

基于蚁群聚类项目评分预测的推荐算法研究

刘贤锋, 刘同存, 李淑明

(桂林电子科技大学 商学院, 广西 桂林 541004)

摘 要: 针对数据稀疏性问题, 提出基于蚁群聚类的项目评分预测方法. 在对 Web 日志分析基础上将用户聚类, 针对目标用户的未评分项目, 找到目标用户的若干最近邻类簇, 利用类簇内其他用户对目标项目的评分预测未评分项目的评分, 从而达到降低数据稀疏性目的. 最后, 结合协同过滤思想设计了相应的推荐算法, 并用从自主开发的旅游电子商务网站上收集的数据进行试验仿真. 实验结果表明, 与其它缓解数据稀疏性的方法相比, 文中的方法显著提高了推荐精度.

关键词: 电子商务; 蚁群聚类; 评分预测; 推荐算法

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2013)01-0131-04

Recommendation Algorithm Based on Item Rating Prediction Using ants Clustering

LIU Xian-feng, LIU Tong-cun, LI Shu-ming

(School of Business, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract: A new recommendation algorithm based on item rating prediction using ants clustering was designed to reduce the sparsity of user rating data. By analysis the web logs of users, ants clustering algorithm was used to cluster users. For the unrated item of target user, a number of nearest neighbor cluster was found, in order to reduce data sparsity, the ratings of other users in the cluster was used to predict the unrated item. Finally, collaborative filtering algorithm was used to provide recommendation services for users. The web server logs of a self-developed e-commerce web site were used for simulation. The result shows that, the recommendation efficiency and accuracy of our method are both obviously superior to other method.

Key words: E-commerce; ants clustering; rating prediction; recommendation algorithm

1 引言

近年来, 电子商务的发展与应用使得互联网商品信息量以指数增长. 为了降低过量商品信息给消费者带来的困扰, 减少消费者的购买成本, 提高消费者对网站的忠诚度, 国内外学者提出了多种信息过滤技术. 其中以协同过滤推荐算法的研究和应用最为广泛^[1].

协同过滤推荐算法是以用户对项目的评分数据为依据, 但随着电子商务系统规模的扩张及用户和项目数的持续增长, 算法的推荐精度受到评分数据

稀疏性的制约, 因为在数据极端稀疏情况下难以精确找到目标用户的最近邻居集^[2]. 为了缓解数据稀疏性, 国内外研究者提出了多种措施^[3-6].

这些方法虽在一定程度上降低了数据稀疏性, 提高了推荐精度, 但仍存在以下不足: ①以项目为主体, 在计算项目相似性基础上构建缺失值填充模型, 忽略了用户的主体地位; ②对未评分项目的评分预测时, 将所有的邻居项目评分平等对待, 忽略了不同用户在不同时间的项目评分对目标项目的影响. 对此, 本文提出基于蚁群聚类的项目评分预测方法, 利用蚁群聚类算法对记录用户浏览行为的 Web 日志

收稿日期: 2012-05-25; **修回日期:** 2012-06-28

基金项目: 国家自然科学基金项目(70862001)

进行分析,将偏好相似的用户聚为一类,针对目标用户的未评分项目,首先找到与目标用户偏好相似的若干类簇,根据类簇内不同用户对项目评分的时间给评分赋予不同权重,在此基础上预测未评分项目的评分,达到降低数据稀疏性目的,最后结合协同过滤思想设计了相应的推荐算法。

2 蚁群聚类基本原理

蚁群聚类源于对蚂蚁尸体堆积和蚁卵分类的研究^[7]。核心思想是将待聚类对象随机分布在二维网格上,用蚂蚁测量当前对象在局部环境内的群体相似度,并通过概率转换函数将相似度转换成拾起、移动或放下的概率,经过多次迭代过程,相似的数据对象便可聚为一类。

设 t 时刻某只蚂蚁在地点 r 处发现数据对象 o_i ,对象 o_i 与其邻域内对象 o_j 的平均相似度表示为:

$$f(o_i) = \max \left\{ 0, \frac{1}{s^2} \times \sum_{o_j \in \text{Neigh}_{s \times s}(r)} \left[1 - \frac{d(o_i, o_j)}{\alpha(1 + (v-1)/v_{\max})} \right] \right\} \quad (1)$$

