结合类别偏好的协同过滤推荐算法

张紫嫣[1] 周 驰[2]

[1](同济大学医学院 上海 200092)

[2](浙江大学软件学院 浙江 杭州 310027)

摘 要 推荐系统是针对如今信息过载现象的一种极为有效的方法,而协同过滤算法自提出以来就在推荐系统中得到了广泛的应用,但是这种方法也存在着推荐精度不高、难以处理稀疏数据等缺点。对此提出一种结合类别偏好的协同过滤推荐算法。在原算法计算用户相似度的基础上,结合用户类别偏好的相似度来计算近邻,从而得到推荐结果。实验结果表明,该方法能较为有效地结合用户的类别偏好,与传统的协同过滤算法相比,有更好的推荐效果。

关键词 协同过滤 用户相似度 类别偏好

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2021.01.049

COLLABORATIVE FILTERING RECOMMENDATION ALGORITHM COMBINED WITH CATEGORY PREFERENCE

Zhang Ziyan[1] Zhou Chi[2]

[1](*School of Medicine*, *Tongji University*, *Shanghai 200092*, *China*)

[2](*School of Software Technology*, *Zhejiang University*, *Hangzhou 310027*, *Zhejiang*, *China*)

Abstract Recommendation system is an extremely effective method for the phenomenon of information overload. Collaborative filtering algorithm has been widely used in recommendation system since its introduction, but it also has a few of shortcomings such as low recommendation accuracy and difficulty in processing sparse data. This paper presents a collaborative filtering recommendation algorithm combined with category preference. On the basis of the original algorithm calculating user similarity, combined with the similarity of user category preference, the nearest neighbor was calculated, and the recommendation result was obtained. The experimental results show that this method can effectively combine the users' category preferences, and has better recommendation effect than the traditional collaborative filtering algorithm.

Keywords Collaborative filtering User similarity Category preference

0 引 言

近几年来,随着网络技术的快速发展,用户数量急剧增加的同时数据信息也大量增多,出现了信息过载、信息迷失等现象,推荐系统应运而生。为使人们能够更加有效地利用资源,接收更符合用户需要的信息,主要采用个性化推荐系统方法解决此类问题。最常见的方法有协同过滤算法[1]、基于内容的推荐算法[2]、二部图网络结构推荐算法[3]、混合推荐算法[4]。

最近邻协同过滤技术[5]是当前应用最为广泛的个性化推荐算法之一,但其存在难以处理稀疏数据、算法较低的可扩展性、推荐结果的难解释性等缺点。针对以上问题,国内外学者们进行了各种深入的探讨和研究,提出了不同的解决方案。早在 1994 年,基于用户的协同过滤算法[6]就被应用在了新闻过滤中。2003 年,亚马逊公司提出了基于 Item 的协同过滤方法[7]。文俊浩等[8]提出了一种基于标签主题的协同过滤算法,从语义层面上计算了用户对项目的偏好概率。林建辉等[9]利用有向网络图构建出用户之间的信任关系,提出一种融合信任用户的协同过滤推荐算法。赵红等[10]通过融合初始资源配置以及协同过滤,形成了

收稿日期:2019-07-31。**张紫嫣**,硕士,主研领域:心血管流行病学。**周驰**,硕士。

更为有效的组合推荐算法。李龙生等[11]将用户行为和物品标签与协同过滤相结合,更好地解决了物品冷启动问题。廖志芳等[12]采用了随机森林来处理用户的属性特征,从而构建出一种新的混合相似度计算模型。卫泽等[13]将用户评分一致频次与评分项目数之比作为惩罚函数引入到相似度的计算中,从而提高了推荐质量。李雪等[14]提出了一种系统主题生成算法,将主题相似度引入到相似度的计算中。但是,如何在海量的数据资源中将信息准确地推荐给用户,依旧是我们目前所面临的一个难题。

由于用户对类别的偏好会导致一定的喜好倾向,所以结合用户类别偏好的相似度计算会比原来的相似度计算更接近真实情况。比如某用户喜欢运动品牌李宁,则他的鞋架上该品牌的鞋子可能会明显多于其余品牌,有相同类别喜好的用户的相似度也会较高。针对此种情况,本文提出一种结合类别偏好的协同过滤算法,通过计算用户对不同类别的偏好程度来得到用户之间的相似度,并结合原有的用户相似度来获得推荐结果,同时还需要对过于热门的类别进行惩罚,避免热门类别的影响,从而提高推荐效果。

1 传统的协同过滤推荐算法

传统的协同过滤算法包括基于用户的协同过滤和基于项目的协同过滤。二者除了相似度计算有所差异之外,本质上都是基于邻域的协同过滤。下面以基于用户的协同过滤为例:

