文章编号:1672-3961(2011)06-0043-07

一种新的基于网络虚拟环境的用户访问模式聚类算法

陈明志[1,2], 陈健[3], 许春耀[3], 余轮[3], 林柏钢[1,2]
(福州大学 1. 数学与计算机科学学院; 2. 网络系统信息安全福建省高校重点实验室;
3. 物理与信息工程学院, 福建 福州 350108)

摘要:为了有效地实现网络虚拟环境的个性化信息推荐,提出一种针对网络三维虚拟环境的用户访问模式聚类算法,即基于多目标粒子群优化的模糊*C*-均值聚类算法(MOPSO-based FCM, MPF)。MPF算法结合了粒子群优化算法(particle swarm optimization, PSO)与模糊*C*-均值算法(fuzzy *C*-means, FCM)的优点,通过PSO的全局空间搜索避免了FCM算法对初始值、噪声数据敏感与容易陷入局部最优等。为了改善聚类效果,在PSO中设计一个基于双目标(最小化类内距离与最大化类间距离)的粒子适应度函数。最后用标准数据集与模拟数据集分别对MPF算法进行性能测试,实验结果表明:本算法在聚类精度方面表现良好。

关键词:网络虚拟环境;用户访问模式聚类;多目标粒子群优化;模糊*C*均值
中图分类号:TP301.6 **文献标志码**:A

A new clustering algorithm for user access patterns based on network virtual environments

CHEN Ming-zhi[1,2], CHEN Jian[3], XU Chun-yao[3], YU Lun[3], LIN Bo-gang[1,2]
(1. College of Math and Computer Science,
2. Key Lab of Information Security of Network Systems (Fujian Province University),
3. College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China)

Abstract: In order to efficiently implement personalized information services in network virtual environments, a new clustering algorithm for user access patterns was proposed, which was the MPF, i.e. the fuzzy *C*-means (FCM) clustering algorithm based on multi-objects particle swarm optimization (MOPSO). The MPF could combine the respective advantages of PSO and FCM. Through the global spatial search of PSO, it could avoid that FCM was susceptible to initial value, noisy data and easily falling into the local optimum. In order to improve the clustering effect, a particle fitness function was designed based on dual-objectives (intra-class distance and inter-class distance) in PSO. Finally, the standard data set and simulation data set were applied to verify the effectiveness of this MPF. Experimental results showed that this algorithm had good performance in clustering precision.

Key words: network virtual environments; user access patterns; multi-objects particle swarm optimization; fuzzy *C*-means

0 引言

近些年个性化信息推荐技术与应用系统在互联网的电子商务、音乐、电影等领域中得到广泛的应用[1-2],如:WebWatcher、SiteSeer、Movie Recommendation等[3-4]。在网络虚拟环境中,由于三维空间的不适应感与资源信息的非直观性,一方面使用户很难从中发现自己感兴趣的信息,另一方面也使得某些有价值的信息由于不为人所知而成为"暗信息"。

收稿日期:2011-07-11
基金项目:福建省自然科学基金项目(2011J01346);福州大学科研基金资助项目(XRC-1039)
作者简介:陈明志(1975-),男,福建古田人,讲师,博士,主要研究方向为智能信息处理,虚拟环境等. E-mail:donres@fzu.edu.cn

因此,文献[5]提出基于网络虚拟环境的智能导航概念,希望为用户提供个性化主动的信息服务,这种信息推荐技术利用用户之间的相似性关系挖掘当前用户潜在感兴趣的对象。因此,如何将具有相似兴趣的用户进行有效聚类是个性化信息推荐成功的关键。用户访问模式聚类是 Web 数据挖掘技术的重要研究方向[6-7]。通过用户对网站的使用信息(Web 日志文件)的处理和研究,得到具有相似访问兴趣的用户群体和用户共同感兴趣的 URL,据此调整站点的结构并进行个性化信息推荐。本研究拟解决在网络三维虚拟环境中用户访问模式聚类的相关问题,如虚拟场景拓扑图、场景关联度、用户访问路径等,特别是用户聚类问题。

1 虚拟场景及用户访问路径处理

三维虚拟环境是由一系列的虚拟场景构成的,本研究定义每一个相对独立的场景为一个浏览点,这类似二维网站的一个网页。具体实现方式:在前期设计虚拟场景时在场景的出入口处或主要通道的关键点设置前后顺序的两个触发点,当虚拟人先后通过这两个触发点的时候,就表明它的浏览方向(是进入还是离开)与浏览意图(具体的访问点)。

