

Since the students in the courses are of different educational backgrounds. Lingras and West decided to use the following five attributions representing each visitor:

1. On campus/Off campus access
2. Day time/Night time access
3. Access during lab/class days or non-lab/class days
4. Number of hits
5. Number of notes downloaded

The values for the first three attributes were either 0 or 1. The last two values were normalized to the interval [0,1] and the last attribute was the most important for clustering visitors.

The total access logs (AllData) have a total size of 21637. We selected 3000 data records (D1) randomly out of the total access logs. Similarly, we got other nine data sets (D2, D3, D4 and so on) with a size of 3000 respectively. The

eleven data sets in all were used for the following experiments. Furthermore the performance of RKMrra is compared to that of two other rough k-means cluster algorithms, which are Lingras and West algorithm and Peters algorithm. To exclude any influence of different selections of the weights, we consider these algorithms with $\omega_l = 0.7$ and $\omega_u = 0.3$, which corresponds to $\vartheta = 54$ in the following experiments. Each algorithm is repeated i-times (iteration factor). When the clustering result doesn't change any more, the cluster algorithm gets to the maximum number of iterations (i_{max}). Among the final results, the experiment with the minimal Davies-Bouldin index (D-B Index) (for more details see [12]) is considered as best.

We focus on the following aspects to evaluate the performance of RKMrra:

- In Section 5.1 we analyze the convergence speed of the algorithm.
- In Section 5.2 we investigate the selections of the thresholds.
- In Section 5.3 we analyze the stability of the algorithm.
- In Section 5.4 we discuss the initial cluster assignment of the algorithm.
- In the last analysis (Section 5.5) we compare the clustering quality of RKMrra with that of the other two algorithms.

5.1 Convergence Speed

In order to evaluate the convergence speed of the cluster algorithms, we conduct 10 experiments on 10 data sets (D1, ..., D10) for each algorithm. The thresholds are chosen $\varepsilon = 0.5, 0.6, 0.7$ for Lingras and West algorithm, $\zeta = 1.1$ for Peters algorithm and $\varepsilon' = 0.7$ for RKMrra. Table 1 shows the final number of iterations when the result of a cluster algorithm remains stable. The last column is the average number of iterations on 10 data sets for each algorithm.

There are some slightly differences among the average number of iterations of the three algorithms. The average number of iterations of Peters algorithm is the smallest, while that of RKMrra is larger and Lingras and West algorithm has the largest average number of iterations.

In general, these three algorithms have similar convergence speeds.

Note that the threshold ε is chosen as three different values (0.5, 0.6 and 0.7) for Lingras and West algorithm. However, the threshold ζ is chosen as a definite value for Peters algorithm, so is the threshold ε' for RKMrra.

We explain why we choose three different values for the threshold ε for Lingras and West algorithm from the following two aspects:

1. Lingras and West algorithm is very sensitive to the threshold ε

For example, the number of the objects in the boundary area increases too much when the threshold ε changes from 0.63 to 0.64 (see Table 2). The same

Table 1. The number of iterations

Algorithm	*D*1	*D*2	*D*3	*D*4	*D*5	*D*6	*D*7	*D*8	*D*9	*D*10	*AverIter*
Lingras	40	24	26	36	27	21	28	34	30	32	29.8
Peters	30	29	26	30	29	19	26	26	34	27	27.6
RKMrra	30	26	25	25	26	34	33	28	30	26	28.3

Table 2. The sensitivity to the threshold ε

Data set	The number of the data objects in the boundary area	ε
D1	8	0.61,0.62
	12	0.63
	2598	0.64,0.65,...

cases happen when experiments are conducted on other data sets for Lingras and West algorithm.

2. To maintain the consistency

In order to maintain the consistency of the analysis when comparing with other algorithms, we limit the number of the data objects in the boundary area within 150.

In conclusion, because of the analysis above, the threshold ε for Lingras and West algorithm must be adjusted for different data sets.

