融合惩罚因子和时间权重的协同过滤推荐算法*

刘超慧,韩传福,陈天成,孔先进
(郑州航空工业管理学院 智能工程学院,河南 郑州 450046)

摘　要:协同过滤算法是一种经典的推荐算法,思想是依据近邻用户或者相似物品对目标进行推荐,常被应用在各类推荐系统中。但传统算法过分考虑热门物品对评分的影响,而忽略了冷门物品对用户兴趣特征度量的贡献,也未考虑用户兴趣动态变化的问题。对此,提出一种新的相似度改进算法,改进后的协同过滤算法将物品热门惩罚因子和时间数据权重进行加权计算,优化了用户相似度计算方法,形成了一种新的相似性度量模型。利用 MovieLens 电影推荐数据集验证改进后的算法,实验结果表明,该算法将推荐平均绝对误差(MAE)与传统算法相比降低了 13.2%,推荐质量有了明显提升。

关键词:协同过滤;推荐算法;惩罚因子;时间权重;个性化推荐;相似度融合
中图分类号:TP311　　**文献标识码**:A　　DOI: 10.19358/j.issn.2096-5133.2020.05.004
引用格式:刘超慧,韩传福,陈天成,等.融合惩罚因子和时间权重的协同过滤推荐算法[J].信息技术与网络安全,2020,39(5):17-21.

Collaborative filtering recommendation algorithm based on penalty factors and time weights

Liu Chaohui, Han Chuanfu, Chen Tiancheng, Kong Xianjin
(School of Intelligent Engineering, Zhengzhou University of Aeronautics, Zhengzhou 450046, China)

Abstract: Collaborative filtering algorithm is a classic recommendation algorithm, which is based on the nearest neighbors or similar objects, and is extensively used in many personalized systems. However, the traditional collaborative filtering algorithms excessively consider the influence of popular objects on the scoring, but ignore the contribution of unpopular objects, and do not consider the dynamic change of user interests. In order to solve these problems, an improved similarity measurement algorithm is put forward, which is based on popular penalty factor and time data weight. The improved algorithm is validated on MovieLens dataset. The experimental results show that the MAE is reduced by 13.2% compared with the traditional algorithm, and the quality of recommendations has been significantly improved.

Key words: collaborative filtering; recommendation algorithm; penalty factor; time weight; personalized recommendation; similarity fusion

0　引言

随着互联网技术的发展和移动终端设备的普及,以用户为核心的信息生产模型造成了信息的爆炸式增长,公众很难在海量信息中迅速、准确地找到所需的信息,面临着严峻的“信息过载”问题。以推荐系统为代表的信息过滤技术,是解决“信息过载”问题的常用方法。推荐系统依据用户的历史行为和数据,通过建立模型来挖掘用户需求和潜在兴趣,进而从海量信息中为用户筛选所需的信息[1]。

协同过滤算法是众多推荐算法中使用最广泛、最有效的算法之一,已成功应用于许多商业推荐系统,但其仍存在着一些亟待解决的问题,例如冷启动、数据稀疏和马太效应。对此许多学者进行了卓有成效的研究工作。于洪等人提出基于时间窗口的时间数据权重,将用户兴趣分为长期和短期两类,更好地反映出了用户兴趣变化规律,提高了推荐精度[2];赵文涛等人提出基于时间的 Logistic 权

*基金项目:河南省科技攻关项目(182102210447);河南省高校省级大学生创新创业训练计划项目(S201910485014);郑州航院教研项目(zhjy18-50)

重函数与用户特征属性进行加权的新的相似度度量模型[3];兰艳等人利用衰减因子建立非线性时间加权函数,赋予评分不同的时间权重,提高了推荐的准确性[4]。上述文献虽然考虑了用户兴趣随时间的变化,却未注意到热门物品对用户评分的影响,对推荐精度有一定的影响。

