文章编号:1000-6788(2007)07-0116-06

基于模糊粗糙 k-均值的用户访问模式的聚类

吴　瑞[1],宁玉富[2]
(1. 山西师范大学数学与计算机学院,临汾 041004; 2. 德州学院计算机系,德州 253023)

摘要:　Web 用户访问过的网页以及在该网页上的浏览时间体现了用户的访问兴趣.为了更好的衡量任意两个用户访问模式之间的相似/相异度,每个用户访问模式都被转换成具有相等长度的模糊向量,其中每个元素要么是 0 要么是模糊语言变量,它体现了用户是否访问过该网页及在该网页上的浏览时间.由于类的边界可能是模糊的,因而使用粗糙 k-均值法对这些代表用户浏览特征的模糊向量进行聚类.最后使用 Davies-Bouldin 指标来衡量聚类的效果.

关键词:　Web 挖掘;Web 聚类;用户浏览模式;粗糙 k-均值
中图分类号:　TP311.13　　**文献标志码**:　A

Clustering User Access Patterns based on Fuzzy Rough k-Means

WU Rui[1], NING Yu-fu[2]
(1. Colledge of Mathematics and Computer Science, Shanxi Normal University, Linfen 041004, China; 2. Department of Computer Science, Dezhou University, Dezhou 253023, China)

Abstract:　The interest of web users can be revealed by their visited web pages and time durations on these web pages during their susfing. In order that similarity/difference between any two patterns can be easily gained, each web access pattern from web logs is transformed as fuzzy vector with same length, in which each element is a fuzzy linguistic variable or 0 representing the visited web page and time duration on this web page. The clusters may not exist crisp boundaries, thus a rough k-means clustering algorithm is proposed to group the fuzzy vectors denoting users' surfing behaviors. Finally, Davies-Bouldin index is provided to measure the clustering exactness.

Key words:　web mining; web clustering; user access patterns; rough k-means

1　引言

WWW 给网站设计者和运营者带来巨大商机的同时,也带来了巨大的挑战.如果一个特定的网站不能在很短的时间内满足用户的需要,这些用户会很快转向其它网站.因此了解网络用户的浏览行为、特征是非常必要的[1].

Web 挖掘的三个主要操作是:web 聚类分析,web 关联分析和 web 序列分析.Web 聚类就是对 web 用户或者对 web 上的信息,如用户访问模式,网页信息等进行聚类.然而 web 聚类不同于传统的聚类算法,其一,web 数据大多是非数值型数据,因而 Runkler 和 Beadek[2] 提出了使用关系聚类的方法对 web 数据进行聚类.其二,错误的和不完整的数据出现的可能性非常高,类与类之间可能存在模糊或不准确的边界[1].一个模式可能以不同程度属于多个类.因此在聚类过程中,人们往往使用模糊或粗糙或两者结合的方法来处理这些不确定的信息[3~6].Krishbapruam[3] 使用模糊关系聚类的方法对 web 用户进行聚类,Arotaritei 等人[4] 认为 web 挖掘中的聚类就是建立一个未知数目的有重叠的集合的模型过程,他们提出了使用模糊的方法对 web 数据进行聚类.业宁等人[5] 提出了一种分析 web 用户行为的聚类算法(FCC),同时给出了一种路径相似度系数计算方法,并使之与雅可比相似系数结合,用于计算用户访问行为的相似度.De[6] 等人使用粗糙集的上近似和下近似表示一个类,然后使用粗糙集的近似理论对 web 访问模式进行聚类.

本文提出了使用粗糙 k-均值的方法在模糊环境下对 web 日志中提取出的用户访问模式进行聚类.一

收稿日期:2006-05-17
资助项目:山西省自然科学基金(2006011039)
作者简介:吴瑞(1971 -),女,博士,副教授,研究方向:不确定理论与 Web 挖掘.

