International Journal of Advanced Science and Technology
Volume 5, April, 2009

An Efficient Approach for Clustering Web Access Patterns from Web Logs

Peilin Shi
Department of Mathematics, Taiyuan University of Technology
Taiyuan, Shanxi 030024, China
Donglingzhen@eyou.com

Abstract

The interests of web users can be revealed by their visited web pages and time duration on these web pages during their surfing. Time duration on a web page is characterized as a fuzzy linguistic variable because linguistic variable makes users easily understand the expression of time duration and can disregard subtle difference between two time durations. Each web access pattern from web logs is transformed as corresponding fuzzy web access pattern, which is a fuzzy vector composed of fuzzy linguistic variables or 0. Each element in fuzzy web access patterns represents visited web page and time duration on this web page. This paper proposed a rough k-means clustering algorithm based on properties of rough variable to group the gained fuzzy web access patterns. Finally, an example and experiment is provided to illustrate the clustering process. Using this approach, users can effectively mine web logs records to discover interesting user access patterns.

Keywords: *clustering, fuzzy variable, rough variable, user access patterns*

1. Introduction

The World Wide Web (WWW) not only provides enormous profits but tremendous challenges for web sites designers and runners. If a particular web site doesn't satisfy the needs of users in a relatively short period of time, the users will quickly move on to another web site [5]. Therefore, it is very important to understand the needs and characteristics of web users. Web usage mining is an application of data mining algorithm to web logs to find trends and regularities in web users' traversal patterns. The results of web usage mining have been used to improve web site design and web server system performance.

Three important aspects of web usage mining, namely clustering, association, and sequential analysis are often used to study important characteristics of web users. Web clustering involves finding natural groupings of web resources or web users. However, there exist some important differences between clustering in conventional applications and clustering in web mining. The patterns from web data are non-numerical, thus Runkler and Beadek [13] proposed relational clustering method to group non-numerical web data. Furthermore, due to a variety of reasons inherent in web browsing and web logging, the likelihood of bad or incomplete data is higher than conventional applications. The clusters tend to have vague or imprecise boundaries [5]. A pattern may belong to more than one candidate clusters by different degrees of the memberships.

Fuzzy theory provides a natural framework for the process in dealing with uncertainty and impreciseness. Several fuzzy learning algorithms for inducing rules from given sets of data have been designed and used to good effect with specific domains, such as [16]. Arotaritei et al. [17] provides asurvey of the available literature on fuzzy Web mining.

Therefore, the role of soft computing technique such as fuzzy theory and rough theory is highlighted. Krishnapuram and Hathaway [2,3] proposed to use fuzzy set theory to group web users into several disjoint clusters. Some other researchers tried to explore clustering method by another soft computing technique, rough theory. The upper and lower approximations of rough sets are used to model the clusters [1,4,5,9]. De [1] tries to use rough approximation to cluster web transactions. Lingras [4] applied the unsupervised rough set clustering based on genetic algorithms to group web visitors. Later, Lingras [5] proposed a modified rough k-means method to cluster web visitors. Mitra [9] proposed a clustering algorithm based on rough c-means and genetic algorithm.

This paper uses a rough k-means clustering method in fuzzy environment to group web access patterns from web logs. A web access pattern represents a unique surfing behavior of a web user, which can be denoted by a set $s_i = \{(url_{i1}, t_{i1}), (url_{i2}, t_{i2}), ...(url_{il}, t_{il})\} (1 \le i \le m)$, where url_{ik} denotes k th visited web page and t_{ik} denotes the time duration on url_{ik}, l is the number of visited web pages during a surfing, m is the number of web access patterns extracted from web logs. If a web page appears in several web access patterns, this implies that these web users show common interests on this web page. However, the difference of time durations on this web page indicates that they show different degrees of interests on this web page during their surfing. If a web user takes more time browsing a web page, we say that he show more interests on it. According to the above consideration, whether a web page is visited or not and the time duration on it, should be considered as two important factors when web access patterns are grouped into several disjoint classes. Because the subtle difference between two time durations can be disregarded, time duration on a web page is depicted by a fuzzy linguistic variable. Then web access patterns can be transformed as fuzzy vectors with the same length. Each element in fuzzy vector represents visited web page and time duration on this web page. Thus we cluster these gained fuzzy web access patterns into several groups. Because these clusters tend to have ambiguous boundaries, a cluster is depicted by a rough variable. A rough k-means method based on the properties of rough variable is adopted to cluster fuzzy web access patterns. The clustering process is illustrated by an example and an experiment shows the clustering result.