假设虚拟环境有 n 个虚拟场景(即浏览点),浏览者是对这 n 个场景进行随机的访问,在三维虚拟环境下,浏览者可以根据导航图进行飞行式的访问。为了实现虚拟场景中用户有效聚类,下面定义场景拓扑结构、场景关联性与用户访问路径差异度等相关概念。

1.1 场景拓扑结构

虽然浏览者可以利用导航图飞行实现瞬时任意场景的到达,但静态上场景还是有对应的拓扑结构图,如图 1 所示,结构图的连接线表示非飞行状态下时场景之间的前后顺序关系,虚线表示用户的访问路径。

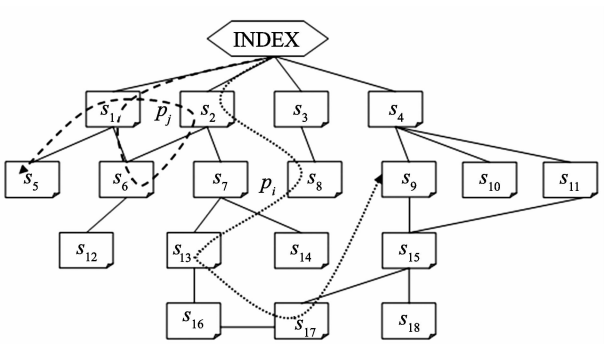

图 1 静态场景拓扑结构图

Fig. 1 Topology map of static virtual scenes

1.2 场景关联性

若有用户访问路径:$p_1 = \{\text{Index} \rightarrow s_1 \rightarrow s_5\}$;$p_2 = \{\text{Index} \rightarrow s_1 \rightarrow s_6 \rightarrow s_{12}\}$;$p_3 = \{\text{Index} \rightarrow s_2 \rightarrow s_6 \rightarrow s_{12}\}$;$p_4 = \{\text{Index} \rightarrow s_2 \rightarrow s_7 \rightarrow s_{13} \rightarrow s_{16} \rightarrow s_{17}\}$,…。

要度量用户访问路径之间的差异性就要先计算访问路径中各场景之间的关联性[8],因此定义在三维虚拟环境下场景之间的关联性。

定义 1 场景置信度

一个场景 s_i 对另一场景 s_j 的置信度 $C(s_i \Rightarrow s_j)$ 或简写为 C_{ij} 可以定义如下:

$$C(s_i \Rightarrow s_j) = \frac{\sup\_\text{count}(s_i \cup s_j)}{\sup\_\text{count}(s_i)}, \qquad (1)$$

其中 s_i、s_j 表示三维虚拟环境中任意的 2 个场景:$\sup\_\text{count}(s_i \cup s_j)$ 表示场景 s_i 和 s_j 出现在同一条访问路径中的次数,即日志数据库中一定时期内浏览者的访问路径统计。

定义 2 置信度矩阵

置信度矩阵表征场景之间互为置信的程度,设矩阵 $\check{\boldsymbol{C}} = (C_{ij})(i,j = 1, 2, \cdots, n)$,其中 $C_{ij} \in [0,1]$ 表示场景 s_i 对另一场景 s_j 的置信度,n 为场景总数。因场景对自身的关联性最高,故矩阵的主对角线应为 1,

$$\check{\boldsymbol{C}} = \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,n} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,n} \\ \cdots & C_{i,j} & \cdots & \cdots \\ C_{n,1} & C_{n,2} & \cdots & C_{n,n} \end{bmatrix}。$$

定义 3 关联度矩阵

关联矩阵表征所有的场景之间的关联性,设关联矩阵 $\check{\boldsymbol{R}} = (R_{ij})$ $(i,j = 1, 2, \cdots, n)$,求置信度矩阵对称位置上 2 个值的平均值,并替换矩阵左下方对应的值,同时删除另一个值,即 $R_{ij} = (C_{ij} + C_{ji})/2$,$R_{ij} \in [0,1]$ 表示场景 s_i 和 s_j 之间的关联度,

$$\check{\boldsymbol{R}} = \begin{bmatrix} 1 & & & \\ R_{2,1} & 1 & & \\ \cdots & R_{i,j} & 1 & \\ R_{n,1} & R_{n,2} & \cdots & 1 \end{bmatrix}。$$

1.3 用户访问路径差异度

当用户以完全不同的路径来浏览网页时,会产生不同的访问路径,设 $p_2 = \{\text{Index} \rightarrow s_1 \rightarrow s_6 \rightarrow s_{12}\}$ 和 $p_4 = \{\text{Index} \rightarrow s_2 \rightarrow s_7 \rightarrow s_{13} \rightarrow s_{16} \rightarrow s_{17}\}$ 分别代表 2 条访问路径中所含的场景集合,则 $p_2 = \{s_1, s_6, s_{12}\}$ 与 $p_4 = \{s_2, s_7, s_{13}, s_{16}, s_{17}\}$ 集合的笛卡尔积 $p_2 \times p_4$ 集合中的每个元素可以代表 2 条路径中场景配对情况,来计算元素的关联度。