谢修娟等人引入物品流行度与位置信息,提高了推荐结果的多样性[5];孙红等人通过添加物品热门惩罚因子,优化了皮尔逊相似度计算,提高了推荐质量[6];AHM H J 等人通过研究物品热门程度的影响,使用启发式算法对用户相似性度量进行优化,缓解了传统协同过滤算法的冷启动问题[7];焦富森等人考虑物品质量和用户评分倾向性对用户打分的影响,提高推荐效果[8]。这些算法虽弥补了传统算法过分考虑热门物品对评分的影响,却未考虑用户兴趣随时间迁移的情况,无法动态追踪用户的兴趣变化。

本文在基于皮尔逊相似度的基础上进行改进实验,提出了一种融合物品热门惩罚因子和时间权重的相似度计算方法,弥补了传统算法的缺陷。在 Movies-100k 数据集上进行实验,实验结果显示融合后的算法可以有效追踪用户兴趣的变化和降低热门物品对用户评分的影响,提高推荐精度。

1 传统的协同过滤算法

协同过滤算法是一种基于用户特征和行为数据的推荐算法,依据用户的过去行为查找用户或物品的最近邻集,以计算用户对物品的偏好,主要包括基于领域、图、关联规则和知识的推荐算法,其中使用最广泛的是基于领域的方法。

1.1 用户相似度计算方法

(1)欧几里得相似度

$$\mathrm{sim}(\mathrm{user}_x,\mathrm{user}_y)=\frac{1}{1+\sqrt{\sum_{i\in I_{x,y}}(R_{x,i}-R_{y,i})^2}} \tag{1}$$

其中,$\mathrm{sim}(\mathrm{user}_x,\mathrm{user}_y)$ 为用户 x 和用户 y 之间的相似度,$I_{x,y}$ 为用户 x、y 的公共评分集,$R_{x,i}$ 为用户 x 对物品 i 的评分,$R_{y,i}$ 为用户 y 对物品 i 的评分。

(2)余弦相似度

$$\mathrm{sim}(\mathrm{user}_x,\mathrm{user}_y)=\frac{\sum_{i\in I_{x,y}}R_{x,i}\cdot R_{y,i}}{\sqrt{\sum_{i\in I_{x,y}}R_{x,i}^2}\cdot\sqrt{\sum_{i\in I_{x,y}}R_{y,i}^2}} \tag{2}$$

(3)修正余弦相似度

$$\mathrm{sim}(\mathrm{user}_x,\mathrm{user}_y)=\frac{\sum_{i\in I_{x,y}}(R_{x,i}-\overline{R_x})\cdot(R_{y,i}-\overline{R_y})}{\sqrt{\sum_{i\in I_x}(R_{x,i}-\overline{R_x})^2}\cdot\sqrt{\sum_{i\in I_y}(R_{y,i}-\overline{R_y})^2}} \tag{3}$$

其中,$\overline{R_x}$、$\overline{R_y}$ 分别为用户 x、y 的平均评分,I_x、I_y 分别为被用户 x、y 评分了的物品集合,修正后的余弦相似度减去用户 x 和用户 y 对物品的历史评分的均值,这样可以减少用户评分习惯带来的误差。

(4)皮尔逊相关系数

$$\mathrm{sim}(\mathrm{user}_x,\mathrm{user}_y)=\frac{\sum_{i\in I_{x,y}}(R_{x,i}-\overline{R_x})\cdot(R_{y,i}-\overline{R_y})}{\sqrt{\sum_{i\in I_{x,y}}(R_{x,i}-\overline{R_x})^2}\cdot\sqrt{\sum_{i\in I_{x,y}}(R_{y,i}-\overline{R_y})^2}} \tag{4}$$

皮尔逊相关系数通常用于计算两个变量之间的相关性,其值域为[-1,1]。当该值大于 0 时,表示两个变量正相关;当该值小于 0 时,表示两个变量负相关;而该值为 0 表示不相关。皮尔逊相似度计算过程考虑了用户的评分偏好,以避免在用户对同一项目进行评分时因不同的评估习惯而引起的差异。

