基于改进相似度的协同过滤算法研究

李 容[1] 李明奇[1] 郭文强[2]

(电子科技大学数学科学学院 成都 611731)[1] (新疆财经大学计算机科学与工程学院 乌鲁木齐 830012)[2]

摘 要 协同过滤利用邻居用户的偏好对目标用户的偏好进行推荐预测,相似度计算是其关键。传统的相似度计算忽略了用户共同评分项目数与用户平均评分的影响,以至于在数据稀疏时不能很好地度量用户间的相似度。提出了两个修正因子来改进传统相似度,同时改进了协同过滤算法,将其应用于电影推荐。仿真结果表明,在电影推荐中,基于改进后相似度计算的协同过滤算法能取得比传统算法更低的MAE值,提高了电影推荐质量。

关键词 协同过滤,Pearson相似度,共同评分项目,电影推荐

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.037

Research on Collaborative Filtering Algorithm with Improved Similarity

LI Rong[1] LI Ming-qi[1] GUO Wen-qiang[2]

(School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China)[1]

(School of Computer Science and Engineering, Xinjiang University of Finance & Economics, Urumqi 830012, China)[2]

Abstract Collaborative filtering recommends and predicts the target user's preferences by using his neighbor user's preference. The calculation of similarity is the key. Traditional similarity calculation ignores the affection from the co-rated item number rated by common users, and their average similarity rating. That causes poor similarity description among users in case of data sparse. In this paper, we proposed two factors to improve the traditional similarity calculation. Meanwhile, the collaborative filtering algorithm was improved with the improved similarity and it is applied to film recommendation. Simulation results show that the improved collaborative filtering algorithm based on the improved similarity can get a lower MAE value than the traditional method, which is helpful to improve the quality of movie recommendation.

Keywords Collaborative filtering, Pearson similarity, Co-rated item, Movie recommendation

1 前言

随着互联网和信息技术的迅速发展,影视行业也得到充分的发挥空间,越来越多的电影丰富着我们的生活。然而,面对数不胜数的电影,人们寻找自己感兴趣的电影的能力却没有得到相应的提高。有些用户选择直接去电影院观看正在热映的电影,而有些用户则直接在电影资源中选择评分较高的电影,但是这样的电影往往不能符合用户的观影兴趣。因此,如何帮助用户发现其真正感兴趣的电影显得非常重要。

为了在用户没有提供明确需求的情况下,主动给用户提供能够满足他们兴趣和需求的信息,推荐系统应用而生,且被视为解决信息过载这一问题的重要的有效的方法之一。推荐系统是根据用户的信息需求、历史的行为数据等,采用一定的推荐算法为目标用户产生推荐。目前已有很多有效可靠的推荐算法,包括基于内容的推荐、基于协同过滤的推荐、基于关联规则的推荐等。其中基于协同过滤的推荐是现如今应用最广泛的一种,最早应用协同过滤技术的是Tapestry推荐系统[1],这个系统主要是解决Xerox公司在PARC的研究中信息过载的问题。

与传统基于内容过滤、直接分析内容进行推荐不同,协同过滤分析用户兴趣,在用户群中找到与目标用户的相似用户,利用这些相似用户对某一商品的评价,形成对指定用户的推荐预测。该算法主要分为基于项目(item)的协同过滤和基于用户(user)的协同过滤,统称基于邻域关系的协同过滤模型。两者的区别在于基于item相似度和基于user相似度做出推荐。本文主要研究基于user相似度的改进推荐算法。

2006年的Netflix大赛是视频推荐邻域的标志性事件。这场比赛将协同过滤推荐、关联规则、奇异值分解等众多推荐算法应用于视频推荐邻域,获得了非常好的推荐效果。同时YouTube等公司也在视频推荐邻域开始专门的研究[11],可见推荐系统在视频网站的重要性。本文将在现有协同过滤推荐算法的基础上,对相似度计算进行改进,提出用户共同评分项目和用户平均评分修正因子,将其融合到传统相似度计算方法中,得到基于改进相似度的协同过滤算法,并将其应用于电影推荐场景中。