定义 4　访问路径差异度

设有 2 条用户访问路径 $p_i=\{s_1^i, s_2^i,\cdots,s_m^i\}$, $p_j=\{s_1^j, s_2^j,\cdots,s_n^j\}$,对 p_i、p_j 的笛卡尔积 $p_i\times p_j$ 集合中的每个元素求关联度 $\boldsymbol{R}(s_l^i, s_k^i)$,将所有的关联度叠加后除 $m\times n$ 即得访问路径 p_i、p_j 的差异度[8]:

$$d(p_i,p_j)=\Big(\sum_{l=1}^{m}\sum_{k=1}^{n}\boldsymbol{R}(s_l^i,s_k^j)\Big)/m\times n。\quad(2)$$

2　聚类的研究策略

聚类是将物理或抽象对象的集合分组成为由类似的对象组成的多个类的过程,其原则是最大化类内的相似性、最小化类间的相似性[9]。因此所谓的用户聚类是指将具有相似访问模式的用户群体归于一类。典型的聚类算法有 *C*-Means、DBSCAN、CLARANS,BIRCH 与 CURE 等[10]。文献[10]认为 *C*-Means 方法属于划分算法即事先需要指定类别数、具有较高聚类效率,但 *C*-Means 方法属于硬聚类技术,即把每个待辨识对象严格地划分到某个确定的类中,不能处理类间重叠问题。但实际上大多数事物并没有严格的属性,类属方面存在着不确定性,所以文献[11]认为 FCM 聚类算法使用隶属度来描述数据对象隶属各个类的不确定性,能够比较客观地反映现实世界,并且对数据的比例变化具有鲁棒性。

考虑到网络虚拟环境中的用户聚类算法是有导师的学习过程,且注重算法的聚类效率,因此 FCM 满足网络虚拟环境中的聚类需求,适合此环境的用户模式聚类,但 FCM 存在缺点:对初始值与噪声数据较敏感,容易陷入局部最优。

因此,引入了 PSO 技术,提出适用于网络虚拟环境的基于粒子群优化的模糊 *C*-均值聚类算法(PSO-based FCM)。先以 PSO 算法求得近似最优解,然后将其作为 FCM 算法的初始值,继续进行局部搜索以求得全局最优解。它既克服了 FCM 算法易陷入局部最优解的缺陷,又能解决 PSO 算法只能找到近似最优解的问题,同时弥补了 FCM 算法对初始值及噪声点比较敏感的不足,且提高聚类效率、加快收敛速度。

聚类准则函数的优化目标是最大化类内的相似性、最小化类间的相似性,显然这是一个多目标优化问题。多数文献[12-13]的聚类准则函数只考虑所有类的类内距离和,没有考虑到类间距离,因此它们的聚类准则函数只考虑最大化类内的相似性,而不能保证最小化类间相似性。

为了实现以上的双优化目标,在求解多目标优化问题的 PSO 算法(multi-objective particle swarm optimization, MOPSO)中设计了 1 个基于双目标优化的准则函数,即适应度函数 $f(p_i)$。当搜索到 $f(p_i)$ 为最大值时,意味着达到类内距离尽可能小、类间距离尽可能大的目标,并且通过调整子函数前的 2 个权值 w_1、w_2 可以给出不同优先级的搜索策略。

3　基于 MOPSO 的 FCM 聚类算法

3.1　模糊 *C*-均值聚类算法

FCM 将 n 个样本 $X_i(i=1,2,\cdots,n)$ 分为 k 个模糊类,并求每类的聚类中心,使得非相似性指标的价值函数达到最小。设有限样本集 $X=\{X_1, X_2,\cdots, X_k\}$ 属于 d 维欧几里德空间 $\mathbf{R}^d$,即 $X_i\in\mathbf{R}^d$, $i=1,2,\cdots,n$ 为样本点。k 为大于 1 的整数,将样本空间 X 分为 k 类,聚类中心 $\check{\boldsymbol{C}}=\{C_1, C_2,\cdots, C_k\}$。可以用一个模糊矩阵 $\boldsymbol{U}=(\mu_{ij})$ 表示分类情况,μ_{ij} 表示第 i 个样本点属于第 j 类的隶属度,显然 μ_{ij} 满足如下条件:

$$\sum_{j=1}^{k}\mu_{ij}=1,\quad \mu_{ij}\in[0,1],$$
$$\forall i=1,2,\cdots,N;\quad j=1,2,\cdots,k。\quad(3)$$