1.2 相似度计算方法性能分析

利用 MovieLens-100K 数据集对四种相似度计算方法进行性能分析,实验结果如图 1 所示。

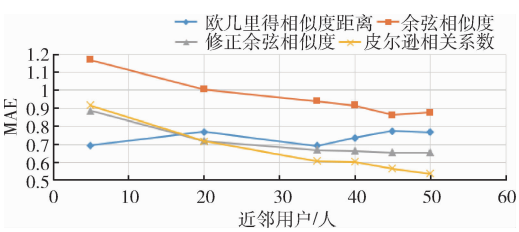

图 1 四种相似度计算方法的对比

实验结果表明,皮尔逊相似度随着近邻人数的增加,平均绝对误差逐渐降低,皮尔逊相似性的结果对分数的绝对值不敏感,更侧重分数间的相对值。与余弦相似度不同的是,皮尔逊相似度考虑了不同用户评分的不同平均值;与修正余弦相似度相比,其去中心化的方式不同。本文以皮尔逊相似度计算方法为基础开展项目研究。

2 基于惩罚因子和时间权重的协同过滤推荐算法

2.1 物品热门惩罚因子相似度改进

协同过滤算法依据用户对物品评估的差异计

算相似度,忽略了物品的热门程度或必需性对相似度的影响。一些大众的、受欢迎的热门物品,并不能充分反映用户的“个性”和潜在的爱好,也不能表明用户之间具有很强的相似性。这时,应减少其对用户相似性度量的影响和贡献。例如,对于给予《战狼2》评分很高的两个观影用户,很难依据这部热门影片判定这两个用户具有相同的兴趣爱好,相反,如果两人都观看过小众的音乐剧影片,则可以反映出用户具有相似的兴趣和爱好。

因此冷门的物品可以更好地反映用户之间的相似性,而且冷门程度越高,相似性越强。考虑到用户的相似度会受热门物品的影响,故添加了物品热门惩罚因子作为加权系数,以抑制热门物品的影响。

本文引入热门物品惩罚因子作为加权系数,改进了皮尔逊相似度计算方法[9],物品出现的次数越多,则表明该物品越大众化,该物品对用户的兴趣相似性的贡献也越小。引入惩罚因子的皮尔逊相似度计算方法如式(5)所示:

$$\mathrm{sim}^1(\mathrm{user}_x,\mathrm{user}_y)=\frac{\sum_{i\in I_{x,y}}\left(\frac{1}{\ln(1+N(i))}\right)(R_{x,i}-\overline{R_x})(R_{y,i}-\overline{R_y})}{\sqrt{\sum_{i\in I_{x,y}}(R_{x,i}-\overline{R_x})^2}\sqrt{\sum_{i\in I_{x,y}}(R_{y,i}-\overline{R_y})^2}} \tag{5}$$

其中,$N(i)$ 表示物品 i 出现的次数。由惩罚项可知,$N(i)$ 越大,$\ln(1+N(i))$ 的倒数越小,则此物品对于用户间的相似度贡献越小。改进后的计算方法可以减弱热门物品的影响。

2.2 时间数据权重

协同过滤算法对用户访问的物品同等对待,没有充分考虑最近访问的物品对用户兴趣的衡量的贡献。最近访问的物品信息更能反映用户的兴趣特征,而早期评估数据应占有较小的比重,因为用户的兴趣会伴随着时间的变化而发生改变,用户感兴趣的物品很可能类似于其最近访问过的物品。因此,引入了基于用户评估时间的数据权重,以增加推荐生成过程中最近访问数据的权重。几个相关的定义如下:

