基于蚁群模糊聚类的协同过滤推荐算法

黄金凤，雷筱珍
（福建交通职业技术学院 信息系，福州 350007）

摘　要：为解决传统协同过滤算法在产生推荐时实时性较差性问题，提出了一种基于蚁群模糊聚类的协同过滤推荐算法. 该算法将分两个步骤产生推荐. 离线时，应用蚁群模糊聚类技术，对基本用户进行聚类；在线时，利用已有的用户蚁群聚类寻找目标用户的最近邻居，并产生推荐. 实验表明，基于蚁群模糊聚类的协同过滤推荐算法能提高推荐产生的速度，即实时性得到了一定的提高.

关键词：推荐算法；协同过滤；蚁群模糊聚类；MAE
中图分类号：TP18；TP311　　**文献标识码**：A　　**文章编号**：1671－119X(2011)04－0055－04

0　引　言

企业通过电子商务网站能增强自身的竞争优势，个人通过使用电子商务网站能感受到足不出户购物的方便与快乐. 但是电子商务网站存在一个亟待解决的问题：推荐适合当前浏览用户的商品给该浏览用户，从而避免用户由于在过多的商品中找到自己所需商品过于耗时耗力而离开. 电子商务推荐系统就是解决此类问题的解决方案. 许多大型电子商务网站早已开始使用电子商务推荐系统，如 Amazon、当当网等.

在推荐系统中，协同过滤（Collaborative Filtering，CF）正迅速成为一项很受欢迎的技术. 协同过滤分析用户兴趣，即用户会对邻居所喜欢的商品产生兴趣. 在用户群中找到指定用户的相似兴趣用户，即邻居用户，综合这些邻居用户对某一项目的评价，形成系统对该指定用户对此项目的喜好程度预测，进而将预测评价最好的前 N 项商品推荐给该指定用户.

当网站的用户和项目数量增加时，CF 的算法复杂度迅速递增，从而使得系统推荐性能不断下降，最终影响推荐的及时性[1][2]. 正是鉴于该问题，本文提出了基于蚁群模糊聚类的协同过滤推荐算法（CF－based ACVC，Collaborative filtering recommendation algorithm based on Ant Colony Vague clustering）. 基本思想是分先离线再在线两阶段. 离线时利用蚁群模糊聚类算法对用户进行聚类，生成若干用户聚类中心，再计算每个用户和各聚类中心的相似性以得到相似性度量矩阵；在线时计算目标用户与各聚类中心的相似性，再以此搜索相似性度量矩阵找到其最近邻居并进行评论预测，最后生成推荐. 同时，仿真实验结果表明，本算法在一定程度上提高了推荐速度和质量.

1　协同过滤算法

协同过滤算法是基于评分相似的最近邻居的评分数据向目标用户产生推荐. 目标用户对未评分项目的评分可以通过最近邻居对该项目评分的加权平均值逼近. 它通过构造用户对项目的偏好数据集来实现.

算法 1 协同过滤算法

协同过滤算法的输入数据通常表述为一个 $m\times n$ 的用户－项目评价矩阵 $R(m,n)$，其中 m 行表示 m 个用户，n 列表示 n 个项目，矩阵元素 $R_{i,j}$ 表示用户 i 对项目 j 的评估值.

首先，计算每个用户对以往评价过的信息资源项目的平均打分：

$$\overline{R}_u=\frac{1}{|I_u|}\sum_{j\in Iu}R_{uj} \tag{1}$$

其中，I_u 为用户 u 的评分向量，$|I_u|$ 为 I_u 的长度，即用户打过分的数字资源数目，R_{uj} 表示用户 u 对项目 j 的评分.

其次，计算目标用户对未评价过的信息资源项目的预测评分值：

$$P_{uj}=\overline{R}_u+k\sum_{i=1}^{N}sim(u,v)(R_{ui}-\overline{R}_v) \tag{2}$$

收稿日期：2011－06－29
作者简介：黄金凤(1971－)，硕士，讲师，研究方向：计算机应用、自动控制.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

其中，P_{uj} 为目标用户 u 对信息资源项 j 的预测值. 算法根据与目标用户相似的 N 个用户的评价进行预测，并非所有用户都参与预测 P_{uj} 值，$sim(u,v)$ 为用户 u 和 v 之间的兴趣相似度，k 为归一化因子.

