一种基于特征属性的Web用户模糊聚类改进算法

应玉龙
(浙江纺织服装学院 信息工程分院,浙江 宁波 315211)

摘 要:为降低传统FCM算法的计算复杂性,提高Web用户聚类的效果,文中提出了一种改进的基于特征属性的Web用户模糊聚类算法。首先通过用户访问页面的次数和时间建立Web用户兴趣度矩阵,并根据商品的特征属性值将Web用户兴趣度矩阵映射为用户对特征属性的偏好矩阵,从而有效降低数据稀疏性;然后以此为数据集,对传统的FCM算法进行了改进,将聚类中心分为活动和稳定两种,忽略稳定聚类中的距离计算以降低计算复杂性。最后通过仿真实验证实了新算法的有效性和可行性。

关键词:特征属性;Web用户;模糊聚类;模糊C均值算法
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2012)07-0095-04

An Improved Web Users Fuzzy Clustering Algorithm Based on Features Property

YING Yu-long
(Information Engineering School, Zhejiang Textile & Fashion College, Ningbo 315211, China)

Abstract: In this paper, present an improved Web users fuzzy clustering algorithm based on features property to reduce the computational complexity of conventional fuzzy c-means clusering algorithm and improve the effect of Web users clusering. First, establish Web user's interest degree matrix through the times and time of user's visited pages, mapping the Web user's interest degree matrix into the user's features property preference matrix according to the features property of item to reduce the data sparseness effectively. Based on the features property preference matrix, improved the conventional fuzzy c-means clusering algorithm. The proposed method first classifies cluster centers into active and stable groups, then skips the distance calculations for stable clusters in the iterative process to reduce the computational complexity of conventional fuzzy c-means clustering algorithm. Finally, the simulation demonstrates the feasibility and validity of the proposed method.

Key words: features property; Web users; fuzzy clustering; FCM

0 引 言

Web用户的浏览行为直接反映了用户的兴趣爱好,通过对Web日志信息的挖掘,发现具有相似访问兴趣的用户群组,可以实现用户的分类管理,提高电子商务个性化服务水平[1]。使用模糊聚类技术,根据用户的浏览行为进行相似用户聚类,是当前Web日志挖掘的研究热点之一[2]。传统的Web用户聚类算法主要根据用户访问页面的次数[3]、时间[4]和路径序列[4,5]等浏览行为,生成用户页面兴趣度矩阵并进行用户相似性度量,从而实现用户聚类。然而,很多时候单纯通过用户对页面的兴趣度来判断用户间是否存在共同兴趣具有一定的片面性[6]。根据商品的特征属性值将用户对页面的兴趣度矩阵映射为用户对商品特征属性的偏好矩阵并进行聚类分析,可以有效降低数据稀疏性,提高聚类的效果。

在Web用户聚类中,模糊C均值算法(Fuzzy c-means, FCM)得到了广泛的应用和研究[7]。FCM算法是一种划分算法,它通过聚类中心的不断迭代直到收敛来实现最优划分[8]。由于传统FCM算法的收敛过程慢于HCM[9],为降低FCM算法的计算复杂性,文中对传统的FCM算法进行改进,利用迭代过程中部分聚类中心趋于稳定的特点,将聚类中心分为活动和稳定两类,迭代过程中忽略稳定聚类中的距离计算,从而有效降低FCM算法的计算复杂性。

收稿日期:2011-11-30;**修回日期**:2012-03-02
基金项目:宁波市自然科学基金(2010A610118);宁波市先进纺织技术与服装CAD重点实验室(2011ZDSYS-A-004)
作者简介:应玉龙(1979-),男,浙江浦江人,硕士研究生,研究方向为数字挖掘、信息推荐、图像检索。