2 相似度的协同过滤算法

2.1 传统协同过滤算法

在协同过滤算法中,为了对目标用户产生推荐,需要利用用户评分矩阵搜索目标用户的邻居,利用邻居用户的兴趣偏

到稿日期:2015-10-13 返修日期:2016-01-25 本文受国家自然科学基金(61163066)资助。

李 容(1990—),女,硕士,主要研究方向为数据挖掘及应用,E-mail:skylir@163.com;李明奇(1970—),男,博士,副教授,主要研究方向为信息论及应用;郭文强(1975—),男,博士,教授,主要研究方向为计算机通信、信号处理。

好产生评分，主要有3个步骤：相似度计算、邻居用户的选择和评分预测[2]。应用到电影场景中，即根据影片的用户评分的相似度来进行推荐，其中相似度计算是基于user协同过滤算法的核心部分。常见的相似度计算方法有余弦(COS)相似度、Pearson相似度、修正余弦(ACOS)相似度等，这些相似度计算方法主要是基于向量的，即将一个用户对所有项目的评分作为一个向量来计算用户之间的相似度；这些相似度计算方法在初期都能准确计算用户之间的相似度，但随着电子商业的迅速发展，用户和项目的数目都急剧增加，导致用户-项目的评分矩阵极度稀疏，上述相似度计算方法的效果随之减低。因此，近年来有很多研究致力于提高稀疏数据下相似度计算的准确率。John S. Bree等人提出了通过惩罚热门物品[3]来提高用户间的相似度；MS Shang提出了一种基于图的相似度计算方法[4]。这些算法在一定程度上改善了相似度的计算，但依然存在一定的缺陷。

在传统相似度计算方法中，Pearson相似度由于其容易理解且便于计算而被广泛使用，具体表达式如式(1)所示：

$$sim(u,v)=\frac{\sum_{i\in I_{uv}}(r_{u,i}-\bar{r}_u)(r_{v,i}-\bar{r}_v)}{\sqrt{\sum_{i\in I_{uv}}(r_{u,i}-\bar{r}_u)^2}\times\sqrt{\sum_{i\in I_{uv}}(r_{v,i}-\bar{r}_v)^2}} \quad (1)$$

其中，$r_{u,i}$表示用户u对项目i的评分，$\bar{r}_u$表示用户u对所有评价过项目的平均评分，I_{uv}表示用户u和用户v共同评分项目集合。

Pearson相似度一般用于计算两个定距变量间联系的紧密程度，它的取值在[−1,1]之间，值为1则表明两个用户对每一个项目均有着完全一致的评价。

2.2 传统相似度计算方法的缺陷

随着行业的发展，用户数和项目数呈指数级增长，这导致用户评分矩阵变得极度稀疏。上述的相似度计算方法虽然已经得到广泛应用，但在稀疏数据场景下依然很难得到真正的最近邻居集。主要的原因总结为以下两点。

1)用户共同评分项目数对相似度的影响

分析传统相似度计算可以发现，两个用户都评分过的项目，即共同评分项目尽管极少，可能只占了用户项目的1%[5]，但却具有极高的相似度。如表1所列的用户评分矩阵利用传统Pearson相似度计算用户u_1与用户u_2和用户u_3的相似度。直观来看，用户u_1和用户u_2仅有2个共同评分项目，而用户u_1和用户u_3有4个共同评分项目，且评分都相近，则$sim(u_1,u_3)$应大于$sim(u_1,u_2)$。然而用户u_1与用户u_3的Pearson相似度仅为0.9683，用户u_1和用户u_2的相似度为1，显然不合理。

表1 用户评分矩阵

	I_1	I_2	I_3	I_4	I_5
u_1	5	3	0	2	1
u_2	5	0	0	3	0
u_3	4	3	0	2	2
u_4	1	2	1	2	1
u_5	4	5	4	5	4

