一种基于改进的层次聚类的协同过滤用户推荐算法研究

张峻玮 杨 洲
(南京理工大学计算机科学与工程学院 南京 210018)

摘 要 为了降低组用户推荐的计算时间,提出了一种改进的层次聚类协同过滤用户推荐算法。由于数据的稀疏性,传统的聚类方法在尝试划分用户群时效果不理想。考虑到传统聚类算法的聚类中心不变组内用户间相关度不高等问题,将用户进行聚类,然后按照分类计算出每个用户的推荐结果,在进行聚类的同时充分利用用户间的信息传递来增强组内用户的信息共享,最后将组内所有的用户的推荐结果进行聚合。最后仿真实验表明,本方法能够有效地提高推荐的准确度,比传统的协同过滤算法具有更高的执行效率。

关键词 推荐系统,协同过滤,层次聚类算法,组推荐,用户推荐
中图法分类号 TP391.03 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.038

Collaborative Filtering Recommendation Algorithm Based on Improved User Clustering

ZHANG Jun-wei YANG Zhou
(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210018, China)

Abstract In order to reduce the computation time of group user recommendation, this paper proposed an improved k-means clustering collaborative filtering recommendation algorithm. Because of the sparsity of data, the effect of the traditional clustering methods is not ideal when trying to divide user group. This paper took into account that invariant group correlation between users in the clustering center of the traditional K-means algorithm is not high, made the user clustering, then according to the classification calculated recommended results of each user in the cluster, made full use of user information transmission between users to enhance information sharing within the group, and polymerized all user recommendation result of the group. Finally, simulation results show that the method proposed in this paper can effectively improve the accuracy of the recommendation, and it is more effective than traditional collaborative filtering algorithm.

Keywords Recommendation systems, Collaborative filtering, K-means algorithm, Group recommended

1 引言

推荐系统能从海量数据中提取用户感兴趣的内容,并得到了广泛应用(例如淘宝,亚马逊)。其中,协同过滤(Collaborative Filtering)算法凭借其简单性和高效性已成为研究的热点[1]。CF算法的原理是通过寻找相似的用户并根据这些用户来推荐商品。聚类的好处是能自动地将用户划分为不同的用户组,而基于用户群的推荐能充分发挥人的社会学特性[2]。此外,基于聚类的推荐算法更适用于为新用户推荐[3]。典型的推荐系统应用网站有Amazon[4]、NetFlix[5]和MovieLens[6]等。协同过滤推荐系统是最常用的推荐系统。当预测用户对产品的偏好时,协同过滤算法通过利用他人对产品的评分值进行评估。协同过滤采用的基本方法是通过某种相似性的度量标准找到与当前用户相似的用户,然后利用这些相似用户对产品的评分值,对待估产品进行评分[7]。

然而在运用推荐算法时,常常会遇到数据稀疏问题。这里的稀疏指的是大部分用户只评论极少部分的商品。比如,数据集movie len的稀疏程度(用户-物品评分矩阵空元素占的百分比)是95.5316%,数据集Epinions的稀疏程度是99.99135%。为解决稀疏问题,学者提出很多高效的算法,大体上可以分为两种:第一种方法是特征的增强,如文献[8-10]使用用户信用度作为特征,并使用pagerank算法来增强它;第二种方法是模型改进,如文献[11]提出一种基于活跃用户的平滑聚类方法。

本文提出了一种改进的层次聚类协同过滤用户推荐算法。由于数据的稀疏性,传统聚类方法在尝试划分用户群时效果不理想,文章考虑到传统的聚类算法的聚类中心不变组内用户间相关度不高等问题,将用户进行聚类,然后按照分类计算出每个用户的推荐结果,在进行聚类的同时充分利用用户间信息传递来增强组内用户的信息共享,最后将组内所有的用户的推荐结果进行聚合。最后仿真实验表明,本文提出的方法能够有效提高推荐的准确度,比传统的协同过滤算法具有更高的执行效率。

2 基于聚类的协同过滤

2.1 协同过滤算法

协同过滤推荐方法的基本思想来源于口碑营销,其应用已有的用户对商品的偏好来预测用户对某商品的喜好程

到稿日期:2014-01-29 返修日期:2014-03-26 本文受国家自然科学基金项目(71272144)资助。
张峻玮(1989—),男,硕士生,主要研究方向为算法设计、推荐系统;杨 洲(1976—),女,博士,副教授,主要研究方向为算法设计。