算法的核心部分是计算用户的兴趣相似度 $sim(u,v)$，相似度量方法主要有三种：余弦相似性、相关相似性及修正的余弦相似性. 鉴于文献[4]的结论，我们选择选用 Pearson 相关相似性度量作为本文的用户相似性度量方式.

设 I 表示用户 u、v 的共同评分项集，则用户 u、v 之间的相似性 $sim(u,v)$ 通过 Pearson 度量，其计算公式：

$$sim(u,v)=\frac{\sum_{i\in I}(R_{u,i}-\overline{R}_u)\cdot(R_{v,i}-\overline{R}_v)}{\sqrt{\sum_{i\in I}(R_{u,i}-\overline{R}_u)^2}\sqrt{\sum_{i\in I}(R_{v,i}-\overline{R}_v)^2}} \tag{3}$$

其中，$\overline{R}_u$、$\overline{R}_v$ 由公式(1)获得.

最后，按 $P_{u,i}$ 值从大到小取前 N 个项目组成推荐集 $I_{rec}=\{i_1,i_2\cdots i_N\}$ 推荐给用户 u，从而完成整个推荐过程.

2 蚁群模糊聚类算法

聚类是基于“物以类聚”的思想，实质是依据项目在属性特征上的相似性对项目进行分类，将相似性高的项目归为一类，而不同类的项目之间的相似性低. FCM 算法[4]是一种基于划分的在模糊集理论上的聚类算法.

蚂蚁觅食过程中，信息传递主要是通过信息素扩散完成的，是以信息素来决定蚂蚁的运动方向. 先前蚂蚁对后面蚂蚁的行为发生影响时，后面蚂蚁一般不在先前蚂蚁的运动轨迹上，而是与该运动轨迹有着或大或小的距离. 当距离比较小时，后面蚂蚁的行为受影响较大，这一特点对数据聚类是十分有用的. 本文借鉴这一原理及文献[5]，提出一种蚁群模糊聚类(基于蚁群算法的模糊聚类)算法.

该算法将数据对象视为具有不同属性的蚂蚁，聚类中心看作是蚂蚁所要寻找的“食物源”，这样，数据聚类过程就可以被看作是蚂蚁寻找食物的过程.

算法的思想是：在蚂蚁从食物源 i 到食物源 j 的过程中，如果找到合适的路径(子解)，它就释放出相应浓度的信息素，该信息素一方面直接影响位于子解的两个聚类中心上的蚂蚁，另一方面它会以该路径为中心向外扩散，影响附近其他蚂蚁的行为，使它们在寻找路径时以更大的概率在下一步选择此路径[6]. 通过这种基于信息素的协作方式，其他数据对象在选择聚类中心时所受的干扰会减小，从而可提高算法的收敛速度. 数据对象的归属根据转移概率的大小来决定；在下一轮循环中，引入聚类偏差的衡量标准，更新聚类中心，计算偏差，再次判断，直到偏差没有变化或在一定误差范围内，算法结束.

算法 2 蚁群模糊聚类算法

令：$X=\{X_i\mid X_i=(X_{i1},X_{i2},\cdots,X_{in}),i=1,2,\cdots,n\}$ 是待聚类的数据集合；

令：$\overline{C}_j$ 为聚类中心，初始值任意分配；

令：$d_{ij}=\|P(X_i-Y_i)\|^2$ 中，d_{ij} 表示 X_i 到 Y_i 之间的加权欧氏距离；P 为权因子，可以根据各分量在聚类中的贡献不同而定.

t 时刻，对于其他数据对象 l，第 k 只蚂蚁从 i 到食物源 j(聚类中心)的路径 (i,j) 上的信息素量 $\tau_{ij}^k(t)$ (如式(4))，此时，蚂蚁将分别以 i 和 j 为中心以 r 为半径向周围扩散信息素，则数据对象 l 由蚂蚁 k 所产生的信息量 $\Delta\tau_{il}^k(t)$，$\Delta\tau_{jl}^k(t)$ 定义为[69]：

$$\tau_{ij}^k(t)=\begin{cases}\dfrac{Q}{\overline{d}_{ij}} & d_{ij}\leqslant R\\ 0 & d_{ij}>R\end{cases} \tag{4}$$