The rest of the paper is organized as follows. Section 2 describes the basic notion of fuzzy variable. Rough variable theory is presented in section 3. The algorithm of clustering web access patterns is proposed using rough k-means method. An example and experimental results are presented in section 4 and section 5 respectively. Finally we conclude in section 6.

2. Reviewal of fuzzy variable and rough variable

2.1. Fuzzy variable

The concept of a fuzzy set was first introduced by Zadeh [15] in 1965. And then many researchers such as Nahmias [10] and Liu [7] enriches the fuzzy theory. In this section, some basic concepts and results of fuzzy variable are reviewed.

Definition 1 *(Nahmias [10] and Liu [7]) A fuzzy variable ξ is defined as a function from a possibility space $(\Theta, P(\Theta), \mathrm{Pos})$ to the set of real numbers, where Θ is a universe, $P(\Theta)$ is the power set of Θ, and Pos is a possibility measure defined on $P(\Theta)$.*

The *possibility*, *necessity*, and *credibility* of a fuzzy event $\{\xi \ge r\}$ can be represented by

$$\begin{aligned} \mathrm{Pos}\{\xi \geq r\} &= \sup_{u \geq \mathrm{r}} \mu_{\xi}(u) \\ \mathrm{Nec}\{\xi \geq r\} &= 1 - \sup_{u < r} \mu_{\xi}(u) \\ \mathrm{Cr}\{\xi \geq r\} &= \frac{1}{2}[\mathrm{Pos}\{\xi \geq r\} + \mathrm{Nec}\{\xi \geq r\}] \end{aligned} \tag{1}$$

respectively, where μ is the membership function of ξ.

The credibility of a fuzzy event is defined as the average value of its possibility and necessity. Thus the credibility measure is self dual. A fuzzy event may fail though its possibility reaches 1 and may hold even though its necessity is 0. However, the fuzzy event must hold if its credibility is 1 and fail if its credibility is 0. It will play an important role in the definition of expected value operator.

Definition 2 *(Liu and Liu [8]) The expected value of a fuzzy variable ξ is defined as*

$$E[\xi] = \int_0^{\infty} \mathrm{Cr}\{\xi \geq \mathrm{r}\}\mathrm{dr} - \int_{-\infty}^{0} \mathrm{Cr}\{\xi < \mathrm{r}\}\mathrm{dr} \tag{2}$$

provided that at least one of two integrals is finite.

Liu and Liu [8]designed a method to calculate the expected value of a fuzzy variable ξ.

Case 1: Let ξ be a discrete fuzzy variable with membership function $\mu(a_i) = \mu_i$ for i=1,2, ⋯,N. Without loss of generalization, we assume that $a_1 < a_2 < \cdots < a_n$. Definition 2 implies that

$$E[\xi] = \sum_{i=1}^{N} \varpi_i \times a_i$$

where $\varpi_1 = \frac{1}{2}(\mu_1 + \max_{1 \leq i \leq N} \mu_i - \max_{1 \leq j \leq N} \mu_j)$

$\varpi_i = \frac{1}{2}(\max_{1 \leq i \leq i} \mu_j - \max_{1 \leq j < i} \mu_j + \max_{i \leq j \leq N} \mu_j - \max_{i < j \leq N} \mu_j)$ $\quad 2 \leq i \leq N-1$

$\varpi_N = \frac{1}{2}(\max_{1 \leq i \leq N} \mu_i - \max_{1 \leq j < N} \mu_j + \mu_N)$

Case 2: Let ξ be a continuous fuzzy variable with a membership function μ. According to Liu and Liu [8], in order to estimate the expected value $E[\xi]$, we sample N points a_i uniformly from the δ-level set of ξ for i=1,2, ⋯,N. Thus we get a new discrete fuzzy variable ξ' and $\mu_{\xi'}(a_i) = \mu_{\xi}(a_i)$ for i=1,2, ⋯,N. We can calculate $E[\xi']$ by the method in case 1. Then we use $E[\xi']$ as the estimation of $E[\xi]$ provided that N is sufficiently large.

Liu and Liu [8] also gave the solving methods for some special fuzzy variables.

Example 1(Liu and Liu [8]): The expected value of a trapezoid fuzzy variable $\xi(r_1 + r_2 + r_3 + r_4)$ is defined as follow

$$E[\xi] = \frac{1}{4}(r_1 + r_2 + r_3 + r_4) \tag{3}$$

The expected value of a fuzzy variable can characterize this fuzzy variable by numerical value.