2)平均评分

传统的相似度计算，包括Pearson和ACOS都是计算两个用户评分向量的线性相关性，而忽略了每个维度上的数值差异，即用户的评分标准不一样。这会导致针对每个项目的具体评分差异会出现这样一种情况：如表1所列，用户u_4和用户u_5的评分向量分别为(1,2,1,2,1)和(4,5,4,5,4)，利用Pearson相似度得出的结果是$sim(u_4,u_5)=1$。但从评分上看，用户u_4对这两个项目都不是很喜欢，而用户u_5比较喜欢。传统相似度计算对数值的不敏感导致了相似度的不准确，需要修正这种不合理性。

以上的问题随着评分矩阵的扩大变得越来越严重，使得传统相似性计算方法不能有效地度量用户间的相似性，降低了推荐质量。

3 基于改进相似度的协同过滤算法

传统相似度计算方法在计算用户相似度时只考虑了共同评分项的评分值，使得推荐系统的推荐质量在数据极度稀疏的情况下不高。通过对传统相似度计算方法的分析，本文的改进相似度计算方法主要基于以下几点：

1)改进相似度需要考虑共同评分项目数对用户相似度的影响，本文利用共同评分项目占用比来度量该影响因子。

2)改进相似度需要考虑两用户平均评分对相似度的影响，本文利用平均评分因子来衡量该影响因子。

3)应用改进相似度计算方法后的电影推荐的平均绝对误差(MAE)应该比传统方法更低。

3.1 共同评分项目数

在用户评分矩阵中，假设两个用户u_1,u_2的偏好方向相似，u_1和u_2的评分项目数分别为N_1和N_2，两者的共同评分项目数为n，则n/N_1和n/N_2都应该比较大。因此利用两者共同评分项目占两者评分项目的比便可有效地改进传统计算方法的缺陷。而对于两个用户，当$N_1\geqslant N_2$时，我们更关心的是共同评分项目数n与N_1的比例。综上所述，本文引入比例修正因子共同评分项目占用比$R(u,v)$，如式(2)所示：

$$R(u,v)=\frac{n}{\max(N_u,N_v)} \quad (2)$$

其中，n表示用户u和v的共同评分项目数，N_u，N_V分别表示用户u和v的评分项目数，R越大，则u和v的整体相似度越高，R越小，则u和v的整体相似度越低。

3.2 平均评分

针对稀疏性数据下相似度计算不够准确的问题，有许多改进相似度的方法，J. Herlocker等人提出设置一定的阈值[2]来修正用户共同评分项目数的问题。例如，当阈值为50时，若两个用户的共同评分数n小于50，则将原来相似度乘以修正因子$n/50$。如果n大于50，则保持原来的相似度，但是这种方法只能修正n较小情况下的相似度。Tanimoto系数考虑了修正用户共同评分数问题，但是该系数针对的是评分值为0和1的情况，且没有考虑用户平均评分的问题。如在式(2)中，R虽然修正了传统相似度算法没有考虑共同评分项目数的缺陷，但是依然没有考虑到用户对每个具体项目的评分差异。因此，本文提出针对用户评分差异的修正因子，引入距离$d(u,v)$用以衡量用户u和v的平均评分差异。计算方法如式(3)所示：

$$d(u,v)=\frac{1}{n}\sum_{i\in I_{uv}}|r_{u,i}-r_{v,i}| \quad (3)$$

其中，$r_{u,i}$表示用户 u 对项目 i 的评分，I_{uv}表示用户 u 和 v 的共同评分项目，n 表示用户 u 和 v 的共同评分项目个数。$d(u,v)$越大，说明两用户的平均评分差异越大，则整体相似度应越低。则修正平均评分因子 $p(u,v)$如式(4)所示：