个用户访问模式代表了一个用户一次特定的浏览行为,它可表示成如下形式:$s_i = \{ url_{i1}(t_{i1}) \rightarrow url_{i2}(t_{i2}) \rightarrow \cdots \rightarrow url_{il}(t_{il}) \}$,其中 $url_{ik}(1 \leq k \leq l)(1 \leq i \leq n)$ 表示第 i 个用户访问的第 k 个网页,t_{ik} 表示访问 url_{ik} 的时间,l 表示一次浏览过程所访问过的所有页面的个数,n 是从 web 日志中提取的用户访问模式的总数. 如果一个网页频繁出现在若干个用户访问模式中,我们认为人们对该网页具有共同的兴趣,然而,不同的用户在该网页上停留不同的时间说明人们对该网页的兴趣程度不同. 基于上述考虑,网页以及网页上的浏览时间是反映用户兴趣的两个重要因素. 由于任意两个网页上的浏览时间的细微差别可以忽略,而且希望表达成人们容易理解的方式,网页上的浏览时间被刻画成相应的模糊语言变量. 每个用户访问模式都被转换成相等长度的模糊向量形式,向量中的每个元素要么是 0 要么是表示浏览时间的模糊语言变量. 最后使用粗糙 k-均值的方法对这些表示用户浏览行为的模糊向量进行聚类,每一个所产生的类代表了一组具有相似浏览习惯的用户访问模式集.

2　预备知识

定义 1　假设 ξ 是从可能性空间 $(\Theta, \mathrm{P}(\Theta), \mathrm{Pos})$ 到实直线 R 的函数,则称 ξ 是一个模糊变量.

定义 2[7]　模糊变量 ξ 的期望值可被定义成如下形式

$$E[\xi] = \int_0^{\infty} \mathrm{Cr}\{\xi \geq r\}\, dr - \int_{-\infty}^{0} \mathrm{Cr}(\xi \leq r)\, dr, \tag{1}$$

其中这两个积分中至少有一个是有限的.

定义 3[8]　一个粗糙变量 η 是一个从粗糙空间 $(\Lambda, \Delta, \mathrm{A}, \pi)$ 到实数集的可测函数. 也就是说,对 R 的任意 Borel 集 B,我们有

$$\{\lambda \in \Lambda \mid \eta(\lambda) \in B\} \in \mathrm{A}. \tag{2}$$

粗糙变量 η 的下近似和上近似可被定义成如下形式:

$$\underline{\eta} = \{\eta(\lambda) \mid \lambda \in \Delta\}, \tag{3}$$

$$\overline{\eta} = \{\eta(\lambda) \mid \lambda \in \Lambda\}. \tag{4}$$

引理 1[8]　由于 $\Delta \subset \Lambda$,显然 $\underline{\eta} \subset \overline{\eta}$.

上近似 $\overline{\eta}$ 中的元素可能属于也可能不属于粗糙变量 η,下近似 $\underline{\eta}$ 中的元素一定属于粗糙变量 η.

3　在模糊环境下使用粗糙 k-均值对用户访问模式的聚类

Web 服务器访问日志文件中包含有大量用户的浏览信息,这些信息揭示了用户的浏览行为. Web 日志原始数据需要进行数据清洗、用户识别、会话识别和事务识别等预处理. 本文只作简单的数据清洗和会话识别. 假定 $W = \{url_1, url_2, \cdots, url_m\}$ 是用户访问过所有页面的全集. 假定第 i 个用户浏览模式可表示成如下形式:$s_i = \{ url_{i1}(t_{i1}) \rightarrow url_{i2}(t_{i2}) \rightarrow \cdots \rightarrow url_{ip}(t_{ip}) \}$,其中 $url_{ik} \in W$,p 为第 i 个用户浏览的页面总数. 假定第 j 个用户的浏览模式为:$s_j = \{ url_{j1}(t_{j1}) \rightarrow url_{j2}(t_{j2}) \rightarrow \cdots \rightarrow url_{jq}(t_{jq}) \}$,其中 $url_{jd} \in W(1 \leq d \leq q)$,$q$ 为第 j 个用户浏览的页面总数. 如果 $url_{ia} \in s_i(1 \leq a \leq p)$,$url_{ib} \notin s_i$,则 i 用户对 url_{ia} 网页感兴趣,而对 url_{ib} 不感兴趣. 如果 $(url_{ia}, t_{ia}) \in s_i(1 \leq a \leq p)$,$(url_{jc}, t_{jc}) \in s_j(1 \leq c \leq q)$,并且 $url_{ia} = url_{jc} = url_k \in W(1 \leq k \leq m)$,$t_{ia} \neq t_{jc}$ 意味着他们对网页 url_k 具有不同的兴趣程度. 如果 $t_{ia} > t_{jc}$,则 i 用户对该网页的兴趣要大于 j 用户的. 然而,强调时间与时间之间的细小差别在现实中是毫无意义的. 而且人们往往习惯用语言术语来刻画事物的特征,如身高为"高/矮",时间用"长/短"等. 本文把网页上的浏览时间刻画成模糊语言变量,它符合人们的正常思维方式,而且可以忽略时间之间的细微差别. 然而为了容易得到任意两个用户浏览模式之间的相似程度,每个用户访问模式需要被转换成等长的模糊向量形式.