5.2 Selections of the Thresholds

To evaluate the influence of the threshold on the algorithm we use the data set (AllData) with the largest size. We analyze the selections of the thresholds from the following two aspects:

1. The range of the threshold ε'

Lingras and West algorithm and Peters algorithm suggested the selections of the threshold ε or ζ respectively. However, they did not point out the range within which ε or ζ should be selected. Therefore, the selections of ε or ζ depend on the decisions of experts or end users. Here we discuss the dependency of the thresholds of the three algorithms. The results are illustrated in Fig. 2.

Lingras and West algorithm shows good performance and similar linear characteristic with RKMrra when the value of the threshold (ε or ε') ranges from 0.2 to 1.0. However, the number of data objects in the boundary area increases too much suddenly so that Lingras and West algorithm delivers no meaningful result for $\varepsilon > 1.0$, so does the same case for Peters algorithm for $\zeta > 1.7$. In contract to Lingras and West algorithm and Peters algorithm, RKMrra still work well even though ε' becomes very large.

In order to better illustrate the range of ε' , we consider four extreme cases as follows:

- When $\varepsilon' = 90$ and $\vartheta = 90$, there exist that$|\underline{C}| > 0$ and $|C^B| > 0$.
- When $\varepsilon' = 90$ and $\vartheta = 90$, there exist that $|\underline{C}| > 0$ and $|C^B| = 0$.
- When $\varepsilon' = 0$ and $\vartheta = 90$, there exist that $|\underline{C}| > 0$ and $|C^B| = 0$.
- When $\varepsilon' = 0$ and $\vartheta = 0$, there exist that$|\underline{C}| > 0$ and $|C^B| = 0$.

where $|\underline{C}$ and $|C^B|$ are the number of data objects in the lower and boundary areas of clusters respectively.

From the four extreme cases above, we conclude that there must exist data objects in the lower area for any ε' from 0 to 90. Therefore, RKMrra needn't

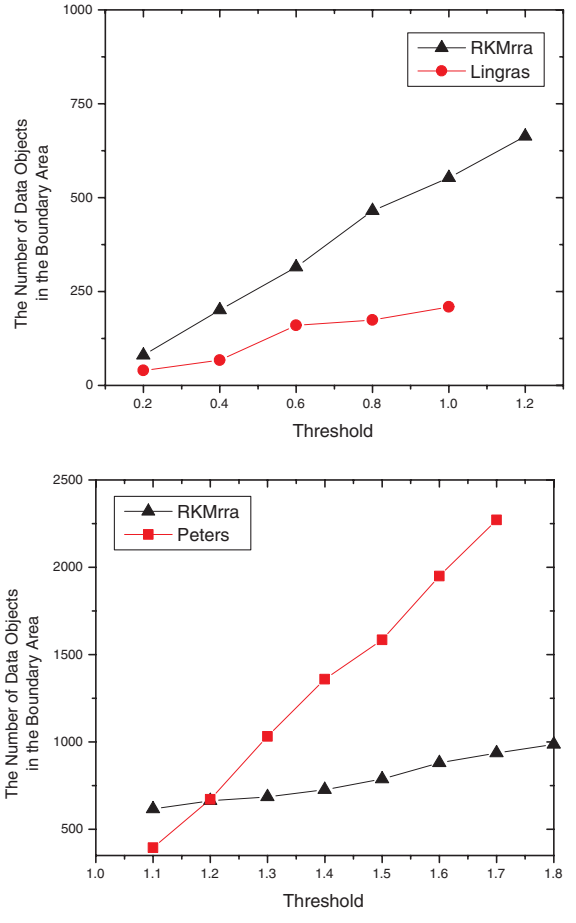

Fig. 2. Dependency on the threshold ε'

force the data objects as the lower members of clusters during the initial cluster assignment. In contrast to RKMrra, it is a necessary step for Peters algorithm (see Eq. (5)).

2. The rationality of the threshold ϑ

The threshold ϑ is used to do two things:

(1) Computing the weights. (see Eq. (7))

(2) Controlling the boundary area. (see Step 3.3)

To evaluate the dependency on the threshold ϑ of RKMrra, we remove the threshold ϑ from Step 3.3 and replace the weights (ω_l and ω_u) in Eq. (7) with the weights in Eq. (2). Table 3 illustrates how the results are influenced by the threshold ϑ . The experiment is conducted on the data set AllData.