度[12]。协同过滤推荐方法是一种有效的推荐方法，且已经被众多的电子商务网站所采用，如亚马逊。相对于基于内容的推荐方法，协同过滤推荐方法的优势是准确性高，不需要过高的商品专业化知识，并且还可以充分利用用户之间的关系网络。然而当系统中没有足够的相似项时(冷启动问题)，协同过滤方法的准确性就降低，此时需要基于内容的推荐方法作为补充。

协同过滤主要划分为基于数据(memory-based)和基于模型(model-based)两种。基于数据的系统算法分为基于用户的系统协同算法(user-based)和基于物品的协同算法(item-based)。这里我们以基于用户协同过滤算法为例。算法的输入是一个评分矩阵，用户是行，表物品是列，矩阵的数值代表用户对物品的评分。算法一般可以分为两个部分。第一部分是评价用户间的相似度，常用的相似度度量方法有 Spearman 相关系数、Tanimoto 相关系数、余旋相似度。其中最通用的相似度计算方法是 Pearson 相关系数，见式(1)。这里 X,Y 指的是用户的评分向量。

$$w=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \tag{1}$$

第二部分是根据相似用户对条目打分，见式(2)。

$$p_{a,i}=\overline{r}_a+\frac{\sum_{u=1}^{k}w_{a,u}(r_{u,i}-\overline{r}_u)}{\sum_{u=1}^{k}w_{a,u}} \tag{2}$$

其中，$p_{a,i}$ 代表评分，w 代表权重，r 代表得分。第二部的核心是如何选取最相似用户，也最为灵活。

2.2 基于层次聚类的协同过滤算法

在相似性用户 F_u 的计算过程中，采用简单的皮尔森相关系数、余弦相似性和欧几里得距离等相似性度量标准。本文采用基于聚类的方法来寻找每个用户的相似性用户集合。

本文采用自下而上的层次聚类方法将用户分割为不同的分类。首先将每个用户划分为一个分类。其次，在每一步中将两个最相似的分类合并成一个稍大的分类。本文定义两个组之间的相似性为来自于两个分类中的用户间的相似性的最小值。最后当所有的分类都不能合并时，算法终止。

在对用户进行聚类分析后，应用分类内部的其他用户作为用户的相似性用户，利用协同过滤算法计算出用户对项的预测值，然后将组用户的所有预测项进行合并得到组用户的推荐结果。

通过聚类使得寻找相似用户变换为寻找相似用户群。用户群是社会信息交互与传递的基本要素。一般来说，大的群中，成员间关心的是共同兴趣。而小的群中，成员间多是私人关系[2]。所以与非聚类的协同算法相比，聚类方法会减少推荐的个性，但能为用户推荐更多的物品。此外还能减少恶意攻击。如果用户 a 的历史数据过少，则聚类算法具有更好的推荐结果。基于聚类的协同过滤算法，预测公式调整为：

$$p_{a,i}=\overline{r}_a+\frac{\sum_{u=1}^{k}w_{a,u}(r_{u,i}-\overline{r}_u)}{\sum_{u=1}^{k}\|w_{a,u}\|} \tag{3}$$

其中，$w_{a,u}$ 既可表示用户间相似度，也可表示图信息。本文使用的是用户相似度。与原公式不同的是，分母是 w 的绝对值，其原因是物品间的相似度会小于 0。

3 改进的层次聚类的协同算法

由于数据的稀疏性，将传统的层次聚类算法运用到推荐系统上存在很多不足，如表 1 所列。

表 1 聚类前后中心对比聚类

初始	149	915	901	457	752
结束	149	915	901	457	752

从表 1 中可以看出，由于数据的稀疏性，往往只有类中心与类中元素可以计算出相似度。而传统的距离是类欧氏距离的，则不会出现这些情况。同时传统的聚类算法都有较强的通用性，这也就意味着缺少相关的领域知识。比如，在推荐系统中存在单用户从属问题：一个用户有多个兴趣，从而从属于多个集合。所以本文在此基础上，改进了思路。通过利用用户与类中心距离与类成员的平均距离，共同判断一个用户是否从属于一个群组。通过遍历簇中所有元素可以减少对类中心用户的依赖性。