定义1 最早评价物品时刻 T_{first}:表示用户第一次评价物品的时间点。

定义2 最后评价物品时刻 T_{final}:表示用户最后一次评价物品的时间点。

定义3 评价某物品时刻 t:表示一个用户评价某一个物品的时间点。

改进的非线性遗忘函数如式(6)所示:

$$f(t)=\mathrm{e}^{-\alpha\cdot\left[\frac{T_{\mathrm{final}}-t}{T_{\mathrm{final}}-T_{\mathrm{first}}}\right]} \tag{6}$$

其中,参数 $\alpha\in[0,1]$,α 的值影响权重随时间的变化速度,值越大权重增长越快;$f(t)$ 值域是 $[1/\mathrm{e},1]$,该遗忘函数符合人的遗忘理论中收敛性的特点,$f(t)$ 的值越大,表明用户对此物品的评分时间越新,对推荐结果的贡献越大。基于时间数据权重的皮尔逊相似度计算方法如式(7)所示:

$$\mathrm{sim}^2(\mathrm{user}_x,\mathrm{user}_y)=\frac{\sum_{i\in I_{x,y}}(R_{x,i}\times f(t)-\overline{R_x^1})(R_{y,i}\times f(t)-\overline{R_y^1})}{\sqrt{\sum_{i\in I_{x,y}}(R_{x,i}\times f(t)-\overline{R_x^1})^2}\sqrt{\sum_{i\in I_{x,y}}(R_{y,i}\times f(t)-\overline{R_y^1})^2}} \tag{7}$$

其中,$\overline{R_x^1}$、$\overline{R_y^1}$ 分别为用户 x、y 所有物品评分乘以时间数据权重后的物品平均评分。

2.3 改进相似度计算模型

物品热门惩罚因子在考虑到用户相似度受热门物品影响的同时,充分考虑冷门物品对度量用户兴趣特征的贡献。依据时间数据权重构造的非线性遗忘函数,增大了最近物品对用户兴趣的影响,考虑了用户兴趣的变化趋势。融合热门物品惩罚因子和时间权重的相似度模型如式(8)所示:

$$\mathrm{sim}^3(\mathrm{user}_x,\mathrm{user}_y)=(1-\beta)\times\mathrm{sim}^1(\mathrm{user}_x,\mathrm{user}_y)+\beta\times\mathrm{sim}^2(\mathrm{user}_x,\mathrm{user}_y) \tag{8}$$

其中 $\beta\in[0,1]$,表示两种相似度融合权重,根据不同推荐系统,可动态调整 β 的值。

2.4 组合 KNN 推荐算法

本文提出的改进算法模型弥补了传统算法未考虑热门物品和时间数据权重对用户相似度计算的缺陷,利用 KNN 算法得到目标用户的近邻集,采用式(9)计算目标用户对物品的预测评分。

$$P_{x,i}=\overline{R_x}+\frac{\sum_{y\in N_x}\mathrm{sim}^3(\mathrm{user}_x,\mathrm{user}_y)(R_{y,i}-\overline{R_y})}{\sum_{y\in N_x}|\mathrm{sim}^3(\mathrm{user}_x,\mathrm{user}_y)|} \tag{9}$$

其中,$p_{x,i}$ 为目标用户 x 对物品 i 的预测评分,N_x 表示目标用户 x 的近邻用户集,选取其前 N 个目标用户未评估的物品组成 Top-N 推荐集。

2.5 算法描述

改进的协同过滤算法兼顾了热门物品对用户

评分的影响和用户的兴趣随时间的变化,较传统算法更符合实际生活,提高了推荐的精度,算法 1 给出了详细的算法描述。

算法 1:基于惩罚因子和时间权重的协同过滤推荐算法

输入:目标用户 u,用户-物品评分矩阵 $\boldsymbol{R}$,近邻用户个数 K,推荐结果个数 N,参数 α、β 的值;

输出:目标用户 u 的 K 个近邻用户,TOP-N 物品推荐集。

(1)利用文中改进的非线性遗忘函数 $f(t)$(式(6)),计算用户-物品评分矩阵 $\boldsymbol{R}^1$;

