融合时间与兴趣相似度的产品推荐方法研究

孔元元,白智远,张 飒,吕 品
(上海电机学院 电子信息学院,上海 201306)

摘 要: 互联网信息技术的发展使得企业可以为众多在线用户实现信息的实时交互。如何挖掘出海量产品数据中隐藏的用户行为、实现个性化推荐服务是企业面临的一个重要问题。本项目使用 PHP、Python、MariaDB 等技术对原始数据进行了清理、集成、标记等预处理,然后以用户消费信息中的产品信息为研究对象,运用传统的协同过滤算法,建立用户与产品信息的 0-1 矩阵,得到产品的偏好推荐。通过测试推荐结果发现,模型效果欠佳。为了提高推荐精度,提出了时间权重与兴趣相似度融合的协同过滤模型。时间权重考虑了用户偏好变化与时间的依赖关系,兴趣相似度用于改进模型的预测精度。在包含 4 万余条电视产品收视数据的数据集上实现了该方法,并将其与传统协同过滤模型进行了对比,发现改进后的协同过滤模型的精度得到了显著改善。最后,基于时间权重与兴趣相似度融合的协同过滤模型的推荐结果,给出了增加用户所使用的机顶盒套餐信息的个性化营销推荐方案。

关键词: 电视产品; 协同过滤算法; 时间权重; 兴趣相似度; 个性化推荐
中图分类号: TP311.1 **文献标识码:** A **文章编号:** 1673-629X(2019)09-0195-05
doi: 10.3969/j.issn.1673-629X.2019.09.037

Research on Products Recommendation Method Integrated with Time Weight and Interest Similarity

KONG Yuan-yuan,BAI Zhi-yuan,ZHANG Sa,LYU Pin
(School of Electronic Information,Shanghai Dianji University,Shanghai 201306,China)

Abstract: The development of Internet information technology has enabled enterprises to realize real-time information interaction for many online users. How to dig out hidden user behaviors in the massive products data and implement personalized recommendation services is an important issue for operators. Firstly,we use MariaDB,PHP and Python to clean,integrate,label the raw datasets. Secondly,taking the products information in the user's consumption information as the research object,the 0-1 matrix of user and TV product information is established based on the traditional collaborative filtering algorithm,and the products preference recommendation is obtained. We found that the accuracy of the traditional collaborative filtering algorithm is not desirable by observing the test recommendation results. To improve the recommendation accuracy,an advanced collaborative filtering model integrated with time weight and interest similarity is proposed. The time weight takes into account the dependence of the user's viewing preference changing with time,and the similarity of interest is used to improve the prediction accuracy of the proposed model. The proposed model is implemented on a data set containing more than 40 000 TV products. Finally,we compared the proposed model against the traditional collaborative filtering model. The experiment suggests that the accuracy of the proposed collaborative filtering model is significantly improved. In addition,based on recommendation results of integrating time weights and interest similarity,the personalized marketing recommendation scheme,with the package information of the set-top boxes added,is put forward to the decision-makers.

Key words: television products; collaborative filtering algorithm; time weight; interest similarity; personalized recommendation

0 引 言

随着互联网和信息技术的快速发展,数字化、网络化越来越普及,大量的信息聚集起来,形成海量信息[1-3]。用户想要从海量的信息中快速找寻自己所感兴趣的内容,已经变得越来越困难,因此,推荐系统便应运而生。目前,推荐技术被广泛应用于电子商务[4]、社交网络[5]、新闻网络[6]等系统中。而推荐技术中,协同过滤推荐技术是应用最早和最为成功的技术之

收稿日期: 2018-07-18 **修回日期:** 2018-11-20 **网络出版时间:** 2019-03-27
基金项目: 2018 年上海市大学生科创项目(A1-0224-18-012-071);上海市教育科学研究项目(C17014/17AR04)
作者简介: 孔元元(1996-),男,研究方向为数据挖掘与机器学习;吕 品,博士,副教授,CCF 会员(60050M),研究方向为数据挖掘、情感分析。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20190327.1620.002.html