给定用户 $u\in U$ 及所属分类 F_u，通过式(2)计算出用户 u 对项 i 的预测值$(i,\hat{r}_{u,i})$，并将所有的结果保存在集合 V_u 中，$V_u=\{(i,\hat{r}_{u,i})\mid r_{u,i}=\text{null for all } u\in G\}$。此后，按照$\hat{r}_{u,i}$的大小对 V_u 进行排序，使得$(i,\hat{r}_{u,i})$值大的元素在前，值小的在后。

对于一组用户 $G=\{u_1,\cdots,u_n\}$，以及相应的 $V_{u_1},\cdots,V_{u_n}$，我们对所有 G 中用户都没有评价的项的预测值应用 $Aggr$ 进行聚合。在对聚合结果进行排序后，返回排名靠前的 k 个项及其预测评分值。改进后的算法如下所示。

输入：聚合函数 Aggr，组用户 $G=\{u_1,\cdots,u_n\}$，其中每个用户 $u\in G$ 带有一个对未知项的评分集合 $V_u=\{(i,\hat{r}_{u,i})\mid r_{u,i}=\text{null for all } u\in G\}$；

输出：k 个二元组$(i,\hat{r}_{G,i})$；

1. For each $i\in I$ 并且 $r_{u,i}=$null for all $u\in G$ do
2. 　从 $V_{u_1},\cdots,V_{u_n}$ 中提取出 $\hat{r}_{u_1,i},\cdots,\hat{r}_{u_n,i}$；
3. 　$\hat{r}_{G,i}=\text{Aggr}(\hat{r}_{u_1,i},\cdots,\hat{r}_{u_n,i})$；
4. End for；
5. scoreK＝对 $\hat{r}_{G,i}$ 进行排序；
6. Return scoreK 的前 k 项；

在现实生活中，一个人可能喜欢足球和书法。针对这种情况，我们的做法是在聚类时允许一个用户被划分到不同的群组中。根据 Pearson 相似度的定义，当数值大于 0.6 时为强相关，大于 0.4 为至少中等相关。因此我们假设一个用户被划分到多个群组的条件是：与多个类中心的聚类大于 0.6，或者与多个群组间有至少 m 个中等相关的元素。

4 实验结果与分析

4.1 数据来源

我们采用的数据集是明尼苏达大学小组收集的 movielens(http://www.cs.umn.edu/Research/GroupLens/)。数据包括 100000 的评分，共有 943 个用户与 1682 个物品。评分的范围是 1－5，其中 5 表示最喜欢。如果用户没有给物品评分，则对应的评分记为 0，其中每个用户至少评价了 20 个物品。这里，我们将数据集的 80%作为训练集，20%作为测试集，并重复实验 5 次。

4.2 评判标准

我们采用最小均误差(Mean Absolute Error)作为评判标准。在推荐系统领域,MAE是最早的也是最广泛的评判推荐好坏的方法。MAE比较预测结果和用户真实评价的离差。MAE的公式如下:

$$MAE=\frac{\sum_{u\in T}|P_u(t)-\tilde{P}_u(t)|}{|T|} \tag{4}$$

其中,$P_u(t)$ 是用户对商品的预测打分,$\tilde{P}_u(t)$是用户对商品的实际打分,T代表测试集,$|T|$代表测试集的大小,数值越低代表准确度越高。其中标准cf的mae的范围是0.82～0.87。

4.3 实验结果分析

本文的目的是改进传统聚类算法并使其更适用于推荐系统领域。考虑到聚类的随机性,本文重复进行聚类算法,并取10次聚类结果的平均值。

图1是本文算法与传统的K-Means算法进行的对比结果。k代表聚类的个数,纵坐标是mse的数值。从图中可知,改进的算法从整体上性能优于K-Means算法。K-Means算法普遍低于标准cf,而改进的算法优于标准cf的baseline。

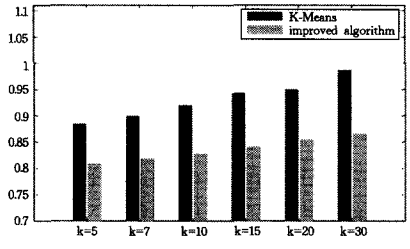

图1 本算法与传统K-Means算法比较

由于聚类算法能大大地提高推荐算法的执行效率,因此它必定要损失推荐的准确性。为了更进一步说明改进的算法在准确性上的优势,在选择算法的聚合函数时,我们假设Min为最小值,Max为最大值,Mean为平均值。分别观察算法在上述3种聚合函数下随着 ω_1 的改变误差的变化情况,实验结果如图2所示。