1 用户特征属性偏好矩阵

1.1 Web用户兴趣度描述

通过对Web日志进行预处理,可以统计得到用户浏览行为的原始信息,包括用户浏览网页的URL地址、浏览时间和浏览次数等。而具有相似浏览模式的用户也具有相似的兴趣度[10]。文中主要使用浏览次数和浏览时间两个指标来定义用户的兴趣度,兴趣度反映了用户对某一网页的偏爱程度,其相关定义如下:

定义1 用户访问事务向量T[11]:用户在一段连续时间内访问的页面序列,可以用一个三元组$T=(U,P,T)$表示,其中$U=\{U_1,U_2,\cdots,U_m\}$(m为用户的个数)表示访问Web页面的用户集合;$P=\{P_1,P_2,\cdots,P_n\}$(n是网页的个数)表示用户访问的页面集合;$T_i=\{T_1,T_2,\cdots,T_n\}$($n$是网页的个数)表示用户访问每个页面的时间集合。

定义2 Web页面兴趣度[12]:设W是网站中所有商品页面的集合,$W_j\in W(0<j\leqslant n,\ n$为商品页面个数),$U$是访问Web站点的所有用户集合,$u_i\in U(0<i\leqslant m,\ m$为用户总数)。用户$u_i$对商品页面$W_j$的兴趣度$I_{ij}$为:

$$I_{ij}=\frac{\sum_{t=1}^{\omega}T_{ijt}}{\sum_{j=1}^{n}\sum_{t=1}^{\omega}T_{ijt}}$$

其中I_{ij}表示用户i对页面j的兴趣度,ω表示用户对该页的浏览次数,T_{ijt}表示用户i第t次访问页面j的停留时间(单位:秒)。

定义3 Web用户兴趣度矩阵:

$$\boldsymbol{M}_{n\times m}=\begin{bmatrix}h_{11}&\cdots&h_{1m}\\ \vdots&h_{ij}&\vdots\\ h_{n1}&\cdots&h_{nm}\end{bmatrix}$$

元素值h_{ij}为用户i对商品页面j的兴趣度,其中$h_{ij}=I_{ij}$,每一行向量$\boldsymbol{h}[i,.]$表示用户i对所有商品页面的访问情况,每一列向量$\boldsymbol{h}[.,j]$表示所有用户对页面j的访问情况。

1.2 特征属性偏好矩阵

Web用户兴趣度矩阵反映了用户对每个商品页面的兴趣度,而每个商品页面的URL地址一般包含了商品的编号,因此Web用户兴趣度矩阵也可以看成是用户对项目(商品)的兴趣度。但是用户浏览(或购买)某一商品的目的通常是出于对该商品所具有的某些特征属性值的偏好。同时,一个用户往往只访问他感兴趣的商品页面,这部分页面只占所有商品页面的很小部分,随着Web站点用户和商品数量的不断增加,导致兴趣度矩阵中的数据存在稀疏性,从而影响了用户聚类的效果。

因此,可以将用户对项目的兴趣度映射到相应的项目特征属性上,产生用户对特征属性的偏好矩阵,从而可以基于特征属性偏好进行用户相似性度量以缓解用户评分数据的稀疏性。

定义4 特征属性偏好矩阵:

$$\boldsymbol{P}_{n\times t}=\begin{bmatrix}p_{11}&\cdots&p_{1t}\\ \vdots&p_{ik}&\vdots\\ p_{n1}&\cdots&p_{nt}\end{bmatrix}$$

元素值p_{ik}表示用户i对特征属性值k的兴趣度,$p_{ik}=\sum_{l=1}^{H}h_{ij}$,其中$H$表示具有相同特征属性值的商品数目,$h_{ij}$表示特征属性值为$k$的用户$i$对商品页面$j$的兴趣度。$t$表示特征属性值个数,一般来说,$t\ll m$。

2 改进的Web用户模糊聚类算法

特征属性偏好矩阵反映了用户对商品某些特征属性值的偏好,是一个模糊矩阵,所以采用模糊聚类技术对特征属性偏好矩阵进行聚类分析,能够更加客观地产生用户分类效果。目前,FCM算法是应用最广泛的模糊聚类算法。