步骤 1 计算用户相似度。计算用户相似度是协同过滤算法的核心部分,有多种不同的方法,如余弦相似度、Pearson 相关系数、Jaccard 相似性度量等。

步骤 2 根据步骤 1 中的相似度计算方法得到用户相似度,进而得到用户的近邻集合,近邻集合的大小通常可以设置为 K,而 K 的取值一定程度上影响了推荐效果。

步骤 3 从用户的近邻集合中得到推荐结果。设目标用户 u 的近邻集合为 $R^K(u)$,则用户对未评分的项目 i 的预测评分 $\hat{r}_{ui}$ 为:

$$\hat{r}_{ui} = \frac{\sum_{v \in R^K(u)} \mathrm{sim}(u,v) \times r_{vi}}{\sum_{v \in R^K(u)} \mathrm{sim}(u,v)} \tag{1}$$

式中:r_{vi} 表示用户 v 对项目 i 的真实评分;$\mathrm{sim}(u,v)$ 表示用户 u 与用户 v 的相似度。

传统的协同过滤算法在计算用户相似度时,往往会受限于数据稀疏的问题,其中的原因是两个用户之间共同喜爱的项目很少甚至没有,这导致了计算用户相似度的时候会与真实的情况存在偏差,从而影响了推荐系统的效果。

2 结合类别偏好的协同过滤

本文通过将协同过滤与用户的类别偏好相结合,来计算用户之间的综合相似度,进而从近邻用户得到推荐结果,来提高推荐的效果。

2.1 余弦相似度

协同过滤算法中经常使用余弦相似度来计算用户相似性。余弦相似度计算方式如下:

$$Sim_{\cos}(u,v) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}} \tag{2}$$

式中:$N(u)$、$N(v)$ 分别为用户 u 和用户 v 有正反馈的项目集合。

2.2 类别偏好

某用户对某一类别的偏好一定程度上会影响用户对 item 的评价结果,可以通过 $F(u,t)$ 来表示用户 u 对类别 t 的偏好程度。

$$F(u,t) = \frac{1}{C_t} \cdot \frac{\sum_{i \in Item(u)} I(i,t)}{Item(u)} \tag{3}$$

式中:$Item(u)$ 为用户 u 有正反馈的 $item$ 集合;$I(\,)$ 为指示函数,用于判断项目 i 是否为类别 t 的项目。

$$I(i,t) = \begin{cases} 1 & i \in Tag2Item(t) \\ 0 & \text{其他} \end{cases} \tag{4}$$

式中:$Tag2Item(t)$ 表示类别 t 所包含的所有 item 集合。

C_t 为对热门 tag 的惩罚系数:

$$C_t = \frac{|Tag2Item(t)| + 1}{\sum_{t \in Tag} |Tag2Item(t)| + |Tag|} \tag{5}$$

这里还额外引入了平滑项,避免出现分子为 0 的情况。

不同的用户有不同的类别偏好,通过计算用户的类别偏好,从而可以依据余弦相似度或其他相似度度量方法来得到用户之间的类别偏好维度上的相似度 Sim_{tag}。

2.3 综合相似度

在得到用户的类别偏好后,同样可以用余弦相似度来计算在类别偏好维度上的用户相似度 Sim_{tag}。在原有的余弦相似度的基础上结合类别偏好的相似度,得到用户综合相似度 $Sim(\,)$:

$$Sim(u,v) = \alpha \times Sim_{\cos}(u,v) + (1-\alpha) \times Sim_{tag}(u,v) \quad (6)$$

式中:α 为权重因子,用来平衡综合相似度中两种不同的用户相似度。

3　实　验

3.1　测试数据集

本实验采用 Movielens-100k 数据集,这是推荐系统的一个经典数据集。该数据集包含了 943 名用户、涵盖了 19 个类别的 1 682 部电影以及 100 000 条评分记录。其中:将 80% 的数据划分为训练集, 剩余 20% 的数据划分为测试集。

3.2　评价指标

本文采用精度(Precision)和召回率(Recall)来评价算法的推荐效果。

(1) 精度:指在推荐结果中,用户真正喜欢的项目数所占的比例。

$$P_{recision} = \frac{TP}{TP + FP} \quad (7)$$

(2) 召回率:指在个性化推荐算法产生的推荐结果中,用户喜欢的项目数占用户真实喜欢项目总数的比例[15]。

$$R_{ecall} = \frac{TP}{TP + FN} \quad (8)$$

式中:TP 为真正例;TN 为真负例;FP 为假正例。

3.3　实验设计及结果分析

实验发现,选取的 α 因子约为 0.8 时,推荐效果最好,所以以下实验的 α 因子均设置为 0.8。

由于不同大小的 K 值会影响推荐的效果,所以对于不同的近邻个数 K,我们取了不同的值来验证算法的有效性。图 1 为在不同 K 值下,结合类别偏好的协同过滤算法在测试集上的精度。