其中, α 为相似度参数, v 为蚂蚁的移动速度, $v_{\max}$ 为该蚂蚁的最大移动速度, $\text{Neigh}_{s \times s}(r)$ 为地点 r 处以 s 为边长的正方形区域, $d(o_i, o_j)$ 为数据对象 o_i 和 o_j 的空间距离。常用的距离度量方法有欧氏距离和余弦函数法,其中以余弦函数法最为常用,本文采用余弦函数法度量:

$$d(o_i, o_j) = 1 - \text{sim}(o_i, o_j) \quad (2)$$

$\text{sim}(o_i, o_j)$ 为数据对象 o_i 和 o_j 的相似度,用两个对象的属性向量余弦夹角相似度度量,即

$$\text{sim}(o_i, o_j) = \frac{\sum_{k=1}^m (o_{i,k} \cdot o_{j,k})}{\sqrt{\sum_{k=1}^m (o_{i,k})^2 \cdot \sum_{k=1}^m (o_{j,k})^2}} \quad (3)$$

m 为对象的属性数, $o_{i,k}$ 为对象 i 的第 k 个属性值。数据对象越相似,则 $\text{sim}(o_i, o_j)$ 的值越趋近于 1,反之则趋近于 0。

聚类过程中,蚂蚁总是拾起与邻域最不相似的点,并将其放在与邻域节点最为相似的方格中,拾起与丢弃的概率由概率转换函数决定,可由 $f(o_i)$ 的函数得到,即:

$$p_p = \left(\frac{k_1}{k_1 + f(o_i)} \right)^2 \quad (4)$$

$$p_d = \begin{cases} 2f(o_i) & \text{if } f(o_i) < k_2 \\ 1 & \text{if } f(o_i) \geq k_2 \end{cases} \quad (5)$$

其中, p_p 、 p_d 分别为拾起和丢弃概率, k_1 、 k_2 为阈值常量, $f(o_i)$ 越大则 p_p 小, p_d 越大,反之亦然。

3 基于蚁群聚类项目评分预测的推荐算法

3.1 用户聚类生成

为了便于用户聚类,首先使用数据过滤、用户识别、会话识别和路径补充等技术对收集的 Web 日志文件进行预处理,删掉噪声数据,转换为便于进行聚类的规范化数据存储格式。考虑到用户可能在不同时间点对网站及商品多次访问,为了方便计算,将同一用户的多次访问会话进行合并,构造关系型事务数据库。设电子商务网站所有商品构成的集合为 $I = \{i_1, \dots, i_2, \dots, i_n\}$, 则构造的用户对商品的浏览事务数据库为 $T = \{uid, \langle f_1, f_2, \dots, f_i, \dots, f_k \rangle\}$, uid 唯一标识一个用户, f_i 为用户对商品 i 的总浏览频率, k 为该用户所浏览的商品总数。

目前,研究和应用较多的聚类算法有 k -均值聚类算法、矢量空间聚类算法、神经网络和模糊聚类算法等^[7], 由于这些算法的聚类结果过度依赖数据元素的初始排列或输入次序,需手工指定聚类中心及数量,并且聚类过程是不可逆的,因此难以适应复杂群体聚类^[8]。蚁群聚类算法克服了上述缺陷,聚类过程可视化,聚类数量在聚类过程自动生成,并具有较强的抗噪声数据能力,能够实现完全分布式控制,自组织性、可扩展性和健壮性均优于传统的聚类算法。因此,本文使用蚁群聚类算法对用户进行聚类。根据蚁群聚类思想,首先将所有用户随机散列在二维空间网格中,利用公式(1)~(5)在构造的关系数据库 T 上实施用户聚类过程,经过多次迭代便可将相似用户聚为一类,并输出聚类结果和数量,记为 $C = \{c_1, c_2, \dots, c_i, \dots, c_n\}$, n 为聚类个数。