2.2. Rough variable

The notion of rough set theory was introduced by Zdzislaw Pawlak [11] in the early 1980s for dealing with the classification analysis of data tables. It has been proved to be an excellent mathematical tool dealing with vague description of objects. A fundamental assumption is that any object from a universe is perceived through available information, and such information may not be sufficient to characterize the object exactly. One way is the approximation of a set by other sets. Thus a rough set may be defined by a pair of crisp sets, called the lower and the upper approximations, which are originally produced by an equivalence relation.

Trust theory is the branch of mathematics that studies the behavior of rough events. Liu [6] gave the definition of the rough variable and basic rough variable theory based on this trust theory.

In order to provide an axiomatic theory to describe rough variable, Liu [6] gave four axioms.

Let Λ be a nonempty set. $\mathcal{A}$ a σ-algebra of subsets of Λ, Δ an element in $\mathcal{A}$, and π a real-valued set function on $\mathcal{A}$. The four axioms are listed as follows:

Axiom 1. $\pi\{\Lambda\} < +\infty$.
Axiom 2. $\pi\{\Delta\} > 0$.
Axiom 3. $\pi\{A\} \geq 0 \quad for \quad any \quad A \in \mathcal{A}$.
Axiom 4. For every countable sequence of mutually disjoint events $\{A_i\}_{i=1}^{\infty}$ we have

$$\pi\{\bigcup_{i=1}^{\infty} A_i\} = \sum_{i=1}^{\infty} \pi\{A_i\} \tag{4}$$

In facts, the set function π satisfying the four axioms is clearly a measure. Furthermore, the triplet (Λ ,$\mathcal{A}$,π) is a measure space.

Definition 3 ([6]) *Let Λ be a nonempty set, $\mathcal{A}$ a σ-alaebra of subsets of Λ, Δ an element in $\mathcal{A}$, and π a set function satisfying the four axioms. Then (Λ, Δ, $\mathcal{A}$,π) is called a rough space.*

Definition 4 ([6]) *A rough variable ξ is a measurable function from the rough space (Λ, Δ, $\mathcal{A}$,π) to the set of real numbers. That is, for every Borel set $\mathcal{B}$ of $\Re$, we have*

$$\{\lambda \in \Lambda \mid \xi(\lambda) \in B\} \in \mathcal{A} \tag{5}$$

The lower and the upper approximation of the rough variable ξ are then defined as follows,

$$\underline{\xi} = \{\xi(\lambda) \mid \lambda \in \Delta\} \tag{6}$$

$$\overline{\xi} = \{\xi(\lambda) \mid \lambda \in \Lambda\} \tag{7}$$

Remark 1 Since $\Delta \subset \Lambda$, it is obvious that $\underline{\xi} \subset \overline{\xi}$.

The lower approximation $\underline{\xi}$ is the set of all the objects that are definitely part of the rough variable ξ. The upper approximation $\overline{\xi}$ is the set of all the objects that possibly belong to the rough variable ξ.

3. Clustering web access patterns by rough *k*-means method in fuzzy environment

In this section, a rough k-means method in fuzzy environment is provided to cluster web access patterns from web logs.

Let $W = \{Url_1, Url_2, \cdots Url_n\}$ be the set of distinct n web pages visited by all users. The web access pattern representing surfing behavior of i th user can be denoted by:

$s_i = \{(Url_{i_1}, t_{i_1}), (Url_{i_2}, t_{i_2}), \cdots, (Url_{i_p}, t_{i_p})\}$ $(1 \le i \le m)$ $(1 \le p \le n)$ where $Url_{i_k} \in W$ and p is the number of web pages visited by i th user. Another web access pattern representing surfing behavior of jth user can be denoted by $s_j = \{(Url_{j_1}, t_{j_1}), (Url_{j_2}, t_{j_2}), \cdots, (Url_{j_q}, t_{j_q})\}$ $(1 \le j \le m)$ $(1 \le q \le n)$ where $Url_{j_k} \in W$ and q is the number of web pages visited by j th user. If $(Url_{i_a}, t_{i_a}) \in s_i$ $(1 \le a \le p)$ and $(Url_{i_b}, t_{i_b}) \notin s_i$, this implies that i th user shows certain degree of interest on Url_{i_a} but has no interest on Url_{i_b} .