FCM 算法采用误差平方和函数作为聚类准则参数,即目标函数 $J(\boldsymbol{U},\check{\boldsymbol{C}})$。

$$J(\boldsymbol{U},\check{\boldsymbol{C}}))=\sum_{i=1}^{n}\sum_{j=1}^{k}\mu_{ij}^m d_{ij}^2,\quad(4)$$

其中,$m\in[1,\infty)$ 为模糊指数, $d_{ij}=\|X_i-C_j\|$ 为第 i 个数据点与第 j 个聚类中心间的欧几里德距离。FCM 算法就是将目标函数 J 最小化的迭代过程,在迭代化过程中的 $\boldsymbol{U}$、$\check{\boldsymbol{C}}$ 取值如下:

$$\mu_{ij}=\begin{cases}\Big[\sum_{h=1}^{k}\dfrac{\|X_i-C_j\|^{2/(m-1)}}{\|X_i-C_h\|^{2/(m-1)}}\Big]^{-1}, & \|X_i-C_h\|\neq 0;\\ 1, & \|X_i-C_h\|=0\cap j=h;\\ 0, & \|X_i-C_h\|=0\cap j\neq h。\end{cases}\quad(5)$$

$$C_j=\frac{\sum_{i=1}^{n}\mu_{ij}^m X_i}{\sum_{i=1}^{n}\mu_{ij}^m}。\quad(6)$$

3.2　基于网络虚拟环境的 MOPSO 算法

PSO 算法与其它的进化算法相比,其最吸引人的特征是简单实现和更强的全局优化能力。

设粒子群的种群规模为 N,决策空间 n 维,其中粒子 i 在时刻 t 的坐标位置可以表示为 $X_i^t=(x_{i1}^t, x_{i2}^t,\cdots, x_{in}^t)$,粒子 i 的速度定义为每次迭代中粒子移动的距离,用 $V_i^t=(v_{i1}^t, v_{i2}^t,\cdots, v_{in}^t)$ 表示,则粒子

i 在时刻 t 的第 j 维子空间中的飞行速度和位置根据下式进行调整:

$$v_{ij}^{t} = wv_{ij}^{t-1} + c_1\mathrm{rand}_1(p_{ij} - x_{ij}^{t-1}) + c_2\mathrm{rand}_2(g_j - x_{ij}^{t-1}), \tag{7}$$

$$x_{ij}^{t} = x_{ij}^{t-1} + v_{ij}^{t}, \tag{8}$$

其中 w 为惯性权值;c_1 和 c_2 为加速因子;r_1 和 r_2 是在[0,1]范围内的2个随机数。通常使用1个常量 $V_{\max}$ 来限制粒子的速度,改善搜索结果。g_j 是代表领袖粒子(Leader)的位置,此处是整个粒子群中的历史最优位置记录(全局极值 g_{best}),也可以是局部粒子群的历史最优位置,此时 g_j 可改为 l_j(局部极值 l_{best}),p_{ij} 是当前粒子的历史最优位置记录(个体极值 p_{best})。

多目标优化问题是指多于1个的数值目标在给定区域上的最优化问题。解决的最终手段:在各子目标之间进行协调权衡和折衷处理,使各子目标函数都尽可能达到最优。对实际应用问题,必须根据问题和决策人员的个人偏好,从 Pareto 最优解集合中挑选出1个或一些解作为问题的最优解。

3.2.1 粒子的编码

聚类样本空间 $X = \{X_1, X_2, \cdots, X_N\}$,其中 X_i 为 d 维。粒子的编码思想为以 PSO 中的1个微粒代表1个簇中心的集合 $p_i(C_{i1}, C_{i2}, \cdots, C_{ic})$,其中 $C_{ij}(j = 1,2,\cdots,C)$ 是与 X_i 同维,代表第 i 个粒子的第 j 类中心点的坐标点。

PSO 是对聚类的簇中心点进行编码,而簇中心点也就是访问路径,根据图1,我们得知访问路径是对图中的结点按某种的访问序列进行组合,生成元素有序的集合。因为这样的簇中心点是在算法迭代过程中生成的,所以对应的访问序列不一定沿静态场景拓扑图中的连接线行进,这是合理的。这是因为在三维虚拟环境中虚拟人飞行漫游浏览者的访问路径可不遵循静态的场景先后顺序的布局。