$$p(u,v)=\frac{1}{1+d(u,v)} \tag{4}$$

其中，p 越大则用户 u 和 v 的整体相似度越高，p 越小则用户 u 和 v 的整体相似度越低。

本文提出的修正因子 $R(u,v)$和 $p(u,v)$综合考虑了 Herlocker，Tanimoto 的相似度计算修正问题，能更全面地修正相似度计算的问题。

3.3 基于改进相似度的协同过滤算法

传统协同过滤算法相似度计算方法是基于用户间共同评分项目的，在评分数据充足的情况下，传统的相似度计算方法能够很好地度量用户间的相似性；而随着影视行业的发展，评分数据极度稀疏的情况下，传统的相似度计算方法很难准确地度量用户间的相似性，使得推荐算法的准确率降低。本文提出的修正因子 $R(u,v)$和 $p(u,v)$全面考虑了共同评分项目数的比例和用户平均评分差异对相似度计算的影响，因此可以有效地缓解数据极度稀疏的情况下传统相似度计算方法不够准确的问题，得到改进的相似度计算方法 $NSim(u,v)$如式(5)所示：

$$NSim(u,v)= sim(u,v)\times R(u,v)\times p(u,v) \tag{5}$$

其中，$sim(u,v)$为传统相似度计算方法。

将改进的相似度计算方法融入到传统的基于协同过滤的项目推荐中，得到基于改进相似度的协同过滤推荐算法。算法具体步骤如下。

1)利用式(5)计算用户间的相似度。

2)利用步骤 1)的结果选择目标用户的邻居用户集，用于预测。

3)利用邻居用户集预测目标用户对项目的评分，计算方法如式(6)所示：

$$p_{u,i}=\bar{r}_u+\frac{\sum\limits_{v\in N_u} sim(u,v)\times|r_{v,i}-\bar{r}_v|}{\sum\limits_{v\in N_u} sim(u,v)} \tag{6}$$

其中，$p_{u,i}$表示用户 u 对项目 i 的预测评分，$\bar{r}_u$ 表示用户 u 的平均评分，N_u 表示用户 u 的邻居用户集，$sim(u,v)$表示用户 u 和 v 的相似度，$r_{v,i}$表示用户 v 对项目 i 的评分。

4)最后根据评分大小产生推荐。

4 实验结果及分析

4.1 数据集

本文采用 GroupLens 提供的 MovieLens 数据集来检验算法。MovieLens 数据集有 3 个不同版本，本文选用最小的数据集。该数据集包含 943 名用户对 1682 部电影的 100000 条评分，评分范围为 1－5，每个用户至少评论过 20 部电影。数据集提供了 5 对训练集和测试集，训练集和测试集分别从 u1. base、u1. test 到 u5. base、u5. test，每对训练集和测试集是按照全集的 80%和 20%来划分的。数据集属性包括用户 Id、电影 Id、评分和评分时间。

4.2 评价标准

本文采用平均绝对误差(MAE)来计算预测误差，对于测试集中的一个用户 u 和项目 i，令 $r_{u,i}$表示用户 u 对项目 i 的实际评分，而$\hat{r}_{u,i}$是推荐算法给出的预测评分，则 MAE 计算公式如式(7)所示：

$$MAE=\frac{1}{n}\sum_{i=0}^{n}|r_{u,i}-\hat{r}_{u,i}| \tag{7}$$

其中，n 是算法预测出的评分集合的大小。MAE 是通过计算用户预测评分与真实评分的误差来衡量算法预测的准确性，故 MAE 值越小，推荐结果的准确率越高，推荐质量越好。

4.3 仿真结果对比

本文将对电影数据集的 5 对训练集和数据集分别进行算法验证，将传统协同过滤算法与基于改进相似度的协同过滤算法在电影推荐中产生的结果进行对比。

图 1 为利用数据集 u1. base 进行训练，利用 u1. test 进行测试的实验结果。横坐标为邻居用户数的大小，此处原点横坐标值为 5，纵坐标为 MAE 值。如图 1 所示，在不同邻居用户数的情况下，采用本文的相似度计算方法，MAE 值明显小于 Pearson，COS，Tanimoto 的相似度计算。其中改进算法的 MAE 值较传统 Pearson 算法降低最为明显，随着邻居用户数的减少，降低程度加大。其中当邻居用户数为 5 时，MAE 值降低程度最大，降低了 4.6%。