2.1 FCM算法描述

FCM聚类算法的描述如下:设数据集$X=\{x_1,x_2,\cdots,x_n\}$,将X分为c类($2\leqslant c\leqslant n$),用模糊矩阵$U=(u_{ij})$表示第$j(j=1,2,\cdots,n)$个数据点属于第$i(i=1,2,\cdots,c)$类的隶属度,目标函数定义如下:

$$F(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m\|x_j-v_i\| \tag{1}$$

式中$v_i(i=1,2,\cdots,c)$为第i个聚类中心,$V=\{v_1,v_2,\cdots,v_c\}$,$\|x_j-v_i\|$表示数据点x_j到聚类中心v_i的欧氏距离,m为加权指数。

FCM是将目标函数$F(\boldsymbol{U},\boldsymbol{V})$最小化的迭代收敛过程,从初始聚类中心开始,通过迭代不断修改聚类中心,直到相邻两次迭代所得的聚类中心变化很小,则认为算法已收敛停止迭代。

算法具体步骤如下:

Step1:给定类别数c,参数m,临界值ξ,ξ为大于0的很小的正数;

Step2:随机初始化聚类中心v_i^p,并令当前迭代次数$p=1$;

Step3:对于给定的聚类中心v_i^p,计算$d_{ij}=\|x_j-v_i\|$,并用式(2)重新计算隶属度u_{ij};

$$u_{ij}=\frac{1}{\sum_{l=1}^{c}\left(\frac{d_{ij}}{d_{lj}}\right)^{\frac{2}{m-1}}} \tag{2}$$

u_{ij}满足条件$\sum_{i=1}^{c}u_{ij}=1,j=1,2,\cdots,n$。

Step4:$p=p+1$,用式(3)修正所有的聚类中心 v_i^p;

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^m x_j}{\sum_{j=1}^{n}u_{ij}^m} \tag{3}$$

Step5:计算误差 $e=\sum_{i=1}^{c}\|v_i^p-v_i^{p-1}\|^2$,如果 $e<\xi$,算法结束;否则 $p=p+1$,转 Step3。

从以上流程中可以看出,FCM 算法的主要计算复杂性在步骤 3 和 4,而步骤 4 的计算复杂性又比步骤 3 小得多,因此,传统 FCM 算法的计算复杂性为 $O(nct)$,其中 t 为迭代次数。

2.2 改进的 FCM 算法

在 FCM 算法的迭代中,一部分聚类中心经过若干次迭代后就趋于稳定,其变化率小于临界值 ξ,而另一部分聚类中心需要更长时间的迭代才能趋于稳定。设第 i 个聚类中心本次和上次迭代过程中的聚类中心分别为 v_i^p 和 v_i^{p-1},$D_i=\|v_i^p-v_i^{p-1}\|$,如果 $D_i<\xi$,则称 v_i^p 为稳定聚类中心,否则为活动聚类中心。相应的,具有稳定聚类中心的聚类称为稳定聚类,否则为活动聚类。随着迭代次数的增加,稳定聚类中心的数量会随之增加[13]。

针对传统的 FCM 算法进行改进,将聚类中心划分为活动的和稳定的两个群组,在迭代过程中对于稳定聚类,跳过其距离计算步骤以减少计算复杂性,具体算法如下:

Step 1:给定类别数 c,参数 m,临界值 ξ;

Step 2:随机初始化聚类中心 v_i^p,并令当前迭代次数 $p=0$;

Step 3:对于给定的聚类中心 v_i^p,计算 $\|x_j-v_i\|$,并用式(2)重新计算隶属度 u_{ij};

Step 4:令 $p=p+1$,用式(3)计算每个聚类中心 v_i^p,并计算与 v_i^{p-1} 的欧式距离 $D_i=\|v_i^p-v_i^{p-1}\|$ $(i=1,2,\cdots,c)$,令 $q=1$,flag=false;