采取离散二进制编码,编码 $\{s_1, s_2, \cdots, s_l, \cdots, s_n\}$,其中 n 为虚拟环境中的的场景数,当 $s_l \in p_i$ 时,$s_l = 1$;当 $s_l \notin p_i$ 时,$s_l = 0$。如图1中虚线所代表的访问路径 $p_i = \{\mathrm{Index} \rightarrow s_2 \rightarrow s_8 \rightarrow s_{13} \rightarrow s_{17} \rightarrow s_9\} \cdots$,其编码为 $p_i = \{0100000110001000 10\}$。在这里为了降低算法的复杂性做了简化处理,即不考虑路径的访问顺序。

3.2.2 粒子的适应度函数

对粒子的适应度函数构造采用了基于表现型共享函数的构造方式。首先计算种群中基于 Pareto 概念下模糊聚类的类间距离度量函数与类内距离度量函数的值,并对2个函数值乘上调整因子(w_1,w_2),然后线性组合成个体的适应度函数。

定义5 类内距离函数

模糊聚类的类内距离为所有样本(访问路径)p_i 与聚类簇中心点 c_j 之间的表现型距离,设有 c 个簇中心点,记 $\mu_{i1}^m d_{i1}^2$、$\mu_{i2}^m d_{i2}^2$、$\cdots$、$\mu_{ic}^m d_{ic}^2$ 分别为样本 p_i 与 c 个簇中心点的分量距离,其中的 μ_{ij}^m 为表示第 i 个样本属于第 j 类的隶属度,$d_{ij} = \| X_i - C_j \|$ 为第 i 个样本与第 j 个簇中心的欧几里德距离,则类内距离函数为

$$f_e = \sum_{i=1}^{N}\sum_{j=1}^{C}\mu_{ij}^m d_{ij}^2。 \tag{9}$$

根据以上定义的访问路径差异度式(2),定义如下的类间距离。

定义6 类间距离函数

设有 C 个的聚类簇中心$(c_1, c_2, \cdots, c_c)$,记 $d(c_j, c_k)$ 为簇中心 c_j 与 c_k 的差异度,则类间距离函数为

$$f_d = \frac{2}{C(C-1)}\sum_{j=1}^{C-1}\sum_{k=j+1}^{C} d^2(c_j, c_k)。 \tag{10}$$

定义7 粒子的适应度函数

考虑类内和类间距离函数,通过求 $f(p)$ 的最大值,类内距离尽可能小、类间距离尽可能大,且调整权值 w_1、w_2 可以给出不同的优先搜索策略,则粒子的适应度函数为

$$f(p) = w_1(1/f_e) + w_2(f_d), \tag{11}$$

其中$(w_1, w_2 \mid w_1 \geqslant 0 \cap w_2 \geqslant 0 \cap w_1 + w_2 = 1)$。

利用这个参数组合构造个体的适应度函数的好处:(1)是评价了个体的优劣性;(2)是通过调整权值 w_1、w_2 可以给出不同优先级的搜索策略;(3)是利用个体表现型上在种群中的密集度保证种群能维持较高的多样性。

3.3 MPF 算法

综上所述,给出 MPF 的实现步骤:

Step 1 初始化随机产生 Z 个粒子 $P = \{p_1, p_2, \cdots, p_i, \cdots, p_Z\}$,其中粒子 p_i 为一个随机产生的簇中心的集合,根据簇中心数 C,可以从样本集 $X = \{X_1, X_2, \cdots, X_N\}$ 中随机选择 C 个向量来初始化粒子 p_i。另初始化2个外部档案为空。

Step 2 迭代次数小于指定值时,重复

根据式(5)计算模糊矩阵 $\boldsymbol{U} = (\mu_{ij})$;

对于每个粒子:

① 根据式(9)与(10)计算 f_e 与 f_d,再依式(11)计算粒子的适应度函数 $f(p_i)$;

② 更新指导粒子外部档案 A_1,从中选择一个粒子作为指导粒子;

③ 比较并更新粒子的个体历史最优与粒子群全局最优;

④ 根据式(7)、(8)更新粒子速度与位置;

⑤ 重复步骤①~④,直到种群中的所有粒子更新完毕,更新历代非劣解外部档案 A_2。

Step 3　输出历代非劣解外部档案 A_2 作为最终解。

Step 4　对非劣解排序,取前 3 个解,即得到 3 个簇中心的集合,根据式(5)得 3 个样本的隶属度矩阵,矩阵合并求均值得一个隶属度矩阵。

4　实验结果与分析

本研究实验环境:操作系统 Windows XP Professional with SP2;硬盘:80G;CPU:Intel Pentium ® Dual Core 3.20 GHz;内存:1.50GB,主板 Intel 955X;显示卡:NVIDIA Quadro FX540。