3.2 项目评分预测

为了能够初步预测用户对未评分项目的评分,首先找到与目标用户最为相似的邻居类簇,利用类簇内其他用户对项目的评分初步评估目标用户对未评分项目的评分,以达到降低数据稀疏性目的。具体过程描述如下。

• 根据类簇密度及目标用户与类簇的相似度选择若干个邻居类簇,类簇 c_i 是否被选择由公式(6)给出^[9]。

$$p_i = \frac{\delta_i \cdot \text{sim}(u, c_i)}{\sum_{j=1}^s \delta_j \cdot \text{sim}(u, c_j)} \quad (6)$$

其中, p_i 为类簇 c_i 是否被选择的概率,如果大于阈值 λ 则选择,否则则弃之, s 为类簇总数。 sim

(u, c_i) 为目标用户 u 与类簇 c_i 的相似性, 由用户与聚类中心的距离度量, 即:

$$\text{sim}(u, c_i) = \sqrt{\sum_{r=1}^d |\text{center}_{c_i, r} - u_r|^2} \quad (7)$$

d 表示项目总数, center_{c_i} 为聚类中心, $\text{center}_{c_i, r}$ 为聚类中心用户对项目 r 的评分, u_r 为用户 u 对项目 r 的评分. δ_i 为类簇 i 的密度, 表示为:

$$\delta_i = \frac{N_i}{T_u} \quad (8)$$

式中, N_i 为类簇 c_i 内的用户数, T_u 为电子商务系统中的用户总数.

• 设目标用户 u , 项目空间 I_u 中未评分项目集合为 N_u , 已评分项目集合为 Y_u , 则 $N_u = I_u - Y_u$. 设目标用户的最近邻类簇集合为 $C_{nu} = \{c_1, \dots, c_l, \dots, c_v\}$, v 表示选择的最近邻类簇个数, 并且 $p_{u, c_1} \geq p_{u, c_2} \geq \dots \geq p_{u, c_l} \geq \dots \geq p_{u, c_v}$. 对于目标项目 $i, i \in N_u$, 邻居类簇内对项目 i 评分过的用户集合为 $U = \{u_1, u_2, \dots, u_g, \dots, u_h\}$ 表示, 则评估用户 u 对目标项目 i 的评分方法为:

$$p_{u, i} = \begin{cases} \frac{\sum_{u_d \in U} \text{sim}(u, u_d) \times r_{u_d, i} \times w_{u_d, i}}{\sum_{u_d \in U} \text{sim}(u, u_d)} & \text{if } U \neq \emptyset \\ 0 & \text{if } U = \emptyset \end{cases} \quad (9)$$

其中, $\text{sim}(u, u_d)$ 为用户 u 与类簇内用户 u_d 的相似程度, 可由公式(3)求得, $r_{u_d, i}$ 为用户 u_d 对目标项目 i 的评分, $w_{u_d, i} = \sum_{u_e \in U} t_{u_e, i} / t_{u_d, i}$, 表示用户 u_d 的评分对目标项目 i 的影响权重, $t_{u_d, i}$ 为用户 u_d 对项目的评分时间距当前时间的差, 最新的用户评分具有较大的权重.

• 经过上述处理, 用户—项目评分数据库中的大多数项目均有一个评分值, 即用户 u 对项目 i 的评分可转换为^[10]:

$$R_{u, i} = \begin{cases} r_{u, i} & \text{if user } u \text{ rate item } i \\ p_{u, i} & \text{if user } u \text{ not rate item } i \end{cases} \quad (10)$$

3.3 产生推荐结果

在经项目评分预测之后的数据集上, 结合协同过滤思想为用户提供推荐服务, 具体推荐过程描述如下:

• 利用 Pearson 相关性度量方法计算用户间的偏好相似性, 计算方法如下:

$$\text{sim}(i, j) = \frac{\sum_{k \in I_{ij}} (r_{i, k} - \bar{r}_i)(r_{j, k} - \bar{r}_j)}{\sqrt{\sum_{k \in I_{ij}} (r_{i, k} - \bar{r}_i)^2} \sqrt{\sum_{k \in I_{ij}} (r_{j, k} - \bar{r}_j)^2}} \quad (11)$$

其中, I_{ij} 表示经用户 i 和 j 共同评分的项目集,

$r_{i, k}$ 表示用户 i 对项目 k 的评分, $\bar{r}_i$ 表示用户 i 对所有商品兴趣度的平均值.