一[4]。协同过滤推荐系统通过计算物品与物品或用户与用户之间的相似度,结合用户的历史喜好,计算出物品对用户的推荐指数,得出推荐结果。

传统的协同过滤推荐系统只考虑用户是否选择了物品,而忽略了用户因时间的推移而产生的兴趣变化。引入用户访问时间的数据权重的改进型协同过滤推荐系统[7]则更好地反映出用户的兴趣变化,改善了系统的准确度。但是没有考虑到用户选择物品的偶然性问题,使得用户在近期偶然选择物品的行为增大了该类型的物品对用户的兴趣度。

本项目以电视产品推荐为例,提出了时间权重与兴趣相似度融合的协同过滤模型。时间权重考虑了用户收视偏好变化与时间的依赖关系;用最近 T 时段内用户观看的电视节目信息表示兴趣相似度。因此,该方法能有效地反映用户收视的动态变化,克服了传统协同过滤算法在计算推荐过程中将用户已观看的每个电视节目同等对待的弊端,实现了基于用户收视信息、电视产品信息和用户基本信息构建的个性化电视产品推荐。

1 时间权重与兴趣相似度融合的协同过滤模型

1992年,Goldberg、Nicols、Oki及Terry提出了协同过滤的概念[8]。随着个性化推荐技术的快速发展,个性化推荐技术广泛应用于各个领域,其中基于协同过滤的推荐技术最受青睐。协同过滤算法采用给用户使用过的对象(如视频、文档、商品等)反馈信息(如评分),这些反馈被记录下来,通过合作的机制分析数据,帮助潜在用户筛选感兴趣的对象[9]。

时间权重与兴趣相似度融合的协同过滤模型包含三个部分:时间权重、兴趣相似度以及融合模型。

1.1 传统协同过滤推荐模型

在传统的协同过滤推荐系统中,输入数据通常表述为一个 $a \times b$ 的用户—资源访问矩阵 R,其中 a 表示用户数,b 表示资源数,R_{ij} 表示第 i 个用户对第 j 个资源的访问记录。矩阵中的每一个数据元素的值表示用户是否访问该资源(1表示访问,0表示未访问)。

传统的协同过滤算法有两种实现方式[10]:基于用户的实现方式与基于物品的实现方式。

相似度是用于比较两个事物近似度的度量,一般基于距离进行计算。由于通常在原始数据集中,产品的名称和种类繁多,用户矩阵与物品相似度矩阵维度过大。因此,为了提高运行效率,本项目组选择杰卡德相似度度量[11]。杰卡德相似度度量如式1所示。

$$J(A_1,A_m)=\frac{|A_1 \cap A_M|}{|A_1 \cup A_M|} \tag{1}$$

其中,$|A_1 \cup A_M|$ 表示喜欢电视产品1与喜欢电视产品M的用户总数;$|A_1 \cap A_M|$ 表示同时喜欢电视产品1和电视产品M的用户数。

1.2 时间权重

在目前的大部分协同过滤推荐算法中,主要是通过计算用户或资源间的相似度,而忽略了用户兴趣的动态变化。但在实际生活中,用户在不同的时间内,兴趣爱好会发生很大的变化。此时,现有的推荐系统无法及时反映出用户兴趣变化的过程,一旦用户的兴趣发生改变,可能导致推荐的个性化资源或者营销方案在很大程度上偏离用户的本质需求。因此,文献[7]提出基于访问时间的数据加权算法以解决传统协同过滤算法的不足。

算法将基于时间的数据加权[7]做如下定义:

$$W_{\text{time}}(u,i)=(1-a)+a\frac{D_{ui}}{L_u} \tag{2}$$

其中,D_{ui} 表示用户 u 访问资源 i 的时间与用户 u 最早访问某资源的时间间隔;L_u 表示用户 u 使用推荐系统的时间跨度,即该用户最早访问某资源的时间与最近访问某资源的时间间隔;$a \in (0,1)$ 称为权重增长指数,改变 a 的值可以调整权重随访问时间变化的速度。a 越大权重增长速度越快,a 的大小可以影响到算法性能。