If $(Url_{i_a}, t_{i_a}) \in s_i$ $(1 \le a \le p)$, $(Url_{j_c}, t_{j_c}) \in s_j$ $(1 \le c \le q)$, $Url_{i_a} = Url_{j_c} = Url_k \in W$ $(1 \le k \le n)$, we say that two web users show common interest on the same web page Url_k. But $t_{i_a} \ne t_{j_c}$ implies they show their interests by different degrees. $t_{i_a} > t_{j_c}$ implies that i th user shows more interests on the web page Url_k than j th user does. However, the subtle difference between t_{i_a} and t_{j_c} (for example, t_{i_a} =56s, t_{j_c} =60s) can be disregarded. Thus time duration on a web page is characterized as a fuzzy linguistic variable, such as *short* , *middle* , *long* etc. In order to easily compute the similarity/difference of web access patterns, each web access pattern is transformed as a fuzzy vector with the same length. Each element in a fuzzy vector is a fuzzy linguistic variable or 0. It represents visited web page and time duration on this web page during a surfing.

3.1. Characterizing user access patterns as fuzzy user access patterns

Suppose there are m users and user transactions $S = \{s_1, s_2, \cdots, s_m\}$, where $s_i (1 \le i \le m)$ discloses a unique surfing behavior of i th user.

Let $W = \{Url_1, Url_2, \cdots Url_n\}$ be the union set of distinct n web pages visited by users,

$U = \{(Url_1, t_{1_1}), \cdots, (Url_1, t_{1_g}), \cdots, (Url_n, t_{n_1}), \cdots, (Url_n, t_{n_h})\}$ be the union set of $s_i (1 \le i \le m)$, where g is the number of all time durations on web page Url_1, h is the number of all time durations on web page Url_n, n is the number of all different web pages visited by users.

Each pattern $s_i \in S$ is a non-empty subset of U . Here, the temporal order of visited web pages has not been taken into account.

Web access pattern $s_i \in S (1 \le i \le m)$ can be represented as a vector

$$V_i = < v_{i1}^t, v_{i2}^t, \cdots, v_{in}^t >,$$

$$\text{where} \quad v_{ik}^t = \begin{cases} t_{ik}, & (Url_k, t_{ik}) \in s_i \\ 0, & \text{otherwise,} \end{cases} \quad (1 \le k \le n). \tag{8}$$

Thus, each pattern $s_i(1 \le i \le m)$ is a real numerical vector with the same length n. Furthermore, time duration with real value is depicted by a fuzzy linguistic variable, which makes people more understandable and can ignore subtle difference between time durations.

All time duration on web pages are clustered into r different fuzzy region according to the method introduced in Wang [14]. Each fuzzy region is characterized as a fuzzy linguistic variable. Their membership functions can be gained by the simulation method [14]. Also the membership functions of time duration can be provided by experienced experts. Assume the membership functions of time duration are shown in Figure 1.

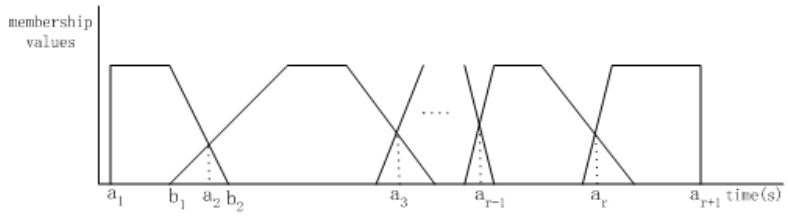

Figure 1. The membership functions of time duration

Assume the first fuzzy region is characterized as a trapezoid fuzzy variable $\xi_1(a_1, a_1, b_1, b_2)$, the last fuzzy region is characterized as ξ_n. From Figure 1, we can get the relation between real numerical v_{ik}^t and fuzzy linguistic variable $\lambda_{ik}(1 \le i \le m)(1 \le k \le n)$, which is as follows:

$$\lambda_{ik} = \begin{cases} 0, & v_{ik}^t = 0 \\ \xi_1, & a_1 \le v_{ik}^t \le a_2 \\ \xi_2, & a_2 < v_{ik}^t \le a_3 \\ \vdots & \\ \xi_r, & a_r < v_{ik}^t \le a_{r+1}, \end{cases} \tag{9}$$

where $\xi_j(1 \le j \le r)$ is the corresponding fuzzy linguistic variable.

Each numerical v_{ik}^t in a numerical vector V_i is transformed as the corresponding fuzzy linguistic variable or 0 according to (8). Thus a fuzzy web access pattern can be denoted as follows

$$f_{vi} = < \lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{in} >, \tag{10}$$

where $\lambda_{ik} \in \{0, \xi_1, \xi_2, \cdots, \xi_r\}(1 \le k \le n)$.