3.1　刻画用户访问模式为模糊向量

假定存在 n 个用户浏览模式,记做 $S = \{s_1, s_2, \cdots, s_n\}$,这里 s_i 代表了第 i 个用户的特定的浏览行为. $W = \{url_1, url_2, \cdots, url_m\}$ 是访问过的页面的全集,令 $U = \{(url_1, t_{11}), \cdots, (url_1, t_{1g}), \cdots, (url_k, t_{k1}), \cdots, (url_k, t_{kh})\}$ 是所有用户浏览页面及页面上的浏览时间的全集,g 是浏览 url_1 上的用户的个数,h 是浏览页

面 url_k 的用户个数.

显然 $s_i \in S(1 \leq i \leq n)$,且 $s_i \subset U$. 这里暂不考虑网页的访问顺序. 则 s_i 可被表示成如下向量形式:

$$V_i = \langle v_{i1}^t, v_{i2}^t, \cdots, v_{im}^t \rangle, \tag{5}$$

其中 $v_{ik}^t = \begin{cases} t_{ik}, (url_k, t_{ik}) \in s_i \\ 0, \text{否则} \end{cases}, (1 \leq k \leq m)$. 这样每个模式均可被转换成具有长度 m 的实向量.

网页上的浏览时间首先被聚成 r 个模糊区间,每个模糊区域对应一个模糊语言变量,每个模糊语言变量的隶属函数可以根据模拟的方法得到,也可根据专家系统给出. 假定隶属函数定义见图 1.

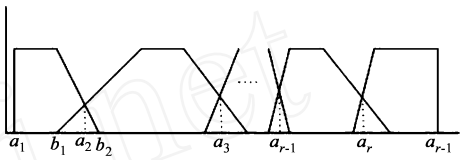

图 1 网页浏览时间的隶属函数

第一个模式区域被刻画成一个梯形模糊变量 $\xi_1(a_1, a_1, b_1, b_2)$,最后一个模糊区域被刻画成梯形模糊变量 ξ_r. 根据图 1,我们得到 v_{ik}^t 与模糊语言变量 $\lambda_{ik}(1 \leq i \leq n)(1 \leq k \leq m)$ 之间的关系.

$$\lambda_{ik} = \begin{cases} 0, & v_{ik}^t = 0 \\ \xi_1, & a_1 \leq v_{ik}^t \leq a_2 \\ \xi_2, & a_2 \leq v_{ik}^t \leq a_3 \\ \cdots & \\ \xi_r, & a_r < v_{ik}^t \leq a_{r+1} \end{cases}, \tag{6}$$

其中 $\xi_j(1 \leq j \leq r)$ 是模糊语言变量.

根据公式(5)和(6),每个用户浏览模式可被转换成如下形式:

$$f_{vi} = \langle \lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{im} \rangle, \tag{7}$$

其中 $\lambda_{ik} \in \{0, \xi_1, \xi_2, \cdots, \xi_r\}$.

3.2 基于粗糙 k-均值的模糊向量的聚类

假定用户事务 S 存在 n 个用户浏览模式,$S = \{s_1, s_2, \cdots, s_n\}$,对于任意两个模式 s_i 和 s_j,它们可转换成如下模糊向量的形式:

$$f_{vi} = \langle \lambda_{i1}, \lambda_{i2}, \cdots \lambda_{im} \rangle,$$
$$f_{vj} = \langle \lambda_{j1}, \lambda_{j2}, \cdots \lambda_{jm} \rangle,$$

其中 $\lambda_{ik} \in \{0, \xi_1, \xi_2, \cdots, \xi_r\}$, $\lambda_{jk} \in \{0, \xi_1, \xi_2, \cdots, \xi_r\}(1 \leq k \leq m)$.

这两个模糊向量的和定义为:

$$sum(f_{vi}, f_{vj}) = \langle E[\lambda_{i1} + \lambda_{j1}], E[\lambda_{i2} + \lambda_{j2}], \cdots, E[\lambda_{im} + \lambda_{jm}] \rangle, \tag{8}$$

这里 $E[\lambda_{ik} + \lambda_{jk}](1 \leq k \leq m)$ 可通过模糊模拟算法(参见文献[7])求出.