以 i 为中心向周围扩散信息素，数据对象 l 由蚂蚁 k 所产生的信息量 $\Delta\tau_{il}^k(t)$ 为：

$$\Delta\tau_{il}^k(t)=\begin{cases}\gamma\cdot\dfrac{Q}{\overline{d}_{ij}}\left(1-\dfrac{d_{il}\cdot(d_{il})^{\omega}}{\overline{d}^{\omega+1}}\right) & \text{if}\quad d_{il}\leqslant R\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

以 j 为中心向周围扩散信息素，数据对象 l 由蚂蚁 k 所产生的信息量 $\Delta\tau_{jl}^k(t)$ 为：

$$\Delta\tau_{jl}^k(t)=\begin{cases}\gamma\cdot\dfrac{Q}{\overline{d}_{ij}}\left(1-\dfrac{d_{jl}\cdot(d_{jl})^{\omega}}{\overline{d}^{\omega+1}}\right) & \text{if}\quad d_{il}\leqslant R\\ 0 & \text{otherwise}\end{cases} \tag{6}$$

其中，$\overline{d}$ 表示数据对象到聚类中心 $\overline{C}_j$ 之间的欧氏距离；ω 为大于 1 的可调常数；γ 是小于 1 的可调常数(置信水平)；R 为预设的聚类半径；设 $\tau_{ij}(0)=0$.

数据对象 X_i 是否归并到聚类中心 $\overline{C}_j$ 由转移概率 P_{ij} 决定：

$$P_{ij}(t)=\frac{\tau_{ij}^{\alpha}(t)\cdot\eta_{ij}^{\beta}(t)}{\sum_{s\in S}\tau_{sj}^{\alpha}(t)\cdot\eta_{sj}^{\beta}(t)} \tag{7}$$

其中，$S=\{X_s\mid d_{sj}\leqslant R,s=1,2,\cdots,n\}$，表示分布在聚类中心 $\overline{C}_j$ 领域内的数据对象的集合；α 表示残留信息的相对重要程度，β 期望值的相对重要程度，η_{ij} 为 t 时刻蚂蚁由城市 i 选择城市 j 的某种启发信息. 若 $P_{ij}(t)\geqslant P_0$ (P_0 为一设定值)，则 X_i 归并到 $\overline{C}_j$；否则，不归并.

令：$CS_j=\{X_i\mid d_{ij}\leqslant R,i=1,2,\cdots,J\}$，表示所有

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

归并到聚类中心 $\overline{C}_j$ 的数据对象的集合,J 为该集合中数据对象的个数.根据公式(8)和公式(9)分别计算新的聚类中心 $\overline{C}_j$ 和隶属矩阵 u_{jl}.第 j 个聚类的偏离误差 ε_J 及此次分析的总体误差 ε 分别由公式(10)和公式(11)给出.

$$\overline{C_j}=\frac{\sum_{l=1}^{J}u_{jl}^{m}X_i}{J} \tag{8}$$

$$u_{jl}=\frac{1}{\sum_{p=1}^{k}\left(\frac{d_{jl}}{d_{pl}}\right)^{2/(m-1)}} \tag{9}$$

$$\varepsilon_J=\frac{1}{J}\sum_{i=1}^{J}(X_i-\overline{C_j}) \tag{10}$$

$$\varepsilon=\frac{1}{J}\sum_{J=1}^{k}\varepsilon_J \tag{11}$$

算法的伪代码如下:

(1) 初始化　设定 $n,k,r,\varepsilon_0,P_0,\alpha,\beta,\tau_{ij}(0)=0$

(2) 初始化隶属矩阵 U、初始化聚类中心.计算 k 个聚类中心 $c_i,i=1,\cdots,k$.或任取不同的数据赋予 C_j.这一步可以用 FCM 算法获得.