Step 5:如果 $D_q<\xi$,则转步骤 6;否则计算 $d_{qj}=\|x_j-v_q\|$ $(j=1,2,\cdots,n)$,令 flag=true(表示存在动态聚类中心);

Step 6:$q=q+1$,如果 $q>k$ 则转步骤 7,否则转步骤 5;

Step 7:如果 flag=true ,计算 u_{ij},转步骤 4;否则程序结束。

改进的 FCM 算法中只计算与活动聚类中心相关的距离,而活动聚类的数量随着迭代次数的增加逐渐减少,令 k_a 表示活动聚类平均数量,则 $1\leqslant k_a\leqslant c$,计算距离 d_{ij} 的概率为 k_a/c,因此,改进的 FCM 算法的计算复杂性为 $O(nct)\times O(k_a/c)=O(n\,k_a t)$,最坏情况下,当 $k_a=c$ 时,计算复杂性为 $O(nct)$。

3 实验结果及分析

本实验的数据取自某服装电子商务网站的服务器日志,对该日志数据进行预处理后取出 4300 条记录作为实验数据。其中,用户数为 56,商品总数为 325,如表 1 所示。

表 1　服务器日志数据集

UID	URL	访问时间(S)
1001	/product/item.htm? id=12519407300	12
1001	/product/item.htm? id=12519407305	3
1001	/product/item.htm? id=23215101709	21
1002	/product/item.htm? id=12359367185	18
1002	/product/item.htm? id=12359367185	10

对服务器日志数据按定义 2 中的公式计算用户对项目的兴趣度,如表 2 所示。

表 2　用户兴趣度表

UID	PID	兴趣度
1001	12519407300	0.3333
1001	12519407305	0.0834
1001	23215101709	0.5833
1002	12359367185	0.2341
1003	12519407305	0.4572
…	…	…

从表 2 可以看出,用户 1001 只浏览了 3 个商品页面,因此该用户对其他商品的兴趣度都为 0。根据商品编号 PID,按照服装风格属性进行归类,得出用户对服装风格的特征属性偏好矩阵如表 3 所示。

表 3　服装风格特征属性偏好矩阵表

UID	瑞丽	百搭	淑女	…	田园	简约
1001	0.4167	0.0000	0.5833	…	0.0000	0.0000
1002	0.6102	0.0128	0.1063	…	0.0104	0.1125
1003	0.2251	0.4571	0.3362	…	0.0000	0.0327
1004	0.0000	0.2289	0.3861	…	0.0328	0.3682
…	…	…	…	…	…	…
1056	0.0258	0.1420	0.3821		0.0388	0.0000

设定参数 $m=2$,临界值 $\xi=0.000001$,初始聚类数 c 分别为 2,4 和 6,从聚类效果和计算性能两个方面对算法进行比较。

3.1 聚类准确性分析

分别以用户兴趣度数据集和特征属性偏好数据集为样本,采用改进的 FCM 算法进行聚类分析,以欧式距离作为相似度衡量标准,计算各聚类簇内各样本间的平均相似度,以此作为聚类准确性评价标准,结果如表 4 和表 5 所示。

从表 5 可以看出,采用特征属性偏好数据集的聚

类结果中,各簇内样本的相似度普遍高于基于用户兴趣度数据集的聚类结果,说明基于特征属性的用户聚类方法具有较好的聚类效果。

表4 不同数据集各聚类簇样本数

聚类数	各聚类簇样本个数	
	用户兴趣度数据集	特征属性偏好数据集
$c=2$	21,35	27,29
$c=4$	8,13,9,16	9,12,15,20
$c=6$	5,9,12,13,8,9	5,8,13,11,9,10