模式 s_i 和 s_j 的相异性(距离)定义为:

$$d(s_i, s_j) \cong d(f_{vi}, f_{vj}) = \sqrt{\frac{\sum_k^m (E[\lambda_{ik} - \lambda_{jk}])^2}{m}} \tag{9}$$

同理利用模糊模拟算法可求解出 $(E[\lambda_{ik} - \lambda_{jk}])^2 (1 \leq k \leq m)$.

本文采用粗糙 k-均值的方法对转换成模糊向量的用户访问模式进行聚类. 由于每个类的边界可是模糊的,每个类定义成一个可测粗糙空间上的粗糙变量 $\eta_i(1 \leq i \leq k)$. 那么每个类的中心点的定义如下:

$$m_i = \begin{cases} w_{low} \dfrac{\sum_{f_{vj} \in \underline{\eta_i}} f_{vj}}{|\underline{\eta_i}|} + w_{up} \dfrac{\sum_{f_{vj} \in (\overline{\eta_i} - \underline{\eta_i})} f_{vj}}{|\overline{\eta_i} - \underline{\eta_i}|} \\ w_{low} \dfrac{\sum_{f_{vj} \in \underline{\eta_i}} f_{vj}}{|\underline{\eta_i}|} \end{cases}. \tag{10}$$

其中参数 w_{low}/w_{up} 决定中心点时的上近似和下近似权重. $0.5 < w_{low} < 1$, $w_{up} = 1 - w_{low}$. $|\underline{\eta_i}|$ 表示处于第 i 类中下近似的模式的个数, $|\overline{\eta_i} - \underline{\eta_i}|$ 表示处在下近似和上近似之间的模式的个数. 中心点 m_i 是一个实值型的向量 $m_i = \langle c_{i1}, c_{i2}, \cdots, c_{im} \rangle$.

模式 s_i $(1 \leq i \leq n)$ 到中心点 m_j $(1 \leq j \leq k)$ 的距离定义如下:

$$d(s_i, m_j) = d(f_{vi}, m_j) = \sqrt{\frac{\sum_{l=1}^{m} (E[\lambda_{il}] - c_{jl})^2}{m}} \tag{11}$$

如果 $d(s_i, m_{k1})$ 和 $d(s_i, m_{k2})$ $(1 \leq i \leq n)$ 的差距很小, 且小于一个给定的阈值, 则模式 s_i 不能清晰的划入一个类中, 它可能属于第 $k1$ 个类和第 $k2$ 个类的上近似. 否则, 如果 $d(s_i, m_{k1})$ 是所有 k 个类中最小的, 则模式 s_i 一定属于第 $k1$ 个类的下近似, 它归属于第 $k1$ 个类是确定的. 显然, 类与类之间存在着交叉. 一个模式最多只能属于一个类的下近似, 但它可属于多个类的上近似.

假定要把模式 $S = \{s_1, s_2, \cdots, s_n\}$ 聚成 k 个类, 分类框架 $C = \{C_1, C_2, \cdots, C_k\}$, 其中每个类可刻画成一个粗糙变量 η_i $(1 \leq i \leq k)$, 可由它的下近似和上近似 $(\underline{\eta_i}, \overline{\eta_i})$ 来表示. 算法描述如下.

算法 1　用户访问模式的聚类算法

输入: 用户事务 $S = \{s_1, s_2, \cdots, s_n\}$, 浏览时间的隶属函数, 阈值 $\delta \in [0, 1]$, w_{low}/w_{up}

输出: 分类框架 C

1) 开始.

2) For each $s_j \in S$, 根据 3.1 节把 s_j 转换成相应的模糊向量 f_{vj}.

3) 对 k 个类初始化中心点 m_i, 随机选择 k 个模式作为中心点.

4) For each f_{vj} $(1 \leq j \leq n)$ do

For $i = 1$ to k do

根据公式 (11) compute $d(f_{vj}, m_i)$

5) For each f_{vl} $(1 \leq l \leq n)$ do

If $d(f_{vl}, m_i) - d(f_{vl}, m_j)$ $(i \neq j) \leq \delta$, then $s_l \in \overline{\eta_i}$, $s_l \in \overline{\eta_j}$;

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

6) $C_1 = \{\underline{\eta_1}, \overline{\eta_1}\}, \cdots, C_k = \{\underline{\eta_k}, \overline{\eta_k}\}$

7) 根据公式 (10), 重新计算每个类的中心点 m_i.

8) 重复 (4) ～ (7) 直到收敛.

9) 输出 C.

10) 算法停止.