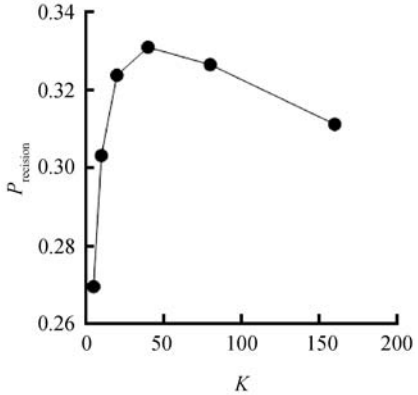

图 1　精度与最近邻个数 K 的关系

图 2 为不同 K 值下,结合类别偏好的协同过滤算法在测试集上的召回率。

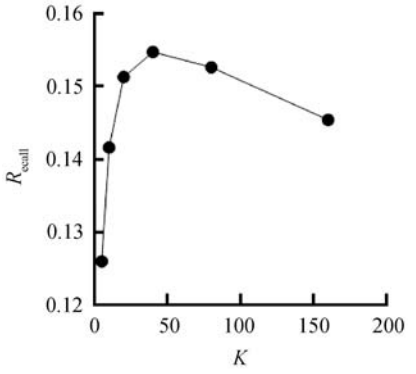

图 2　召回率与最近邻个数 K 的关系

结果表明,当 K 在 20 ~ 80 之间时,精度和召回率均表现较好,说明 K 值在这个范围内的推荐效果较好。而在实际应用中,如果 K 值过大,会导致计算时间的大量增加,从而影响用户的体验,所以选取一个合适的 K 值在推荐系统中极为重要。因此,本文针对小 K 值,选取了 3 ~ 30 之间的多个 K 值来进行实验,将算法改进前后的精度和召回率进行对比。图 3、图 4 是协同过滤算法改进前后的精度和召回率的对比。

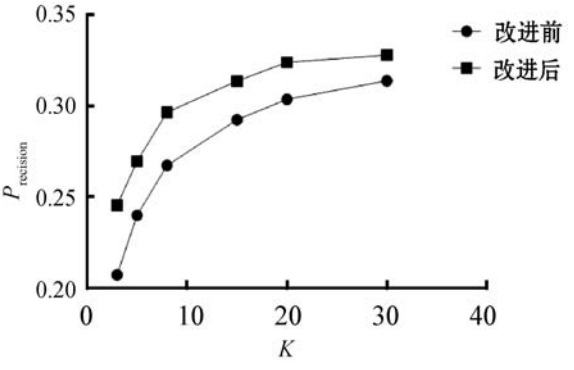

图 3　算法改进前后精度对比

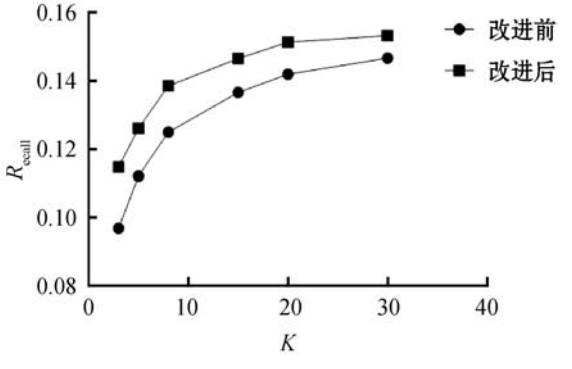

图 4　算法改进前后召回率对比

实验结果表明,在相同 K 值的情况下,结合类别偏好的协同过滤与传统的协同过滤相比,在精度和召回率这两个评价指标上均有较为明显的提升,这表明了改进后的算法能较好地结合用户的类别偏好,使得近邻计算更为合理,从而提高推荐效果。

4 结 语

推荐系统是针对信息过载现象的一种常见处理手段,而协同过滤算法则是其中最为基本的技术之一。本文提出一种结合类别偏好的协同过滤推荐算法,在原有的计算用户相似度的基础上结合了用户的类别偏好相似度,使得算法对用户相似度的计算更为准确,也更能反映真实情况。实验表明,该方法能够有效地结合类别偏好,提高推荐效果。

参 考 文 献

[1] Liang X, Xia Z, Pang L, et al. Measure prediction capability of data for collaborative filtering[J]. Knowledge and Information Systems, 2016, 49(3): 975 – 1004.

[2] 杨武,唐瑞,卢玲. 基于内容的推荐与协同过滤融合的新闻推荐方法[J]. 计算机应用,2016,36(2):414 – 418.