分 2 种情况进行性能评价:(1)利用标准的数据集对聚类算法相关性能进行测评;(2)根据虚拟环境系统的模拟用户使用情况设计多组的模拟数据,用此对 2 种算法进行性能测评。

采用国际著名的机器学习资源库 UCI 提供的有关评价聚类与推荐系统的数据集 Netflix Prize,从中选取 5% 作为测试数据。

4.1　标准数据集测评

为了表明 MPF 算法的性能改进程度,需要将其与经典算法进行比较。选择基本 *C*-Means 算法,FCM 算法与 MPF 算法从用户聚类精度(users clustering precision, UCP)[14] 与计算时间 2 个指标进行实验对比。

首先生成大小(data set size, DSS)分别为 200 kb、400 kb、600 kb、800 kb、1 000 kb、1 200 kb 的 6 组数据用于以上 3 个算法的训练和扩展性测试。再随机生成 3 组大小相同(600 kb)的数据集用于测试算法的平均性能。

4.1.1　算法训练与精度测评

应用 6 组不同大小的数据集训练时所得聚类精度如图 2 所示。

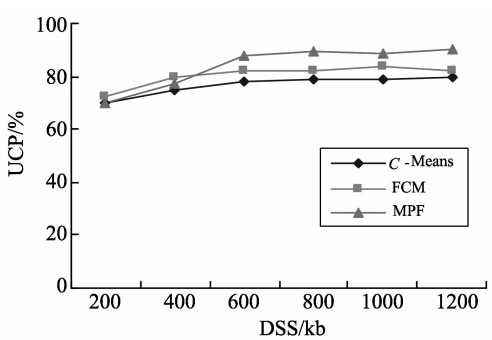

图 2　数据集规模与聚类精度的关系

Fig. 2　Relationship between data set size and users clustering precision

在完成第 1、2 组训练集测试时发现 3 个算法的精度都不是太理想,于是调整算法相关参数;在第 3 组数据集测试后 MPF 算法的精度有了明显提高,而 *C*-Means 与 FCM 算法精度提高不大;当第 4、5、6 组测试后 3 个算法的聚类精度均改变不大,这说明算法训练完成。

现应用另外 3 组大小均为 600 kb 的数据集测试算法的平均聚类精度,结果如表 1 所示。

表 1　聚类精度比较

Table 1　Comparison of users clustering precisions

算法	UCP/%		
	C-Means	FCM	MPF
第 1 组	77.6	76.4	82.3
第 2 组	70.0	79.3	90.2
第 3 组	79.6	80.5	87.3
平均值	75.7	78.73	86.6

以上得知:*C*-Means 聚类精度在 76% 左右,FCM 聚类精度在 79% 左右,MPF 聚类精度在 87% 左右。可见 MPF 算法的聚类精度在 3 个算法中是最高的。

4.1.2　算法耗时测评

训练时算法的计算时间(computing time, CT)见图 3。

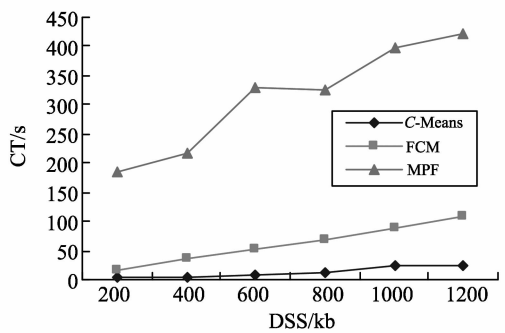

图 3　数据集规模与计算时间的关系

Fig. 3　Relationship between data set size and computing time

以上结果表明:(1)随着数据规模的增大,3 种算法计算时间的增长基本呈现出线性的上升趋势,表明 MPF 算法具有较好的扩展性。(2)*C*-Means 算法的计算耗时最少,FCM 耗时居中,而 MPF 最多。经分析,原因是 FCM 相比 *C*-Means 算法增加了模糊处理,故计算时间多些;而 MPF 则是在 FCM 的基础上进行 PSO 的全局空间的搜索,因此比前两者有较大的时间耗费也就不足为奇了。另由于 PSO 搜索具有一定的随机性,故每组耗时不呈严格的单调递增。这是因为聚类计算是离线进行的,所以计算时间在此处是 1 个次要的指标,而聚类精度是算法的关键指标。

4.2 模拟数据集测试

为了测试 MPF 算法在网络虚拟环境中的工作性能,分别从 2 个指标,即聚类准确率与精度来进行比较。参考文献[14]定义网络虚拟环境的用户聚类精度:分析测试数据集,从中找出用户 i 与相应的场景浏览点 S_i 集。然后利用算法计算对用户进行聚类,决定用户 i 的归属类,并得出其相应的预测浏览点 S'_1,则用户聚类精度如下:

$$UCP = \frac{1}{n}\sum_{i=1}^{n}\frac{\| S_i - S'_i \|}{\| S_i \|}。 \quad (12)$$