(2)计算目标用户 u 的关联用户集 R_u;

(3)利用本文提出的融合型皮尔逊相似度计算模型(式(8)),计算与 u 相似度最高的前 K 个近邻用户集 N_u;

(4)利用组合 KNN 推荐算法预测目标用户未评价物品的评分 $p_{x,i}$;

(5)选择评分较高的前 N 项物品 TOP-N 为目标用户 u 推荐结果集。

3 实验结果及分析

3.1 测试数据集

本文选取明尼苏达大学 GroupLens 小组发布的 MovieLens-100K 数据集进行实验分析(https://grouplens.org/datasets/movielens/100k/),数据主要来自研究人员对电影中不同用户群体的评分调查,u.data 文件包含从 1 682 个电影项目 943 个用户评分中选取的 100 000 条评估数据。选取 70% 用作训练集,剩余 30% 用作测试集。为了有效缓解数据稀疏性,这些用户至少参与了 20 部电影的评估。采用五级评分制,评分越低,用户对电影的喜爱程度就越低。

3.2 算法评估标准

平均绝对误差(Mean Absolute Error,MAE)是推荐系统中最常用的标准,用以衡量推荐算法的质量。其原理是把实验结论与实际结果之间的偏差当作度量,推荐的准确率与 MAE 值的大小成反比,即 MAE 值越小表示推荐算法的质量越好,MAE 计算方法如式(10)所示:

$$\mathrm{MAE}=\frac{\sum_{i\in I_x}|R_{x,i}-P_{x,i}|}{|I_x|} \tag{10}$$

其中,$|I_x|$ 为被用户 x 评分的物品集大小。

3.3 实验结果

为了比较本文提出的算法与传统算法的推荐精度,设计了 3 组实验进行分析。

(1)实验 1:计算分析权重值 α 与平均绝对误差 MAE 的关系。实验中目标用户的近邻用户集数 KN = 25,$\alpha\in[0,1]$,按步长 0.1 进行实验分析,实验结果如图 2 所示。

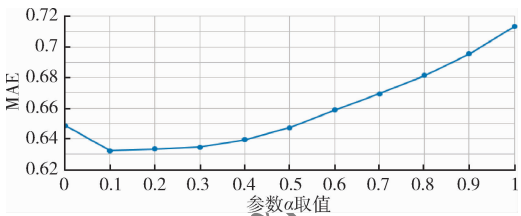

图 2 α 与 MAE 的关系

实验结果显示,在 $\alpha\in[0,0.1]$ 内,MAE 随 α 的增大而减小,且 α 取 0.1 时 MAE 值最小,故本实验选取 α = 0.1 作为非线性遗忘函数的参数取值。

(2)实验 2:计算式(8)中两种相似度融合权重 β 与平均绝对误差 MAE 的关系。实验中目标用户的近邻用户集数 KN = 25,$\beta\in[0,1]$,按步长 0.1 进行实验分析,实验结果如图 3 所示。

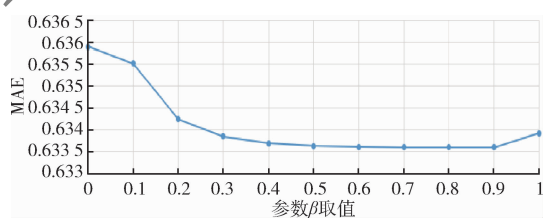

图 3 β 与 MAE 的关系

根据实验结果可知,在 $\beta\in[0,0.8]$ 内,MAE 随 β 的增大而减小,且 β 取 0.8 时 MAE 值最小,故本实验选取参数 β = 0.8 作为改进相似度的融合权重取值。