1.3 兴趣相似度

式2中,$W_{\text{time}}(u,i)$ 的值随着用户 u 访问资源时间间隔呈线性变化,用户近期访问数据的权重总是大于早期访问数据的权重,从而突出了近期数据的重要性。

在实际生活中,不同用户的兴趣变化速度和规律不同,此外用户的兴趣经常存在反复,所以早期访问的资源对于用户生成个性化推荐也很重要。如果只单纯地使用基于时间的数据权重,而削弱推荐系统中早期资源的作用,极有可能对推荐效果产生负面影响。即用户观看的时长在一定程度上可以反映用户的喜好程度。为此,本项目提出基于兴趣相似度的数据权重,如下:

$$W_{\text{interest}}(u,i)=(1-a)+a\frac{T_{ui}}{D_u} \tag{3}$$

其中,T_{ui} 表示用户 u 选择产品 i 的次数;D_u 表示用户 u 做出选择的总次数,且 $T_{ui} \leqslant D_u$;$a \in (0,1)$ 为权重增长因子。在电视节目推荐的应用中,T_{ui} 为用户 u 观看节目 i 的时长,D_u 为用户 u 观看的总时长。

1.4 融合模型

将两个权重函数用式4进行矩阵的融合。

$$W_{\text{fusion}}=\beta \times W_{\text{time}}(u,i)+(1-\beta) \times W_{\text{interest}}(u,i) \tag{4}$$

其中,$\beta \in [0,1]$,β 和 $1-\beta$ 分别代表两种权重值所占的比例,通过选择合适的 β 将两种加权方法结

合起来,从而进一步提高推荐算法的准确率。

为了提高模型精确度,选择向量的夹角余弦进行相似度度量,计算产品的相似度矩阵。夹角余弦相似度度量如式 4 所示。假设把用户评分看作 n 维空间上的向量,两个用户 x,y 的评分向量分别为 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$,则它们之间的余弦相似性[12] 计算如下:

$$\mathrm{sim}(x,y)=\cos(\boldsymbol{X},\boldsymbol{Y})=\frac{\boldsymbol{X}\cdot\boldsymbol{Y}}{\|\boldsymbol{X}\|\times\|\boldsymbol{Y}\|}=\frac{\sum_{c=1}^{n}R_{x,c}R_{y,c}}{\sqrt{\sum_{c=1}^{n}R_{x,c}^{2}}\sqrt{\sum_{c=1}^{n}x_{y,c}^{2}}} \tag{5}$$

其中,$R_{x,c}$、$R_{y,c}$ 分别为用户 x 和用户 y 对项目 c 的评分。

由于考虑到不同用户的评价标度的不同,通过对项目的平均评分对余弦相似度进行方法改进。设用户 x 和 y 共同评分项目集合用 $I_{x,y}$ 表示,则用户 x 和用户 y 之间的相似性计算如式 6 所示。

$$\mathrm{sim}(x,y)=\frac{\sum_{c\in I_{x,y}}(R_{x,c}-\overline{R_x})(R_{y,c}-\overline{R_y})}{\sqrt{\sum_{c\in I_x}(R_{x,c}-\overline{R_x})^2}\sqrt{\sum_{c\in I_y}(R_{y,c}-\overline{R_y})^2}} \tag{6}$$

其中,$\overline{R_x}$ 和 $\overline{R_y}$ 分别表示用户 x 和用户 y 在已评价项目上的平均评分。

2　基于 Python 的时间权重与兴趣相似度融合协同过滤模型的实现

Python 作为一种编译型语言,拥有大量的第三方库,可用来方便地读取、切片、转置、计算数据。给定模型中数据与数据之间的关联与计算方法,作为 Python 的输入数据即可进行训练,实现模型。为了更直观地表述模型的实现方法,提出的数据处理流程如图 1 所示。