4.2.1 聚类准确率

构建 1 个网络虚拟环境,其中设定有效浏览点数为 35 处,参加测试用户数为 30 人。实验要求在 20 d 内,按照各自的兴趣浏览虚拟场景,总浏览时间不少于 40 h。20 d 后将记录在日志数据库中的所有用户使用信息进行人工分析与整理,获取相似的用户群与聚类中心。随机抽取 6 组数据作为测试样本,再采用 3 个算法对这 6 组数据集进行计算,然后根据算法计算结果与手工整理结果对比求得聚类准确率(clustering accuracy, CA),如图 4 所示。

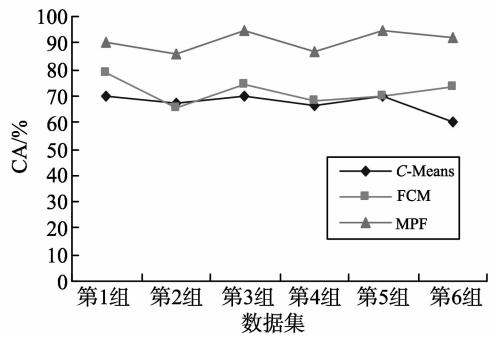

图 4 聚类准确率

Fig. 4 Clustering accuracy

从上图得知:MPF 聚类算法对虚拟环境的模拟数据集具有最好的聚类准确率。

4.2.2 聚类精度

从图 5 中数据对比得知:MPF 不仅对标准数据集有良好的表现,而且对网络虚拟环境的模拟数据集依然发挥稳定,并有较高的聚类精度。

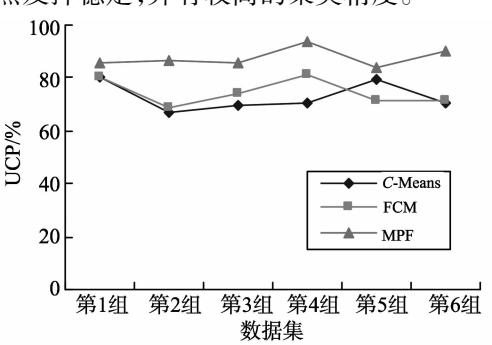

图 5 用户聚类精度

Fig. 5 Users clustering precision

5 结论

本研究提出面向网络虚拟环境的基于 MOPSO 的 FCM 算法。体现以下 2 个工作点:

(1) 在用户聚类方面,提出基于多目标粒子群优化的 FCM 聚类算法(即 MPF),先以 MOPSO 算法求得近似最优解,然后将其作为 FCM 算法的初始值,继续进行局部搜索以求得全局最优解,该方法有效解决 FCM 对初始值与噪声数据的敏感,容易陷入局部最优等问题。

(2) 为了改善聚类效果,实现类内距离尽可能的小,同时类间距离尽可能大的目标,设计了 1 个基于双目标(即类内距离函数与类间距离函数)优化的准则函数,即适应度函数 $f(p_i)$,当搜索到 $f(p_i)$ 为最大值,就可满足最大化类内的相似性、最小化类间的相似性的原则。而且通过调整距离函数前的 2 个权值 w_1、w_2 可以给出不同优先级的搜索策略。

最后用标准数据集与模拟数据集对聚类算法进行测评,实验结果表明:本研究所提的聚类算法在聚类精度与准确率等方面都有良好表现。

参考文献:

[1] CHANGCHIEN S W, LEE C, HSU Y. On-line personalized sales promotion in electronic commerce [J]. Expert Systems with Applications, 2004, 27(1):35-52.

[2] 许海玲,吴潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报. 2009, 20(02): 350-362.
XU Hailing, WU Xiao, LI Xiaodong, et al. Comparison study of internet recommendation system [J]. Chinese Journal of Computers, 2009, 20(02): 350-362.

[3] HERLOCKER J L, KONSTAN J A, RIEDL J. Explaining collaborative filtering recommendations [C]// Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. New York: ACM Press, 2000: 241-250.

[4] LEKAKOS G, CARAVELAS P. A hybrid approach for movie recommendation [J]. Multimedia Tools and Applications, 2008, 36 (1-2): 55-70.