转换 n 个用户访问模式为等长模糊向量的时间为 $O(n)$, 而计算每个模糊向量到聚类中心的时间为 $O(kn)$, 重新计算每个类的中心需要 $O(kn')$ $(n' < n)$. 如果在聚类中心选择适当的情形下, 收敛很快, 时间复杂度为 $O(n)$, 相反收敛慢, 时间复杂度接近 $O(n^2)$. 为了便于计算任意两模式之间的相似/相异性, 每个模式被转换成等长形式, 因此需要 mn 大小的存储空间. 为了避免过度浪费不必要的空间, 可以过滤去过长或过短的用户模式. 算法的准确性检验见实验比较.

4　实例分析

本节将通过一个例子对该算法进行描述. 假定经过预处理后的用户访问模式见表 1.

假定根据专家系统给定的网页上的浏览时间的隶属函数见图 2.

由图 2, 网页上的浏览时间被刻画成三个梯形模糊语言变量, *short* (5, 5, 30, 60)、*middle* (30, 60, 90, 120) 和 *long* (90, 120, 150, 150). 则实值型的浏览时间 v_{ik}^t 与模糊语言变量 λ_{ik} $(1 \leq i \leq n)$ $(1 \leq k \leq m)$ 之间的关系如下.

$$\lambda_{ik}=\begin{cases}0, & v_{ik}^{t}=0\\ short, & 5\leq v_{ik}^{t}\leq 45\\ middle, & 45<v_{ik}^{t}\leq 105\\ long, & 105<v_{ik}^{t}\leq 150\end{cases}. \tag{12}$$

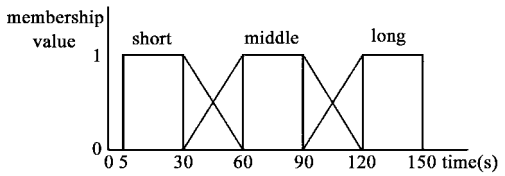

图 2 浏览时间的隶属函数

表 1 用户浏览模式集

cid	用户浏览序列
1	(A ,30) , (B ,42) , (D ,118) , (E ,91)
2	(A ,92) , (B ,89) , (F ,120)
4	(A ,50) , (B ,61) , (D ,42) , (G,98) , (H ,115)
5	(A ,70) , (C ,92) , (G,85) , (H ,102)
6	(A ,40) , (B ,35) , (D ,112)
7	(A ,52) , (B ,89) , (G,92) , (H ,108)

令 $S=\{s_1,s_2,s_3,s_4,s_5,s_6\}$,$U=\bigcup_{i=1}^{6}s_i$,$W=\{A,B,C,D,E,F,G,H\}$. 则每个模式 $s_i\in S(1\leq i\leq 6)$ 根据 3.1 节可表示成如下模糊向量的形式.

$$s_1=\langle short,short,0,long,middle,0,0,0\rangle$$
$$s_2=\langle middle,middle,0,0,0,long,0,0\rangle$$
$$s_3=\langle middle,middle,0,short,0,0,middle,long\rangle$$
$$s_4=\langle middle,0,middie,0,0,0,middle,middle\rangle$$
$$s_5=\langle short,short,0,long,0,0,0,0\rangle$$
$$s_6=\langle middle,middle,0,0,0,0,middle,long\rangle$$

假定要把这六个用户访问模式聚成 3 类,每个类是一个粗糙空间上的可测的粗糙变量 $\eta_i(1\leq i\leq 3)$,且有 $w_{up}=0.3$,$w_{low}=0.7$,$\delta=0.1$.

首先选择 s_1、s_3 和 s_5 作为初始的 3 个类的中心点,$m_1=s_1$,$m_2=s_3$,$m_3=s_5$. 计算 $d(f_{vl},m_i)(1\leq l\leq 6)(1\leq i\leq 3)$,如果 $d(f_{vl},m_i)-d(f_{vl},m_j)(i\neq j)\leq\delta$,那么 $s_l\in\overline{\eta_i}$,$s_l\in\overline{\eta_j}$;否则如果 $d(f_{vl},m_i)$ 是 3 个类中距离最小的,则 $s_l\in\underline{\eta_i}$.

一次循环后,聚类结果为 $s_1\in\underline{\eta_1}$,$s_2\in\overline{\eta_1}$,$s_2\in\overline{\eta_2}$,$s_2\in\overline{\eta_3}$,s_3,s_4,$s_6\in\underline{\eta_2}$,$s_5\in\underline{\eta_3}$. 重复执行直到收敛,$s_1\in\underline{\eta_1}$,$s_5\in\underline{\eta_1}$;s_3,s_4,$s_6\in\underline{\eta_2}$,$s_2\in\underline{\eta_3}$,则最后聚类结果为 $C_1=\{s_1,s_5\}$,$C_2=\{s_3,s_4,s_6\}$,$C_3=\{s_2\}$.