3.2. Algorithm for clustering fuzzy web access patterns based on rough k-means

Assume there exists m web access patterns in user transactions S, $S = \{s_1, s_2, \cdots, s_m\}$. Given any two web access patterns $s_i (1 \le i \le m)$ and $s_j (1 \le j \le m)$, according to the section 3.1 they can be denoted as follows

$$f_{vi} =< \lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{in} >,$$

and

$$f_{vj} =< \lambda_{j1}, \lambda_{j2}, \cdots, \lambda_{jn} >,$$

where $\lambda_{ik} \in \{0, \xi_1, \xi_2, \cdots, \xi_r\}$,and $\lambda_{jk} \in \{0, \xi_1, \xi_2, \cdots, \xi_r\}$, $(1 \le k \le n)$.

Their sum can be defined as

$$sum(s_i, s_j) \approx sum(f_{vi}, f_{vj}) =< \lambda_{i1} + \lambda_{j1}, \lambda_{i2} + \lambda_{j2}, \cdots, \lambda_{in} + \lambda_{jn} > \quad (11)$$

The distance between s_i and s_j can be defined as

$$d(s_i, s_j) \approx d(f_{vi}, f_{vj}) \approx \sqrt{\frac{\sum_{k=1}^{n} (E[\lambda_{ik}] - E[\lambda_{jk}])^2}{n}}. \quad (12)$$

This paper adopts a rough \$k\$-means algorithm for clustering web access patterns. Each web access patterns is transformed as corresponding fuzzy web access pattern. Lingras and West [5] considered each cluster as an interval or rough set. Here, i th cluster is characterized as a rough variable $\eta_i (1 \le i \le k)$ defined on a measurable rough space (S, S_i, A_i, π_i), where $S_i \subset S$. Then the centroid m_i of i th cluster is computed as

$$m_i = \begin{cases} \omega_{low} \dfrac{\sum_{f_{vj} \in \underline{\eta_i}} f_{vj}}{|\underline{\eta_j}|} + \omega_{up} \dfrac{\sum_{f_{vj} \in (\overline{\eta_i} - \underline{\eta_i})} f_{vj}}{|\overline{\eta_i} - \underline{\eta_i}|} & \overline{\eta_i} - \underline{\eta_i} \ne Null; \\ \omega_{low} \dfrac{\sum_{f_{vj} \in \underline{\eta_i}} f_{vj}}{|\underline{\eta_j}|} & otherwise, \end{cases} \quad (13)$$

where the parameter $\omega_{low} / \omega_{up}$ controls the importance of the patterns lying within the lower/up approximation of a cluster in determining its centroid. $0.5 < \omega_{low} < 1$ and $\omega_{up} = 1 - \omega_{low}$. $|\underline{\eta_i}|$ indicates the number of patterns in the lower approximation of i th cluster, while $|\overline{\eta_i} - \underline{\eta_i}|$ is the number of patterns in the rough boundary lying between the two approximations $\underline{\eta_i}$, $\overline{\eta_i}$. Thus, the centroid m_i of i th cluster is a real-valued vector denoted by $m_i = \langle c_{i1}, c_{i2}, ..., c_{in} \rangle$.

The distance between pattern $s_l (1 \le l \le m)$ and the centroid m_i is defined as follows

$$d(s_l, m_i) \cong d(f_{vl}, m_i) \cong \sqrt{\frac{\sum_{k=1}^{n} (E[\lambda_{LK}] - c_{ik})^2}{n}} \quad (14)$$

If the distance between $d(s_l, m_i)$ and $d(s_l, m_j)$ $1 \le l \le m$ is less than a given threshold, this implies that pattern s_l doesn't crisply belong to a cluster. It may belong to the upper approximations of i th cluster and i th cluster. If there doesn't exist $d(s_l, m_i) - d(s_l, m_j)$ less than given threshold and $d(s_l, m_i)$ is minimum over the k clusters, pattern s_l must belong to the lower approximation of i th cluster. Its membership to i th cluster is crisp. Evidently, there exist overlaps between clusters. A web access pattern s_l can be part of at most one lower approximation. If $s_l \in \overline{\eta_i}$ of i th cluster, then simultaneously $s_l \in \underline{\eta_i}$. If s_l is not a part of any lower approximation, then it belongs to two or more upper approximations.