(3)实验 3:分析比较邻近人数与 MAE 值的关系。利用相同的实验数据,对传统的协同过滤算法与本文的融合算法的推荐精度进行比较,实验结果如图 4 所示。

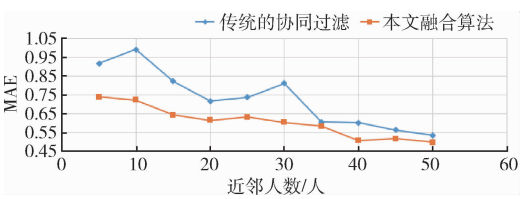

图 4 推荐算法的 MAE 比较

实验结果显示,随着近邻用户数量的增加,两种推荐算法的平均绝对误差 MAE 均呈现下降趋势,并趋于平稳。其原因是近邻用户数量增加时,推荐计算参考的用户也增多,精准度也逐渐上升。本文提出的融合算法比传统算法更稳定,随着近邻用户数量的增加,平均绝对误差逐渐降低,最高降低了26.9%,而当近邻用户较少时,推荐质量也优于传统的协同过滤算法,这表明融合算法在一定程度上可以缓解传统推荐算法的冷启动问题。

4 结束语

本文对比分析了传统协同过滤算法中常用的四种用户相似度计算,在指出其不足的基础上进行改进。充分考虑冷门物品对度量用户兴趣特征的贡献,引入了物品热门惩罚因子;增大最近物品对用户兴趣影响的权重,以反映用户兴趣的动态变化,提出了基于时间数据权重的非线性遗忘函数。进一步提出了基于惩罚因子和时间权重的用户相似度融合算法,实验结果表明,平均绝对误差(MAE)减小幅度较大,可更好地发现用户的兴趣,进行更有效的推荐,克服了传统算法的不足。

该算法引入了物品热门惩罚因子和时间权重,在提高准确率的同时,也增加了计算的复杂。未来的研究方向包括以下两个方面:其一,在不降低推荐精度的前提下提高推荐的广度,以便减少用户的审美疲劳,防止推荐马太效应;其二,进行画像分析,为用户提供更好的个性化推荐服务。

参考文献

[1] 赵小文. 基于协同过滤的推荐算法研究[D]. 西安:西安电子科技大学,2019.

[2] 于洪,李转运. 基于遗忘曲线的协同过滤推荐算法[J]. 南京大学学报(自然科学版),2010,46(5):520-527.

[3] 赵文涛,成亚飞,王春春. 基于 Logistic 时间函数和用户特征的协同过滤算法[J]. 计算机应用与软件,2017,34(2):285-289,312.

[4] 兰艳,曹芳芳. 面向电影推荐的时间加权协同过滤算法的研究[J]. 计算机科学,2017,44(4):295-301,322.

[5] 谢修娟,莫凌飞,李香菊,等. 融合位置信息和物品流行度的协同过滤算法[J]. 河海大学学报(自然科学版),2019,47(6):568-573.

[6] 孙红,韩震. 融合物品热门因子的协同过滤改进算法[J]. 小型微型计算机系统,2018,39(4):638-643.

[7] AHN H J. A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem[J]. Information Sciences,2008,178(1):37-51.

[8] 焦富森,李树青. 基于物品质量和用户评分修正的协同过滤推荐算法[J]. 数据分析与知识发现,2019,3(8):62-67.

[9] JANNACH D,ZANKER M,FELFEMIG A,et al. Recommender systems:an introduction[D]. Cambridge:CUP,2010.

(收稿日期:2020-03-20)

作者简介:

刘超慧(1981 -),男,硕士,副教授,主要研究方向:富媒体资源推荐。

(上接第 5 页)

[12] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016:770-778.

(收稿日期:2020-03-11)

作者简介:

鲁志敏(1994 -),男,硕士研究生,主要研究方向:立体视觉与数字电路设计。

袁勋(1995 -),男,硕士研究生,主要研究方向:立体视觉与集成电路工程。

陈松(1979 -),通信作者,男,博士,副教授,主要研究方向:VLSI 计算机辅助/架构体系设计、立体视觉算法与系统设计。E-mail:songch@ ustc. edu. cn。

版权声明