When ε' is small, the two algorithms have the same clustering results no matter whether the threshold ϑ is removed. As ε' becomes very large, such as the maximum value (90), the Davies-Bouldin indexes of both algorithms change a

Table 3. The dependency of the threshold ϑ

Algorithm	Whether removing the parameter ϑ	ω_l	ω_u	ε'	The number of the boundary objects	D-B index
$RKMrra$	No	0.7	0.3	1.1	616	0.626
$RKMrra'$	yes					
$RKMrra$	No	0.7	0.3	1.2	663	0.628
$RKMrra'$	yes					
...						
$RKMrra$	No	0.7	0.3	90	12433	1.175
$RKMrra'$	yes			90	16532	6.766

lot. Moreover, the algorithm of removing the threshold ϑ (RKMrra') has a larger Davies-Bouldin index comparatively. Therefore, it is reasonable to suggest the threshold ϑ to compute the weights and control the boundary area for RKMrra.

In general, in comparison to Lingras and West algorithm and Peters algorithm, the parameter ϑ is reasonable suggested by RKMrra. At the same time, the thresholds (ε' and ϑ) of RKMrra are selected within a reasonable range. As far as the selection of the thresholds within the specified range is concerned, the setting of the thresholds has actually been relaxed. Furthermore, during the initial cluster assignment, it isn't a necessary step to force the data objects as the lower members of clusters for RKMrra.

5.3 Stability

We use 10 data sets (D1, ..., D10) to conduct the experiments. The thresholds are chosen as in Section 5.1. Since the algorithms adjust the assignment of the data objects gradually, we find that the Davies-Bouldin index of each algorithm changes a lot in the experiments. Before the clustering result remains stable, the Davies-Bouldin index sometimes increases and sometimes decreases. That means the Davies-Bouldin index does not always monotonously increase or decrease. For example, the Davies-Bouldin index (D-B index) of an algorithm increases from $i = 5$ to $i = 10$ firstly, then decreases from $i = 10$ to $i = 15$ and increases again from $i = 15$ to $i = i_{max}$. If we use the number of jumps to record the change of the D-B index, then the number of jumps of the example equals 3. Here we analyze the stability of the algorithm from the number of jumps of the D-B index in relation to iteration factors.

Table 4 shows the jumps of the D-B index of the algorithms run on ten different data sets. The iteration factor is set to $i = 5, 10, 15, 20, \ldots, i_{max}$ respectively to

Table 4. The number of jumps of the D-B index

$Algorithm$	$D1$	$D2$	$D3$	$D4$	$D5$	$D6$	$D7$	$D8$	$D9$	$D10$	$AvJump$
$Lingras$	1	2	3	4	1	1	2	2	2	3	$\frac{21}{10} = 2.1$
$Peters$	1	1	2	3	1	1	2	1	2	3	$\frac{17}{10} = 1.7$
$RKMrra$	1	2	1	1	4	1	1	2	2	1	$\frac{16}{10} = 1.6$

calculate the number of jumps of the D-B index. The last column of the table is the average number of jumps.

Obviously, RKMrra with the smallest average value has the best stability. Peters algorithm has a close value with the proposed one and Lingras and West algorithm is the most instable.

5.4 The Initial Cluster Assignment

The number of the data objects in the boundary area (the boundary objects) changes a lot for different iteration factors (see Table 5). We use the ratio of the boundary objects for $i = 5(i_{min})$ to those for $i = i_{max}$ to demonstrate this change. Figure 3 shows the ratios of different algorithms run on five data sets (D1, ..., D5). The ratios on other five data sets (D6, ..., D10) are similar to the ratios on the five data sets (D1, ..., D5).