[5] 陈明志,许春耀,余轮. 三维网站智能导航系统的设计与实现[J]. 计算机工程与设计, 2010(10): 1-5.
CHEN Mingzhi, XU Chunyao, YU Lun. Design and implementation of intelligent navigation system based on 3D website [J]. Computer Engineering and Design, 2010 (10): 1-5.

[6] 郭岩,白硕,杨志峰,等. 网络日志规模分析和用户兴趣挖掘[J]. 计算机学报, 2005, 28(09):1483-1495.

GUO Yan, BAI Shuo, YANG Zhifeng, et al. Analyzing scale of Web logs and mining users' interests[J]. Chinese Journal of Computers, 2005, 28(09): 1483-1495.

[7] 邢东山,沈钧毅,宋擒豹. 从 Web 日志中挖掘用户浏览偏爱路径[J]. 计算机学报, 2003, 26(11): 1518-1523.
XING Dongshan, SHEN Junyi, SONG Qinbao. Discovering preferred browsing paths from Web logs[J]. Chinese Journal of Computers, 2003, 26(11): 1518-1523.

[8] 朱兴亮,游中胜,王勇. Web 用户访问路径的差异性度量方法研究[J]. 计算机科学, 2006, 33(7): 104-106.
ZHU Xingliang, YOU Zhongsheng, WANG Yong. Research on Web user access path's difference measurement [J]. Chinese Computer Science, 2006, 33(7): 104-106.

[9] 姜园,张朝阳,仇佩亮,等. 用于数据挖掘的聚类算法[J]. 电子与信息学报, 2005, 27(04): 655-662.
JIANG Yuan, ZHANG Zhaoyang, QIU Peiliang, et al. Clustering algorithms used in data mining[J]. Jounral of Electronics & Information Technology, 2005, 27(04): 655-662.

[10] HAN J, KAMBER M. Data mining: concepts and techniques[M]. Massachusetts: Morgan Kaufmann Publishers, 2006: 48-55.

[11] 张敏,于剑. 基于划分的模糊聚类算法[J]. 软件学报, 2004, 15(6): 858-868.
ZHANG Min, YU Jian. Fuzzy partitional clustering algorithms[J]. Journal of Software, 2004, 15(6): 858-868.

[12] 王玲,贺兴时. 基于 PSO 的模糊 *C* 均值聚类算法[J]. 甘肃联合大学学报:自然科学版, 2008, 22(02): 78-81.
WANG Ling, HE Xingshi. PSO-based fuzzy *C*-mean clustering algorithm[J]. Journal of Gansu Lianhe University:Natural Sciences, 2008, 22(02): 78-81.

[13] 许磊,张凤鸣. 基于 PSO 的模糊聚类算法[J]. 计算机工程与设计, 2006, 27(21): 4128-4129.
XU Lei, ZHANG Fengming. Fuzzy clustering algorithm based on PSO[J]. Computer Engineering and Design, 2006, 27(21): 4128-4129.

[14] MARTI′N-GUERRERO J D, PALOMARES A, BALAGUER-BALLESTER E, et al. Studying the feasibility of a recommender in a citizen web portal based on user modeling and clustering algorithms [J]. Expert Systems with Applications, 2006, 30(2): 299-312.

(编辑:陈燕)

工学版编委 Musharraf Zaman 教授访问编辑部

应山东大学自然科学学报编辑部的邀请,《山东大学学报(工学版)》国际编委、美国俄克拉荷马大学教授 Dr. Musharraf Zaman 于日前访问自然学报编辑部。

2011 年 11 月 23 日下午,在《山东大学学报(工学版)》主编、山东大学土建与水利学院院长李术才教授、姚占勇教授、孙仁娟博士、葛智博士的陪同下,Dr. Zaman 与学报编辑部靳光华主任和工学版全体编辑人员进行了座谈。会上,编辑部靳光华主任向 Zaman 教授颁发了编委聘书,靳光华主任和工学版执行主编胡春霞分别介绍了编辑部和工学版的具体情况。Zaman 教授就如何提高期刊载文质量、提高期刊引用率、选题组稿等方面提出建设性的意见和建议,就如何提高工学版学术质量、加快国际重要数据库收录等问题进行了商讨,并愉快地接受了约稿。

Zaman 教授现任美国俄克拉荷马大学终身教授,俄克拉荷马大学工学院负责研究和研究生教育的副院长。美国土木工程师协会(ASCE)成员、美国交通研究委员会(TRB)大学代表、国际计算方法和岩土力学发展协会会员 (IACMAG)理事、国际计算方法和岩土力学发展协会会员 (IACMAG)仲裁委员会联席主席、国际岩土力学杂志主编、国际岩土工程编委、路面研究和科技杂志编委、山东大学学报(工学版)编委等。

(工学编辑室)