• 根据计算的相似性找到目标用户 u 的最近邻居集合 $N_u = \{u_1, u_2, \dots, u_j\}$, 其中 $u \notin N_u$ 且 $\text{sim}(u, u_1) > \text{sim}(u, u_2) > \dots > \text{sim}(u, u_j)$.

• 根据目标用户的最近邻居对项目的评分预测用户未评分项目的值, 并选择分值较高的 N 个项目推荐给目标用户, 计算方法如下:

$$P_{u, i} = \bar{r}_u + \frac{\sum_{k \in N_u} \text{sim}(u, k) \times (r_{k, i} - \bar{r}_k)}{\sum_{k \in N_u} \text{sim}(u, k)} \quad (12)$$

$\bar{r}_u$ 和 $\bar{r}_k$ 表示目标用户 u 和邻居用户 k 对项目的平均评分, $\text{sim}(u, k)$ 表示目标用户 u 与邻居用户 k 对项目评分的相似度, $r_{k, i}$ 表示邻居用户 k 对项目 i 的评分.

4 实验仿真与结果分析

本实验在 PC 机 (Pentium T4400, CPU 2.2Hz, 内存 2GB), Windows XP 操作系统平台上进行, 使用 C# 程序设计语言在 Microsoft Visual Studio 2008 开发环境下编程算法程序.

本文以从自主开发的旅游电子商务网站 (<http://www.youdaike.com>) 上收集的数据为例进行实验仿真. 为了便于实验, 将用户对商品的平均评分作为其对该商品的分值, 最后得到 1135 条用户评分数据, 包含 105 个用户对 132 件商品的评分, 评分数据的稀疏性为 $\epsilon = 1 - 1135 / (105 \times 132) = 0.918$, 可见评分数据极为稀疏.

将评分数据划分为训练集和测试集, 对此, 引入变量 x , 表示训练集占整个数据集的百分比. 比如, $x = 0.8$ 表示数据集中 80% 的评分数据作为训练集, 20% 评分数据作为测试集. 在本实验中, 将 x 设为 0.8.

在推荐算法的推荐质量评价标准中, 最常用的是平均绝对差 (MAE), 它是通过计算预测的用户对项目评分与用户对该项目的真实评分之间的偏差来度量算法预测的准确性, MAE 越小, 预测的越精确, 推荐精度也越高. 假设利用我们的算法预测用户对 k 个项目的评分为 $\{p_1, p_2, \dots, p_k\}$, 用户真实评分为 $\{r_1, r_2, \dots, r_k\}$, 则平均绝对偏差可描述为:

$$\text{MAE} = \frac{\sum_{i=1}^k |p_i - r_i|}{k} \quad (13)$$

本文将实验过程分为用户聚类、评分预测和推荐三个阶段. 鉴于蚁群聚类算法涉及多个参数, 因

此,实验首先逐渐调整不同参数的取值进行用户聚类,以选择合适的参数为后续实验奠定坚实的基础.通过多次实验发现,当 $\alpha=1$ 、 $v=0.5$ 、 $v_{\max}=2$ 、 $s=1$ 、 $k_1=k_2=0.5$ 时,取得较好的聚类效果.根据输出的聚类结果,将阈值 λ 设为0.5,利用本文提出的方法对评分数据集中的缺失值预测,最后在处理后的数据集上实施推荐过程.

为了验证本文提出基于蚁群聚类项目评分预测推荐算法(ACR-Based CF)的有效性,分别与传统的推荐算法(CF)、基于项目评分预测的推荐算法(IR-Based CF)、基于奇异值分解的推荐算法(SVD-Based CF)、进行比较.选取的邻居个数从4逐渐增加到20,间隔为4,实验结果如图1所示.