As can be seen from Fig. 3, there is a significant difference between RKMrra and the other two algorithms. The ratio of RKMrra is either greater or less than one, while the ratios of the other two algorithms are both less than or equal one. This shows that the number of the boundary objects of RKMrra for i_{max} is greater than those for i_{min} on some data sets. However, there don't exist such cases for Lingras and West algorithm and Peters algorithm.

Table 5. The ratio of the boundary objects for i_{max} to those for i_{min}

Algorithm	D1		D2		D3		D4		D5	
	i_{max}	i_{min}	i_{max}	i_{min}	i_{max}	i_{min}	i_{max}	i_{min}	i_{max}	i_{min}
Lingras	12	109	9	98	14	77	4	12	48	117
Peters	40	40	47	98	41	63	47	85	70	84
RKMrra	52	69	55	39	44	78	57	94	66	102

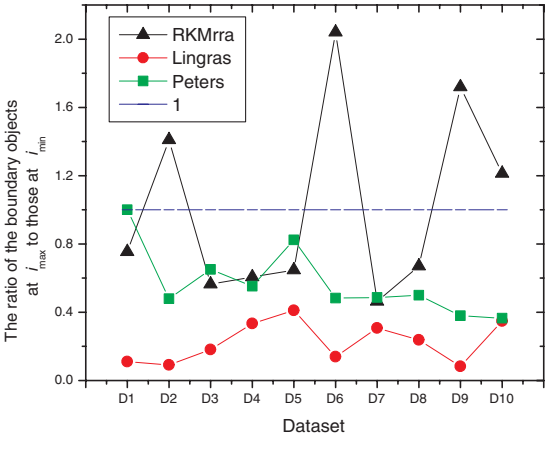

Fig. 3. The ratio of the boundary objects for i_{max} to those for i_{min}

The reason is that RKMrra just adjusts the assignment of the data objects in the boundary area for each iteration factor. In contrast to RKMrra, Lingras and West algorithm and Peters algorithm need to restrict the boundary area within a certain range for the initial cluster assignment firstly, then select the data objects around the limited boundary area.

5.5 Clustering Quality

As introduced above, among the rough k-means cluster algorithms, the one with the minimal Davies-Bouldin index is considered as best. In order to evaluate the clustering quality, We use 10 data sets (D1, . . . , D10) to conduct the experiments. The thresholds are chosen as in Section 5.1. An interesting phenomenon is found among the results of the experiments: The Davies-Bouldin index increases with the number of data objects in the boundary area. That shows the clustering quality of an algorithm is better as the data objects in the boundary area decrease.

The boundary area in rough k-means clustering is also referred to as the security zone [7]. Because the data objects in the boundary area have the possibilities to belong to more than one clusters and require a second look before making a final decision. Hence, the cluster algorithm with the maximum number of data objects in the boundary area indicates the highest security requirements.

Strictly speaking, unlike the classical k-means cluster algorithms, the rough k-means can be interpreted as two layer interval clustering approaches with lower and upper approximations. Therefore, there isn't a kind of cluster validity criterion for the rough k-means cluster algorithms virtually.

Taken into consideration two factors (the D-B index and the number of the boundary objects) analyzed above, here we use the ratio of the Davies-Bouldin index to the number of data objects in the boundary area to evaluate the clustering quality.The results are depicted in the Fig. 4. The algorithm with the minimum ratio is considered as best.

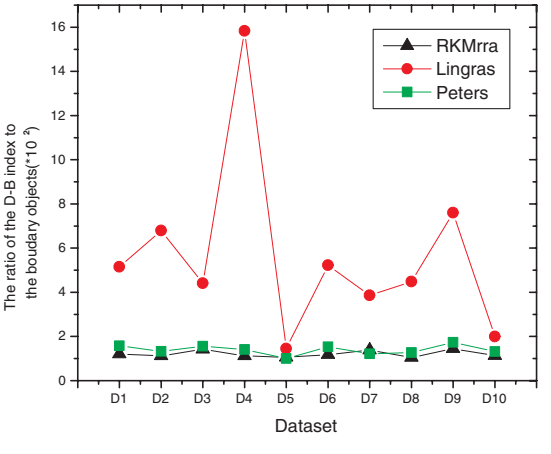

Fig. 4. The clustering quality

Obviously, RKMrra has the similar clustering quality with Peters algorithm. In contract to Lingras and West algorithm, the ratios of RKMrra and that of Peters algorithm change smoothly and have small ratios. Therefore, RKMrra and Peters algorithm have better clustering quality.