Assume these patterns $\{s_1, s_2, s_3, s_4, \ldots s_m\}$ are grouped into k clusters, cluster scheme $C = \{C_1, C_2 \ldots C_m\}$, where $C_I = \{\overline{\eta_i}, \underline{\eta_i}\}$ $(1 \le i \le k)$.

Algorithm

Input: User transactions S composed of m different web access patterns $\{s_1, s_2, s_3, s_4, \ldots s_m\}$, membership functions of time duration, threshold $\delta \in [0,1]$, parameters ω_{low}, ω_{up}

Output: Cluster scheme C.

step1: start.

step2: For each $s_i \in S$, denote s_i by a corresponding fuzzy vector f_{vi} according to the method introduced in section 3.1. Each element in fuzzy vector f_{vi} is a fuzzy linguistic variable or 0.

step3: Assign initial means m_i for the k clusters. Here, choose randomly k web access patterns as the initial means.

step4: For each fuzzy web access pattern $f_{vl}(1 \le l \le m)$ do

For i=1 to k do

Compute (f_{vl}, m_i) according to Eq.(14).

step5: For each pattern $s_l(1 \le l \le m)$ do

If (f_{vl}, m_i)-d(f_{vl}, m_j) is less than some threshold δ then $s_l \in \overline{\eta_i}$ and $s_l \in \overline{\eta_j}$

else if the distance d(f_{vl}, m_i) is minimum over the k clusters, then $s_l \in \overline{\eta_i}$.

step6: $C_1 = \{\underline{\eta_1}, \overline{\eta_1}\}, \cdots, C_k = \{\underline{\eta_k}, \overline{\eta_k}\}$.

step7: Compute new mean for each cluster using Eq.(13).

step8: Repeat steps 4-7 until convergence, i.e., there are no more new assignments.

step9: output C.

step10: stop.

4. An example

An example is provided to illustrate this clustering process using rough k means method in fuzzy environment. All users browsing information is stored in web logs. There are several preprocessing tasks that must be performed prior to applying clustering algorithms to the data collected from web logs. As for our algorithm, we just need data cleaning and simple session identification. Data cleaning refers that all data irrelevant to the algorithm, such as the files with the suffixes of gif, jpeg, jpg, map, swf, cgi, ect., are deleted because only the HTML files are involved in our research. Session identification refers that a session is end if time duration on a web page is above a certain threshold. The preprocessed web access data is shown in Table 1. Each element $(U_r l_{ik}, t_{ik})$ in a web record represents the visited web page and time duration on this web page during the surfing by i th user.

Assume the membership functions of time duration on web pages by experts systems are shown in Figure 2.

Time durations are grouped into three fuzzy regions, one of which is characterized as a fuzzy linguistic variable. From Figure 2, we can get three fuzzy linguistic variables *short* (0,0,30,60), *middle* (30,60,90,120) and *long* (90,120,150,150). Their expected values can be gained by Eq(3).

Table1.User access patterns from the log data

Client Id	Browsing sequences
1	(A,30),(B,42),(D,118),(E,91)
2	(A,92),(B,89),(F,120)
3	(A,50),(B,61),(D,42),(G,98),(H,115)
4	(A,70),(C,92),(G,85),(H,102)
5	(A,40),(B,35),(D,112)
6	(A,52),(B,89),(G,92),(H,108)

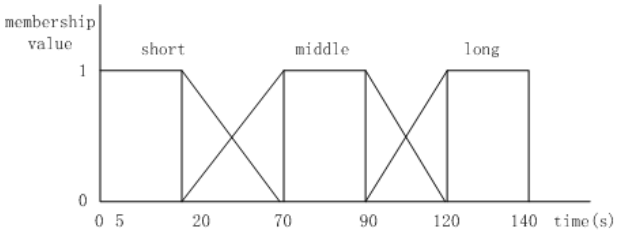

Figure 2. The membership functions of the time duration

(3). The following relations between real numeric time duration v_{ik}^t with fuzzy linguistic variable λ_{ik} can also be gained from Figure 2.

$$\lambda_{ik} = \begin{cases} 0, & v_{ik}^t = 0 \\ short, & 0 \le v_{ik}^t \le 45 \\ middle, & 45 < v_{ik}^t \le 105 \\ long, & 105 < v_{ik}^t \le 150 \end{cases} \quad (15)$$

Let $S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$ be the set of user transactions, U be the union set of distinct items accessed from user transactions S and $W = \{A, B, C, D, E, F, G, H\}$ be the union set of distinct web pages visited by all users.