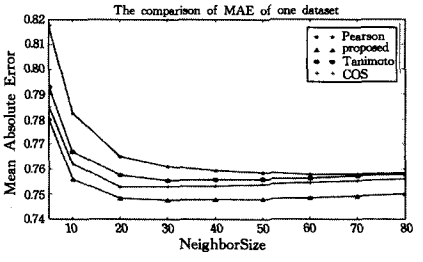

图 1 同一个数据集 MAE 值对比

图 2 为对 5 对训练集和测试集分别进行仿真的结果，横坐标为数据集的编号，纵坐标为数据集的平均 MAE 值。如图 2 所示，改进算法具有良好的稳定性，对不同数据集的 MAE 降低程度相当，平均降低了 1.9%。

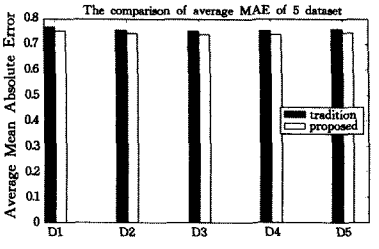

图 2 不同数据集平均 MAE 值对比

从仿真结果中可以看出，通过 $R(u,v)$与 $p(u,v)$对传统相似度进行改进，得到改进后的协同过滤算法，并将其应用到电影推荐中，所计算出的 MAE 值相对于传统算法都普遍较小，确实在推荐精度和质量上有一定的改进。

结束语 协同过滤算法通过相似度计算寻找邻居用户，再利用邻居用户对目标项目的兴趣度来进行推荐。本文通过分析传统相似度计算方法的不足，提出考虑用户共同评分项

(下转第 240 页)

[2] Park J, Kim B I. The school bus routing problem: A review [J]. European Journal of Operational Research, 2010, 202(2): 311-319

[3] Dang Lan-xue, Chen Xiao-pan, Kong Yun-feng. Review of School Bus Routing Problem: Concept, Model and Optimization Algorithms [J]. Journal of Henan University (Natural Science), 2013, 43(6): 682-691(in Chinese)
党兰学，陈小潘，孔云峰. 校车路径问题模型及算法研究进展[J]. 河南大学学报(自然科学版)，2013，43(6)：682-691

[4] Zhang Fu, Zhu Tai-ying. Optimization Design for the School Bus Stations and Routing [J]. Mathematics in Practice and Theory, 2012, 42(4): 141-146(in Chinese)
张富，朱泰英. 校车站点及线路的优化设计[J]. 数学的实践与认识，2012，42(4)：141-146

[5] Kinable J, Spieksma F C R, Berghe V G. School bus routing-a column generation approach [J]. International Transactions in Operational Research, 2014(21): 453-478

[6] Xu Wen-long, Li Xiao-juan, Gong Hui-li, et al. An algorithm for school bus optimal path planning [J]. Geospatial Information, 2011, 9(4): 67-71(in Chinese)
许文龙，李小娟，宫辉力，等. 校车最优路径规划算法[J]. 地理空间信息，2011，9(4)：67-71

[7] Dang Lan-xue, Hou Yan-e, Kong Yun-feng. Spatio temporal Neighborhood Search for Solving Mixed-load School Bus Routing Problem[J]. Computer Science, 2015, 42(4): 221-225(in Chinese)
党兰学，侯彦娥，孔云峰. 时空相关的混载校车路径问题邻域搜索[J]. 计算机科学，2015，42(4)：221-225

[8] Kusuma S, Anan M, Gerrit K J, et al. Heterogeneous VRP Review and Conceptual Framework [J]. Lecture Notes in Engineering and Computer Science, 2014, 2210(1): 1052-1059

[9] Ripplinger D. Rural school vehicle routing problem [J]. Transportation Research Record: Journal of the Transportation Research Board, 2005, 1992: 105-110

[10] Ke X. School bus selection, routing and Scheduling[D]. Canada, Windsor: University of Windsor, 2005