6 Conclusion

In this paper we introduce Lingras and West rough cluster algorithm and Peters refined one at first and then comment on them to put forward some problems. In order to solve these problems, the characteristics of the data objects in the boundary area are analyzed. This led to the suggestion of a reasonable rough approximation. The reasonable rough approximation is proposed from the angle aspect, instead of from the distance aspect, although there are some relationships between them. Based on the reasonable rough approximation suggested, a rough k-means cluster algorithm is proposed.

A challenge of the rough k-means is resolved to some extent: the selection of the initial parameters ω_l and ω_u. Since the parameters are limited within a reasonable range according to the threshold ϑ. Furthermore, they are closely related with the width of the boundary area. By tuning the initial parameters, experts can interpret the clustering results according to the given width of the boundary area.

At last, RKMrra is applied to Web logs. The paper describes the design of the experiments to compare RKMrra with Lingras and West algorithm and Peters algorithm with respect to five characteristics. The results show that RKMrra discovers meaningful clusters of Web users and its rough approximation is more reasonable.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No.60475019) and the Ph.D. programs Foundation of Ministry of Education of China (No.20060247039). We are very grateful to Pawan Lingras for supplying with the Web access logs.

References

1. Cooley, R., Mobasher, B., Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. Tools with Artificial Intelligence. In: Proceedings of the Ninth IEEE International Conference, pp. 558–567. IEEE-CS Press, Los Alamitos (1997)
2. Zhong, N., Yao, Y., Ohsuga, S., Liu, J. (eds.): WI 2001. LNCS (LNAI), vol. 2198, pp. 1–17. Springer, Heidelberg (2001)
3. Zhong, N., Liu, J., Yao, Y.Y. (eds.): Special issue on Web Intelligence (WI). IEEE Computer 35(11) (2002)
4. Ji, J., Liu, C., Sha, Z., Zhong, N.: Online Personalized Recommendation Based on a Multilevel Customer Model. International Journal of Pattern Recognition and Artificial Intelligence 19(7), 895–917 (2005)

5. Pawlak, Z.: Rough Set Theory and Its Applications to Data Analysis. Cybernetics and Systems. An International Journal 29, 661–688 (1998)
6. Lingras, P., West, C.: Interval Set Clustering of Web Users with Rough K-means. Journal of Intelligent Information System 23(1), 5–16 (2004)
7. Peters, G.: Some Refinement of K-means Clustering. Pattern Recognition 39, 1481–1491 (2006)
8. De Kumar, S., Radha Krishna, P.: Clustering Web Transactions Using Rough Approximation. Fuzzy Set and Systems 148, 131–138 (2004)
9. Mitra, S.: An Evolutionary Rough Partitive Clustering. Pattern Recognition Letters 25, 1439–1449 (2004)
10. Asharaf, S., Murty, M.N., Shevade, S.K.: Rough Set Based Incremental Clustering of Interval Data. Pattern Recognition Letters 27, 515–519 (2006)
11. Hogo, M., Snorek, M., Lingras, P.: Temporal Versus Latest Snapshot Web Usage Mining Using Kohonen Som and Modified Kohonen Som Based on the Properties of Rough Sets Theory. International Journal on Artificial Intelligence Tools 13(3), 569–592 (2004)
12. Bezdek, J.C., Pal, N.R.: Some New Indexes of Cluster Validity. IEEE Trans. Systems Man Cybernet Part-B 28, 301–315 (1998)
13. Kohavi, R.: Mining e-Commerce Data: the Good, the Bad, and the Ugly. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 1(1), pp. 5–32 (2001)
14. Yao, Y.Y.: Information Granulation and Rough Set Approximation. International Journal of Intelligent Systems 16(1), 87–104 (2001)