The web access pattern $s_i \in S(i=1,2,\cdots,6)$ can be represented as a fuzzy vector according to the method introduced in section 3.1. Each element in fuzzy vector is a fuzzy linguistic variable among $\{short, middle, long\}$ or 0.

Similarly, the six web access patterns from Table 1 can be denoted by as follows:

$s_1 =< short, \quad short, \quad 0, \quad long, \quad middle, \quad 0, \quad 0, \quad 0 >$
$s_2 =< middle, \quad middle, \quad 0, \quad 0, \quad 0, \quad long, \quad 0, \quad 0 >$
$s_3 =< middle, \quad middle, \quad 0, \quad short, \quad 0, \quad 0, \quad middle, \quad long >$
$s_4 =< middle, \quad 0, \quad middle, \quad 0, \quad 0, \quad 0, \quad middle, \quad middle >$
$s_5 =< short, \quad short \quad 0, \quad long, \quad 0, \quad 0, \quad 0, \quad 0 >$
$s_6 =< middle, \quad middle, \quad 0, \quad 0, \quad 0, \quad 0, \quad middle, \quad long, \quad >$

Assume the six web access patterns are grouped into 3 clusters. Each cluster is characterized as a rough variable $\zeta_k (1 \le k \le 3)$. ω_{up}=0.3, ω_{low}=0.7, δ=0.1.

Firstly we randomly choose s_1, s_3 and s_5 as the centroids of the three clusters, thus we can get follow equations.

$m_1 = s_1$
$= \langle \lambda_{11}, \lambda_{12}, ... \lambda_{18} \rangle$
$\Box \langle E[short], E[short], 0, E[long], E[middle], 0, 0, 0 \rangle.$
Similarly,

$m_2 = s_3$
$\Box \langle E[middle], E[middle], 0, E[short], 0, 0, E[middle], E[long] \rangle.$

$m_2 = s_5$
$\Box \langle E[short], E[short], 0, E[long], 0, 0, 0, 0 \rangle.$

Here, $E[short], E[middle], E[long]$ can be gained by Eq.(3).

For each pattern $s_l (1 \le l \le 6)$, compute $d(s_l, m_i)(1 \le k \le 6, 1 \le i \le 3)$.

If $d(s_l, m_i) - d(s_l, m_j) \le \delta$ then $s_l \in \zeta_i, s_l \in \zeta_j$, and s_l can not be a member of any lower approximation of rough variables. Else if $d(s_l, m_i)$ is minimum over the 3 clusters, we can get $s_l \in \underline{\zeta_i}$.

After one cycle, we can get the following results shown in Figure 3.

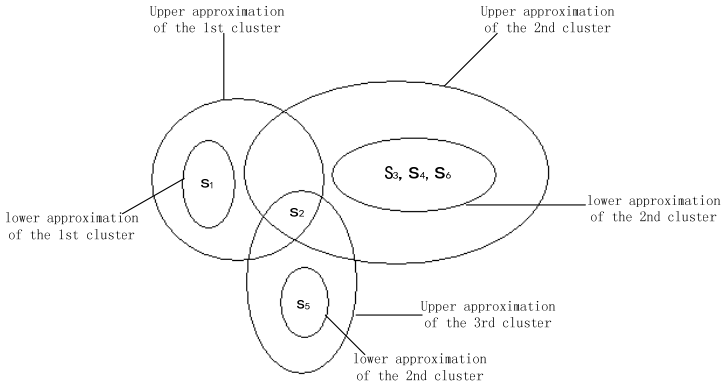

Figure 3. The first clustering result

After the algorithm stops, the clustering result is shown in Figure 4. s_1 and s_5 are clustered into one group. Their surfing behaviors are similar. Similarly, s_3, s_4 and s_6 are partitioned into one cluster. s_2 has little similarity with the patterns in other clusters.

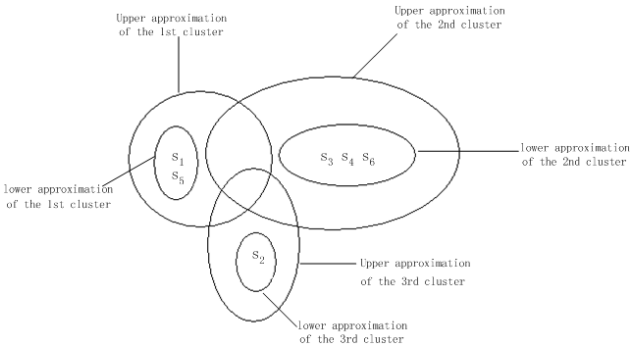

Figure 4. The last clustering result

From Figure 4, we can see that the three clusters have no overlaps. Their boundaries are crisp.