5 实验分析

我们从 *web* 服务器上下载一日志文件,页面总数为 20,经过数据清洗和简单的用户浏览行为的识别后,用户的浏览路径数为 1,020. 假定要把这些用户浏览模式聚成 3 类,每个类是一个粗糙空间上的可测的粗糙变量 $\eta_i(1\leq i\leq 3)$,且有 $w_{up}=0.3$,$w_{low}=0.7$,$\delta=0.1$. 则有聚类结果见图 3.

从图 3 可看出如下结论:类与类之间是有重叠的;其中 36 种模式一定属于第 1 类;有 79 种模式接近于第 1 类;有 8 种模式可能属于第 1 类,也可能属于第 2 类,它们的归类是模糊的;其中 12 种模式的模糊性更大,它们可能是这 3 类中的任何一种;更多的用户更倾向于第 2 类的浏览模式,该聚类方法显然比传统的聚类方法(它们的类间界限是清晰的)更自然更易被人理解.

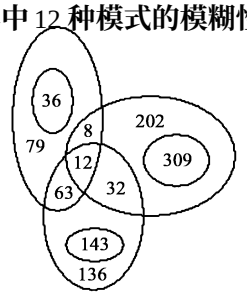

图 3 聚类结果图

这里使用 Davies-Bouldin 指标[9] 来衡量聚类的效果,其定义如下:

$$\frac{1}{c}\sum_{k=1}^{c}\max_{l\neq k}\left\{\frac{S(U_k)+S(U_l)}{d(U_k,U_l)}\right\} \tag{15}$$

其中 $S(U_k)$ 表示第 k 个类的类内距离,$d(U_k,U_l)$ 表示类间距离.

对以上相同数据再使用模糊 k-均值方法和基于粗糙近似法进行聚类后根据公

式(15)计算其 Davies-Bouldin 值,比较结果见表 2.

表 2　比较结果

聚类方法	Davies-Bouldin 值
模糊粗糙 k-均值聚类	0.476
模糊 k-均值聚类	0.583
基于粗糙近似聚类	0.692

6　结论

Web 用户访问模式的聚类由于其海量的数据以及很多不确定因素的存在使之成为一个极具挑战的工作.本文提出了模糊环境下基于粗糙 k-均值的方法对用户的访问模式进行聚类.由于很多不确定因素的存在,如类与类之间可能不存在清晰的边界,粗糙变量被用来代表一个边界模糊的类,这样类与类之间存在交叉是符合现实规律的.这种方法有助于从 web 日志中发现一些有趣的规律,从而根据用户的浏览行为建立个性化的网站.

参考文献:

[1] Lingras P, West C. Interval set clustering of web users with rough k-means[J]. Journal of Intelligent Information Systems, 2004, 23(1): 5 - 16.

[2] Runkler T, Beadek J. Web mining with relational clustering[J]. International Journal of Approximate Reasoning, 2003, 32: 217 - 236.

[3] Krishnapram R, Joshi A. Low compexity fuzzy relational clustering algorithms for web mining[J]. IEEE Transactions on Fuzzy Systems, 2001, 9: 595 - 607.

[4] Arotaritei D, Mitra S. Web mining: a survey in the fuzzy framework[J]. Fuzzy Sets and Systems, 2004, 148: 5 - 19.

[5] 业宁,李威,等. 一种 Web 用户行为聚类算法[J]. 小型微型计算机系统, 2004, 25(7): 1364 - 1367.
Ye Ning, Li Wei, et al. An clustering algorithm for web user behaviors[J]. MINI-MICRO Systems, 2004, 25(7): 1364 - 1367.

[6] De S, Krishna P. Clustering web transactions using rough approximation[J]. Fuzzy Sets and Systems, 2004, 148: 131 - 138.

[7] Liu B, Liu Y. Expected value of fuzzy variable and fuzzy expected value models[J]. IEEE Transactions on Fuzzy Systems, 2002, 10(4): 445 - 450.

[8] Liu B. Theory and Practice of Uncertain Programming[M]. Heidelberg: Physica-Verlag, 2002, 102 - 108.

[9] Bezdek J, Pal N. Some new indexes forcluster validity[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part-B, 1998, 28: 301 - 315.