[3] Yu F, Zeng A, Sébastien G, et al. Network-based recommendation algorithms: A review[J]. Physica A: Statistical Mechanics and its Applications, 2016, 452: 192 – 208.

[4] 宋文君,郭强,刘建国. 一种改进的混合推荐算法[J]. 上海理工大学学报,2015(4):327 – 331.

[5] 黎明,徐德智. 一种结合基于项目和用户的个性化推荐算法[J]. 小型微型计算机系统,2011(4):611 – 613.

[6] Resnick P, Iacovou N, Suchak M, et al. GroupLens: An open architecture for collaborative filtering of netnews[C]//1994 ACM Conference on Computer Supported Cooperative Work, 1994.

[7] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1): 76 – 80.

[8] 文俊浩,袁培雷,曾骏,等. 基于标签主题的协同过滤推荐算法研究[J]. 计算机工程,2017,43(1):247 – 252.

[9] 林建辉,严宣辉,黄波. 融合信任用户的协同过滤推荐算法[J]. 计算机系统应用,2017,26(6):124 – 130.

[10] 赵红,郑骏. 融合初始资源与协同过滤的二部图推荐算法[J]. 计算机应用与软件,2019,36(1):291 – 295.

[11] 李龙生,艾均,苏湛,等. 结合用户行为和物品标签的协同过滤推荐算法[J]. 计算机应用与软件,2018,35(6):248 – 253.

[12] 廖志芳,符本才,孔令远,等. 一种新颖的混合相似度计算模型[J]. 计算机应用与软件,2018,35(1):175 – 182.

[13] 卫泽,周登文. 基于用户的优化协同过滤推荐算法[J]. 计算机与数字工程,2017,45(4):613 – 615,628.

[14] 李雪,高心丹. 一种基于系统主题挖掘的协同过滤算法[J]. 小型微型计算机系统,2018,39(4):664 – 667.

[15] 苏健民,郭伟超. 结合用户偏好和项目属性的网络结构推荐算法[J]. 黑龙江大学自然科学学报,2018,35(2):229 – 236.

(上接第 292 页)

[5] Yu H, Liu Z, Wang G. An automatic method to determine the number of clusters using decision-theoretic rough set[J]. International Journal of Approximate Reasoning, 2014, 55(1): 101 – 115.

[6] 张婷,张红云,王真. 基于三支决策粗糙集的迭代量化的图像检索算法[J]. 南京大学学报(自然科学版),2018,54(4):714 – 724.

[7] Zhang H R, Min F. Three-way recommender systems based on random forests[J]. Knowledge-Based Systems, 2016, 91: 275 – 286.

[8] Liu D, Liang D C, Wang C C. A novel three-way decision model based on incomplete information system[J]. Knowledge-Based Systems, 2016, 91: 32 – 45.

[9] Lin G, Liang J, Qian Y, et al. A fuzzy multigranulation decision-theoretic approach to multi-source fuzzy information systems[J]. Knowledge-Based Systems, 2016, 91: 102 – 113.

[10] Qian Y, Zhang H, Sang Y, et al. Multigranulation decision-theoretic rough sets[J]. International Journal of Approximate Reasoning, 2014, 55(1): 225 – 237.

[11] 刘丹,徐立新,李敬伟. 不完备邻域多粒度决策理论粗糙集与三支决策[J]. 计算机应用与软件,2019,36(5):145 – 157.

[12] Zhou B. Multi-class decision-theoretic rough sets[J]. International Journal of Approximate Reasoning, 2014, 55(1): 211 – 224.

[13] Li W, Huang Z, Jia X, et al. Neighborhood based decision-theoretic rough set models[J]. International Journal of Approximate Reasoning, 2016, 69: 1 – 17.

[14] Sun B, Ma W, Zhao H. Decision-theoretic rough fuzzy set model and application[J]. Information Sciences, 2014, 283: 180 – 196.

[15] Song J, Tsang E C C, Chen D, et al. Minimal decision cost reduct in fuzzy decision-theoretic rough set model[J]. Knowledge-Based Systems, 2017, 126: 104 – 112.

[16] Zhang X, Mei C, Chen D, et al. Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition, 2016, 56: 1 – 15.

[17] 姚晟,吴照玉,陈菊,等. 基于决策理论粗糙集的一种新属性约简方法[J]. 微电子学与计算机,2019,36(5):76 – 81.

[18] 姚晟,徐凤,吴照玉,等. 基于邻域粗糙互信息熵的非单调性属性约简[J]. 控制与决策,2019,34(2):353 – 361.

[19] 徐风. 数值型数据的粗糙集模型与特征选择研究[D]. 合肥:安徽大学,2018.