(3)Repeate

(4)取不同于 C_j 且未被标识过的 X_{ij},计算 P_{ij}

(5)if($P_{ij}\geqslant P_0$)

(6)标识 X_i 并归并到 C_j

(7)if(所有数据均被处理)

(8)计算 $\varepsilon_J,\varepsilon_0$

(9)更新 $\tau_{ij},\Delta\tau_{il},\Delta\tau_{jl}$

(10)计算新的聚类中心$\overline{C_j}$及隶属矩阵 u_{jl}

(11) end if

(12) end if

(13)until($\varepsilon_J<\varepsilon_0$)

算法的输出是 K 个聚类中心点向量和 $K*N$ 的一个模糊划分矩阵,这个矩阵表示的是每个样本点属于每个类的隶属度.根据这个划分矩阵按照模糊集合中的最大隶属原则就能够确定每个样本点归为哪个类.聚类中心表示的是每个类的平均特征,可以认为是这个类的代表点.

3　基于蚁群模糊聚类的协同过滤算法(CF－based ACVC)

离线阶段,先对用户进行蚁群模糊聚类,产生若干聚类中心和一个类别隶属矩阵,这个矩阵表示每个用户属于每个类的隶属度.根据这个划分矩阵按照模糊集合中的最大隶属原则就能获得用户与聚类中心的相似性度量矩阵.在线阶段,只要计算目标用户与各个聚类中心的相似性,再通过对比离线时获得的相似性度量搜索矩阵搜索目标用户的最近邻居,并产生推荐.基于聚类的实质,离线阶段获得的聚类数目远远小于用户数目,在线阶段系统只需计算目标用户与少量的聚类中心的相似性,因此,提高了推荐的实时性.

3.1　离线时的蚁群模糊聚类

先用公式(3)代替蚁群模糊聚类算法中的公式(9)的右半部分计算隶属矩阵 μ_{jl},再利用算法 2 获得用户聚类中心 $C(k,n)$ 以及每个用户属于每个类的隶属度矩阵 $U(k,n)$.

用户聚类中心 $C(k,n)$ 表示 n 个项目的 k 个聚类中心,其元素 c_{ij} 表示用户聚类中心 i 对项目 j 的评分,也是第 i 类中的所有用户对项目 j 的平均评分.

基本用户的类别隶属度矩阵 $U(k,n)$ 表示 k 个基本用户的 n 个用户聚类,其元素 u_{ij} 表示用户 i 对聚类中心 j 的相似性,即用户 i 和第 j 个用户聚类中心之间的 Pearson 相关相似性度量.

3.2　在线时查找目标用户的最近邻居并产生推荐

在离线处理结果的基础上,在线阶段,利用算法 1 完成整个推荐过程.

4　仿真实验及分析

通过仿真实验验证本文提出的算法,并文献[2]的算法进行了比较.

4.1　实验环境及实验数据集

实验所用 PC 机的配置为 Intel(R) Core(TM) 2 Duo CPU T9550 2.66GHz、2GB RAM,Windows XP 操作系统,Access 数据库,算法用 Matlab 实现.

实验数据集采用 MovieLens 数据集[8].MovieLens 数据库集是美国 Minnesota 大学 GroupLens 项目组提供的,用于接收用户对电影的评分并提供相应的电影推荐列表.该数据集包含了 943 位用户对 1682 部电影作出的 1000000 条评分数据,分为 5 个 base 数据集和 5 个 test 数据集.我们使用十折交叉验证(10－fold cross－validation)方法进行实验.每次选择一对 base 数据集和 test 数据集,使用 base 数据集中的记录作为基本用户,对 test 数据集中的目标用户进行推荐测试.

4.2　评价标准

实验采用统计精度度量方法中广泛使用的平均绝对误差 MAE(Mean Absolute Error)来衡量算法的预测精度.MAE 是测试集中所有用户对资源打分的实际值与预测值的偏差的绝对值的平均[8].

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

MAE 值越小，说明推荐算法的预测精度越高.

4.3 实验结果及分析

为了验证本文提出算法的有效性，我们进行了仿真实验并与文献[2]的算法进行了对比.