[11] Thangiah S R, Fergany A, Wilson B, et al. School Bus Routing in Rural School Districts[C]// Proceedings of the 9th International Conference on Computer-Aided Scheduling of Public Transport. Spring, 2008: 209-232

[12] De Souza L, Siqueira P H. Heuristic Methods Applied to the Optimization School Bus Transportation Routes-A Real Case[C]// 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems. 2010, 6093: 247-256

[13] Park J, Tae H, Kim B I. A Post-improvement Procedure for the Mixed Load School Bus Routing Problem [J]. European Journal of Operational Research, 2012, 217(1): 204-213

[14] Vidal T, Crainic T G, Gendreau M, et al. Heuristics for multi-attribute vehicle routing problems: A survey [J]. European Journal of Operational Research, 2013, 231(1): 1-21

[15] Penna P H V, Subramanian A, Ochi L S. An iterated local search heuristic for the heterogeneous fleet vehicle routing problem [J]. Journal of Heuristics, 2013, 19(2): 201-232

[16] Hansen P, Mladenovic N. Variable Neighborhood Search: Principles and Applications [J]. European Journal of Operations Research, 2001, 130(3): 449-467

[17] Dang Lan-xue, Wang Zhen, Liu Qing-shong, et al. Heuristic Algorithm for Solving Mixed Load School Bus Routing Problem [J]. Computer Science, 2013, 40(7), 248-253(in Chinese)
党兰学，王震，刘青松，等. 一种求解混载校车路径的启发式算法[J]. 计算机科学，2013，40(7)：248-253

(上接第 208 页)

目占用比和平均评分因子的改进相似度计算方法，得到一种新的协同过滤算法。仿真结果表明，该算法有效地缓解了传统相似度计算引起的推荐结果不准确的问题；将该算法应用到电影推荐系统中，提高了电影推荐质量。

参 考 文 献

[1] Goldberg D, Nichols D, Oki B M, et al. Using Collaborative Filtering to Weave an Information Tapestry[J]. Communications of the ACM, 1992, 35(12): 61-70

[2] Herlocker J, Konstan J A, RIED J. An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms[J]. Information Retrieval Journal, 2002, 5(4): 287-310

[3] 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. 1998: 43-52

[4] Shang M S, Zhang Z K, Zhou T, et al. Collaborative filtering with diffusion-based similarity on tripartite graphs[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(6): 1259-1264

[5] Resnick P, Iacovou N, Suchak M. Grouplens: an open architecture for collaborative filtering of net news[C]// Proceedings of ACM CSCW 94 Conference on Computer-Supported Cooperative Work. 1994: 175-186

[6] Wang J, de Vries A P, Reinders M J. Unifying user-based and item-based collarborative filtering approaches by similarity fusion[C]// Proceeding of SIGIR. 2006

[7] Li Hua, Wang Gen-long, Gao Min. A novel similarity calculation for collarborative filtering[C]// Proceeding of the 2013 International Conference on Wavelet Analysis and Pattern Recongnition. 2013: 14-17

[8] Wang Wei-jie, Yang Jing, He Liang. An improved Collaborative Filtering based on item similarity odified and common ratings [C]// International Conference on Cyberworlds. 2012: 213-235

[9] Lee H C, Lee Y J. A study on the improved collaborative filtering algorithm for recommender system[C]// The 5th Iternational Conf. on Software Engineering Research, Management and Applications. 2007: 297-304

[10] Zhu Xu-zhen, Tian Hui, Cai Shi-min. Personalized recommendation with corrected similarity[J]. Journal of Statistical Mechanics Theory & Experiment, 2014, 2014(7)

[11] Sun Hai-feng, Gan Ming-xin, Liu Xin, et al. Review on dominating websites for movie recommender systems [J]. Journal of Computer Application, 2013, 33(S2): 119-124(in Chinese)
孙海峰，甘明鑫，刘鑫，等. 国外电影推荐系统网站研究与评述[J]. 计算机应用，2013，33(S2)：119-124