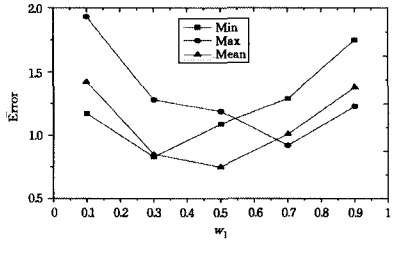

图2 参数 w_1 对结果的影响

从图中可以看出,在3种聚合函数下,算法的误差都随着 ω_1 的增加先降低后增加。这说明基于邻居的评分预测和基于用户支持率的评分预测都对评分预测的结果有影响。此外在Min聚合函数下,算法在 $\omega_1=0.3$ 处取最小值,这说明了本文算法能有效地进行推荐。

因此,从所有的实验结果来看,本文提出的方法能够有效提高推荐的准确度,比传统的协同过滤算法具有更高的执行效率。

结束语 推荐系统可以有效地解决信息过载问题。在推荐系统中,组推荐可以为一组用户推荐旅游的目的地、就餐的餐馆以及共同喜爱的电影等。然而,如果组内的用户过多,推荐算法将消耗很长的计算时间,不能及时地给出推荐结果。为了提高组推荐算法的性能,本文提出了一种有效的基于用户聚类的组用户推荐算法。首先,将用户进行聚类,然后按照分类计算出每个用户的推荐结果,最后将组内所有用户的推荐结果进行聚合。实验表明,本文提出的基于用户聚类的组用户推荐算法的执行效率很高,明显优于传统的协同过滤推荐算法。

参 考 文 献

[1] Sarwar B, Karypis G, Konstan J. Analysis of recommendation algorithms for e-commerce[C]// Proceedings of the 2nd ACM conference on Electronic commerce. ACM Press, 2000: 158-167

[2] Pham M C, Cao Y, Klamma R. A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis[J]. Journal of Universal Computer Science, 2011, 17(4): 583-604

[3] Harper F M, Sen S, Frankowski D. Supporting social recommendations with activity-balanced clustering. [C]// Proceedings of the ACM Recommender System conference. ACM, 2007: 165-168

[4] Massa P, Avesani P. Trust-aware Collaborative Filtering for Recommender Systems[C]// Proceedings of Federated International Conference on Move to Meaningful Internet. Springer, 2004: 492-508

[5] Massa P, Avesani P. Trust-aware Recommender Systems[C]// Proceedings of the 2007 ACM Conference on Recommender systems. ACM, 2007: 17-24

[6] Chowdhury M, Thomo A. Trust-Based Infinitesimals for Enhanced Collaborative Filtering[C]// Proceedings of the 15th International Conference on Management of Data. Computer Society of India, 2009

[7] Sun D, Zhou T, Liu J. Information filtering based on transferring similarity[J]. Physical Review E, 2009, 80(1): 173-177

[8] Gurrin, He C A, Kazai Y A. A Performance Prediction Approach to Enhance Collaborative Filtering Performance[C]// Proceedings of European Conference on Information Retrieval. Springer, 2010: 382-393

[9] Breese J S, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering[C]// Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. ACM, 1998: 43-52

[10] McFee B, Barrington L, Lanckriet G. Learning Similarity from Collaborative Filters[C]// Proceedings of the International Society of Music Information Retrieval Conference. ACM, 2010: 345-350

[11] Herlocker J L, Konstan J A, Borchers A. An algorithmic framework for performing collaborative filtering[C]// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999: 230-237

[12] 黄创光,印鉴,汪静,等. 不确定近邻的协同过滤推荐算法[J]. 计算机学报,2010,33(8):1369-1377

[13] 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(2):350-362

[14] 刘建国,周涛,郭强,等. 个性化推荐系统评价方法综述 [J]. 复杂系统与复杂性科学,2009,6(3):1-10

[15] 傅鹤岗,王竹伟. 对基于项目的协同过滤推荐系统的改进[J]. 重庆理工大学学报:自然科学版,2010,24(9):69-74