表5 不同数据集各聚类簇样本间平均相似度

聚类数	各聚类簇内样本平均相似度	
	用户兴趣度数据集	特征属性偏好数据集
$c=2$	0.4562, 0.5318	0.6257, 0.6325
$c=4$	0.5276, 0.6083	0.6701, 0.6257
	0.5524, 0.5819	0.8002, 0.7105
$c=6$	0.7324, 0.6918	0.8429, 0.7802
	0.7539, 0.7225	0.8103, 0.7726
	0.6651, 0.8027	0.7539, 0.8122

3.2 计算性能比较

以特征属性偏好数据集作为聚类数据源,对传统FCM和改进后的FCM算法从距离计算次数方面进行比较,结果如表6所示。

表6 距离计算次数表

聚类数	传统FCM		改进FCM	
	迭代次数	距离计算量	迭代次数	距离计算量
$c=2$	23	2576	25	1960
$c=4$	31	6944	33	5360
$c=6$	46	15456	45	10836

从表6中可以看出,虽然改进FCM算法的迭代次数比传统FCM算法有了略微的增加,但由于距离计算次数的大量减少,使得改进FCM算法的计算性能总体上优于传统FCM算法。

4 结束语

文中提出了两种策略以提高Web用户聚类的效果。一是根据项目的特征属性将用户对项目的兴趣度映射为用户对特征属性的偏好,以此降低数据稀疏性,提高用户聚类的准确性;二是根据迭代过程中部分聚类中心趋于稳定的特点,对FCM算法进行改进,将聚类中心分为活动和稳定两类,忽略稳定聚类中的距离计算,以此降低FCM算法的计算复杂性。实验结果表明了该算法的有效性。

本实验数据中,只根据服装风格这一特征属性对用户项目兴趣度进行了映射,其实在实际应用过程中,可以将服装风格、颜色、款式等多种特征进行组合,以提高Web用户聚类的准确性。下一步工作中,将结合用户Web浏览日志、购买记录、项目评分等多种数据源,建立更科学的用户兴趣评价模型,优化用户聚类算法,提高聚类有效性。

参考文献:

[1] 王宏超,陈未如,刘 俊.基于客户聚类的商品推荐方法的研究[J].计算机技术与发展,2008,18(7):212-214.

[2] Kim Y S. Weighted order-dependent clustering and visualization of web navigation patterns[J]. Decision Support System, 2006,13(8):38-70.

[3] 吴 瑛,王秋生.模糊C均值聚类算法在Web使用挖掘上的应用研究[J].计算机技术与发展,2008,18(6):32-35.

[4] 陈 敏,苗夺谦,段其国.基于用户浏览行为聚类Web用户[J].计算机科学,2008,35(3):186-187.

[5] 张文东,易铁虎.基于兴趣相似性的Web用户聚类[J].山东大学学报(理学版),2006,41(3):54-57.

[6] 李 聪,梁昌勇.基于属性值偏好矩阵的协同过滤推荐算法[J].情报学报,2008,27(6):884-890.

[7] Berget I, Vazirgiannis B H. Web mining for web personalization[J]. ACM Transactions on Internet Technology, 2003, 3(1):1-27.

[8] 李 雷,罗红旗,丁亚丽.一种改进的模糊C均值聚类算法[J].计算机技术与发展,2009,19(12):71-73.

[9] Fan J L, Zhen W Z, Xie W X. Suppressed Fuzzy C-means Clustering Algorithm[J]. Pattern Recognition Letters, 2003, 48(5):1607-1612.

[10] Hoppner F, Klawonn F, Kruse R, et al. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition[M]. [s.l.]: John Wiley & Sons, 1999:1503-1517.

[11] Fu K S. Digital Pattern Recognition[M]. New York: Springer-Verlag, 1976:47-94.

[12] Lai J Z C, Liaw Y C, Liu J. A fast VQ codebook generation algorithm using codeword displacement[J]. Pattern Recognition, 2008, 41(1):315-319.

[13] Song Qinbao, Martin S. Mining web browsing patterns for e-commerce[J]. Computer in Industry, 2006, 57(7):622-630.