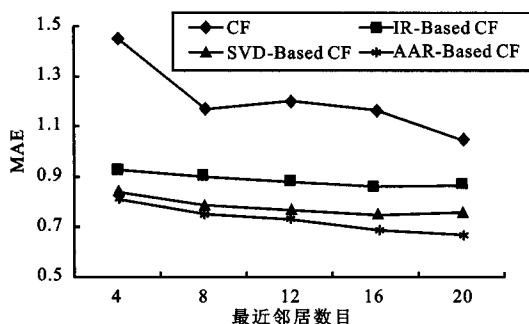


图1 算法推荐精度比较

由图1可知,随着选取最近邻个数的增加,无论是传统的还是文中提出的推荐算法的平均绝对差(MAE)均不同程度的降低,推荐质量不断提高.在选取的最近邻居数相同情况下,本文提出的基于蚁群聚类项目评分预测的推荐算法(ACR-Based CF)的MAE明显小于传统的推荐算法,推荐精度显著提高,说明文中提出的算法优于传统的推荐算法.

5 结束语

本文提出基于蚁群聚类项目评分预测的推荐算法,在对用户访问Web日志分析基础上,利用蚁群聚类算法对用户聚类,从用户相似性角度对未评分项目进行初步预测,最后结合协同过滤思想设计了相应的推荐算法.实验结果表明,本文提出的项目评分预测方法显著提高了商品推荐质量和精度.

本文的研究是从用户相似性角度对缺失值评估,进一步拓展了数据稀疏性研究思路,但仍存在不足之处,主要表现在:①在聚类时将用户的多个访问会话平等对待,忽略了不同会话期对用户相似性度量产生的影响;②仅依据用户对项目的评分时间设置权重,忽略了用户的相似性影响,理论上,与目标

用户越相似且最新的项目评分应具有较大的权重;③实验数据来源网站日访问量不是很高,收集的数据量相对较少,有待在大规模数据集上深入检验本文方法的有效性.针对这些不足之处,作者将继续对其展开深入的研究.

参考文献:

- [1] Kim BM, Li Q, Park CS, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information Systems, 2006, 27(1):79-91.
- [2] Panagiotis S, Alexandros N, Apostolos N, et al. Collaborative recommender systems: Combining effectiveness and efficiency[J]. Expert Systems with Applications, 2008(34):2995-3013.
- [3] Garcia I, Sebastia L, Onaindia E. On the design of individual and group recommender systems for tourism[J]. Expert Systems with Applications, 2011, 38(6):7683-7692.
- [4] Vozalis MG, Margaritis GM. Using SVD and demographic data for the enhancement of generalized Collaborative Filtering[J]. Information Sciences, 2007, 117(15):3017-3037.
- [5] Lee TQ, Park Y, Park YT. A time-based approach to effective recommender systems using implicit feedback[J]. Expert Systems with Applications, 2008, 34(4):3055-3062.
- [6] 张光卫,李德毅,李鹏,等.基于云模型的协同过滤推荐算法[J].软件学报,2007,18(10):2403-2411.
- [7] Yang Y, Kamel MS. An aggregated clustering approach using multi-ant colonies algorithms[J]. Pattern Recognition, 2006, 39(7):1278-1289.
- [8] 徐选华,范永峰.改进的蚁群聚类算法及在多属性大群体决策中的应用[J].系统工程与电子技术,2011,33(2):346-349.
- [9] Shinde SK, Kulkarni U. Hybrid personalized recommender system using centering-bunching based clustering algorithm[J]. Expert Systems with Applications, 2012, 39(1):1381-1387.
- [10] 刘枚莲,刘同存,李小龙.基于用户兴趣特征提取的推荐算法研究[J].计算机应用研究,2011,28(5):1664-1667.

作者简介:

刘贤锋 男,(1973-)硕士,讲师.研究方向为电子商务.

刘同存 男,(1984-),硕士研究生.研究方向为电子商务.

李淑明 女,(1973-),硕士,讲师.研究方向为电子商务.