在对 MovieLens 数据集的数据的预处理过程中，我们发现原始类别的前 7 类中，每个原始类别都包含相对较多的基本用户，而其他类则包含相对少量的用户，所以在验证本文算法时，我们取聚类数目 k 的值为 7. 采用最近邻用户数 K 分别取 10、15、20、25、30、35、40，用户相似性度量方法采用 Pearson 相关系数，运行本文提出的算法 CF－based ACVC 和文献[2]的算法(IRP－CF)，计算在不同最近邻用户数时两种算法各自的 MAE. 实验结果如图 1 所示.

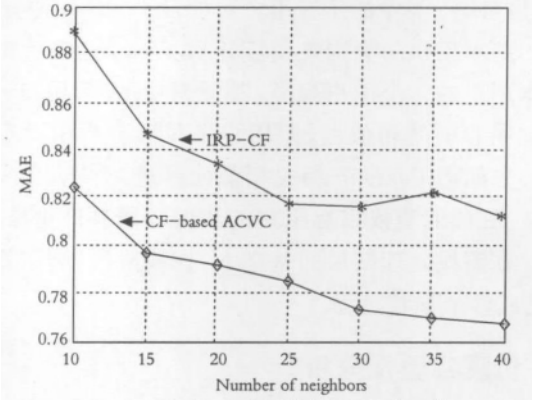

图 1 两种算法的评分预测质量比较

由图 1 可知 CF－based ACVC 具有更小的 MAE

5 结 论

由前文对 CF－based ACVC 算法的分析可知，传统协同过滤算法是直接计算目标用户与所有的 m 个基本用户之间的相似性，而本文算法首先计算目标用户与基本用户聚类中心之间的相似性，基本用户聚类中心的个数相比于所有基本用户是小了很多，所以本文提出的算法提高了在线搜索目标用户的最近邻居的速度，从而在一定程度上提高了推荐系统的实时性. 实验结果表明，本文算法在一定程度上提高了在线时的推荐生成速度，同时推荐质量也有一定的提高. 实验结果还表明对于离线时基本用户的聚类，虽然我们并没有特别要求，比如聚类过程中 k 的选取等，CF-based ACVC 仍然能够在一定程度上加快推荐产生速度.

参 考 文 献

[1] 邓爱林，左子叶，朱扬勇. 基于项目聚类的协同过滤推荐算法[J]. 小型微型计算机系统，2004，25(9)：1665－1670.

[2] B M Sarwar. Sparsity, Scalability, and Distribution in Recommender system[D]. Minneapolis, MN: University of Minnesota, 2001.

[3] H J Ahn. A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-starting Problem[J]. Inforamtion Sciences, 2008, 178(1): 37－51.

[4] KANADE PM, HALL LO. Fuzzy Ants as a Cluste-ring Concept[A]. Proceedings of the 22nd Interna-tional Conference of the North American Fuzzy In-formation Processing Society[C]. USA, 2003: 227－ 232.

[5] 姜长元. 基于混合信息素递减的蚁群算法[J]. 计算机工程与应用，2007，43(32)：62－64.

[6] 马溪骏，潘若愚，杨善林. 基于信息素递减的蚁群算法[J]. 系统仿真学报，2006，18(11)：3297－3300.

[7] http://www.grouplens.org/node/73[EB/OL].

[8] Mobasher B, Jin X, Zhou Y Z. Semantically Enhanced Collaborative Filtering on the Web, Springer-Verlag, 2004: 57－76.

Collaborative Filtering Recommendation Algorithm Based on Ant Colony Vague Clustering

HUANG Jin-feng, LEI Xiao-zhen

(Information Department, Fujian Communicton Technology College, Fuzhou 350007, China)

Abstract: To overcome the difficulty of timely of collaborative filtering algorithm used for generating recommendation, a collaborative filtering recommendation algorithm based on Ant Colony Vague clustering is presented. The algorithm separates the procedure of recommendation into offline and online phases. In the offline phase, the basal users are clustered by Ant Colony Vague clustering technology; while in the online phase, the nearest neighbors of an active user are found according to the basal user clusters, and the recommendation to the active user is produced. The experimental results show that the presented algorithm can improve the performance of CF sysytems in both the recommendation quality and efficiency.

Key words: recommendation algorithm; collaborative filtering; Ant Colony Vague clustering; MAE

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net