5. An experiment

We download a log file from a web server. The information resource is shown in Table 2.

Table 2. Information of experimental source

Web URLs	20
Expreimental period	2006/5/10 1-2006/5/17
Records	5,650
Records after data cleaning	1,020

Assume these patterns are divided into 3 clusters. Membership functions of time duration on web pages are shown in Figure 2. ω_{up}=0.3, ω_{low}=0.7, δ=0.1. After the convergence, the result is shown in Fig.5.

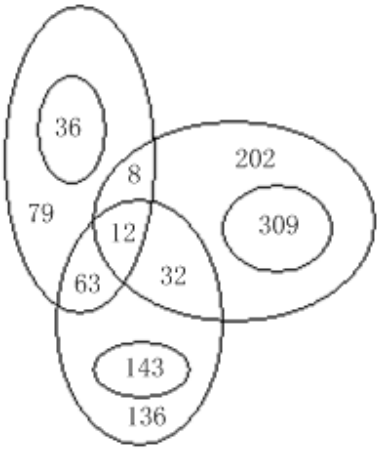

Figure 5. Clustering result

From Figure 5, there exist overlaps between clusters.

6. Conclusion

Soft computing techniques, such as fuzzy theory and rough set theory, are suitable for handling the issues related to understandability of patterns, incomplete/noisey data, and can provide approximate solutions faster.

In this paper, a novel approach based on rough k-means in fuzzy environments is proposed to cluster the web transactions. This approach is useful to find interesting user access patterns in web log. These user access patterns will aid the web site designer and be helpful in building up adaptive web server according to the user individual behavior.

Acknowledgement

This work was supported by National Natural Science Foundation of China Grant No.70802043 and Shanxi Provincial Natural Science Foundation of China Grant No. 2008011029-2.

References

[1] S. De, P. Krishna, Clustering web transactions using rough approximation, Fuzzy Sets and Systems, 148(2004) 131-138.
[2] R. Krishnapram, A. Joshi, and etc, low compexity fuzzy relational clustering algorithms for web mining, IEEE Transactions on Fuzzy Systems, 9(2001), 595-607.
[3] R. Hathaway, J. Beadek, Switching regression models and fuzzy clustering, IEEE Transactions on Fuzzy Systems, 1(3)(1993), 195-204.
[4] P. Lingras, Rough set clustering for web mining, Proceedings of the 2002 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'02), Vol: 2, pp.1039-1044.
[5] P. Lingras, C. West, Interval set clustering of web users with rough k-means, Journal of Intelligent Information Systems 23(1)(2004) 5-16.
[6] B. Liu, Fuzzy random dependent-chance programming, IEEE Transactions on Fuzzy Systems, 9(5)(2001) 721-726.
[7] B. Liu, Theory and Practice of Uncertain Programming, Physica-Verlag, Heidelberg, 2002.
[8] B. Liu, Y. Liu, Expected value of fuzzy variable and fuzzy expected value models. IEEE Transactions on Fuzzy Systems. 10(2002) 445-450.
[9] S. Mitra, An evolutionary rough partitive clustering, Pattern Recognition Letters, 25(2004) 1439-1449.
[10] S. Nahmias, Fuzzy variable. Fuzzy Sets and Systems. 1(1978) 97-101.
[11] Z. Pawlak, Rough sets, International Journal of Comput. Inform. Sci. 11(1982) 341-356.
[12] Z. Pawlak, Rough sets-Theoretical aspects of reasoning about data, Kluwar Academic Pulishers, Dorfrecht, 1991.
[13] T. Runkler, J. Beadek, Web mining with relational clustering, International Journal of Approximate Reasoning, 32(2003) 217-236.
[14] X. Wang, M. Ha, Note on maxmin u/E estimation, Fuzzy Sets and Systems, 94(1998) 71-75.
[15] L. Zadeh, Fuzzy sets. Information and Control, 8(1965) 338-353.
[16] T. Hong, C .Kuo, S. Wang, A Fuzzy AprioriTid Mining Algorithm With Reduced Computational Time. Applied Soft Computing. 5 (2004) 1-10.
[17] D. Arotaritei, S, Mitra. Web Mining: A Survey In The FuzzyFramework. Fuzzy Sets and Systems. 148 (2004) 5-19.