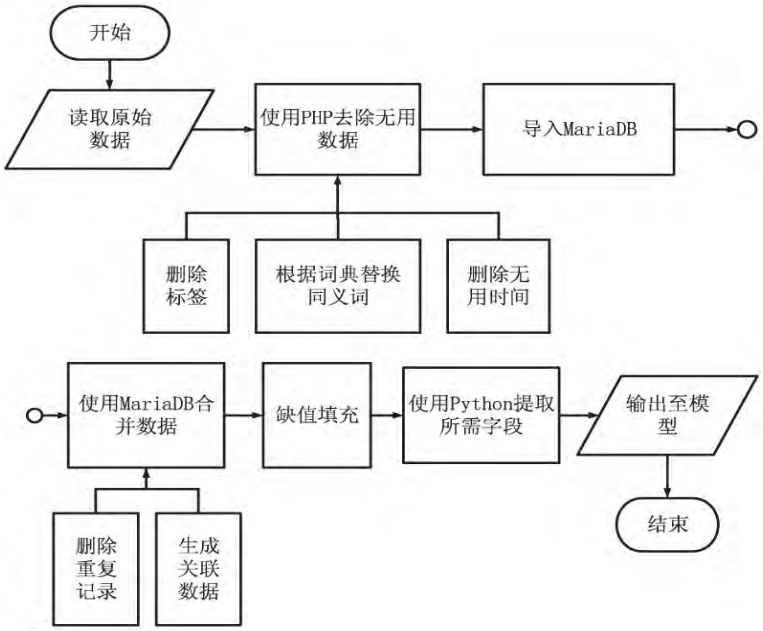

图 1　数据处理流程

2.1　数据清理

考虑到通常的原始数据中存在着大量错误的信息与无用的信息,这些信息会影响处理的效率与模型的训练效果。因此,在读取原始数据后,需要先进行预处理[13]。

原始信息中通常包含的无用信息有无用的标签、同义词以及无用的时间信息,这里无用的时间信息不包括后期模型所需的访问时间。因此在图 1 中,本项目组首先使用 PHP 对原始数据进行过滤与转化,生成格式化后的数据进行下一步处理。

2.2　合并数据

格式化后的数据中可能仍然存在重复的记录,而且数据集中关键信息的单位可能也存在不统一的问题,因此还需要将数据输入 MariaDB 进行合并去重的处理。同时,也要将模型所用到的不同数据关联起来,生成统一的数据集。

2.3　缺值填充

数据集中还有可能存在着缺值的情况。常见的缺值填充方法有: 均值、众数填充,线性回归填充与直接剔除[14]。由于直接剔除会影响缺值记录对用户产生的影响,因此可根据项目对精准度的要求,使用均值、众数填充或线性回归填充[15]。

2.4　训练模型

在完成对数据的处理之后,将数据输入模型,对模型进行训练,计算出用户与物品之间的相似度。最后将测试用的数据集输入模型,即可得出针对每个用户

的个性化推荐结果。

3 实例验证与分析

本节将以电视产品个性化推荐为例，基于 Python，结合用户的历史收视数据对每个用户进行个性化推荐。再利用测试集数据对推荐结果进行验证，与传统协同过滤模型进行对比，得出改进前后模型的精度差异。

3.1 实验设置

实验采用的硬件环境：Inter(R) Core(TM) i7 2.8 GHz 四核处理器；软件环境：操作系统 macOS 10.13，开发环境 Python3.6。实验过程中根据每个用户在数据集中的访问记录为其计算推荐集，如果推荐集中某个资源 i 出现在该用户数据集中的访问记录里，则表示生成了一个正确推荐[16]。为此，使用评估系统效果的准确率(precision) 、召回率(recall) 和推荐覆盖率(coverage) ，作为对比传统推荐算法和改进推荐算法的精度标准[17]。

3.2 实验结果对比分析

由于融合时间与兴趣相似度的协同过滤模型反映了用户兴趣随时间变化的特征，因此，分析两种算法的结果后，得到图 2 和图 3 所示的用户电视产品信息推荐表。

	用户号	产品名称	推荐指数
2	10003	变形金刚2	1.426298
3	10003	熊出没之冬日乐翻天	1.436697
4	10003	爱情公寓三	1.439366
5	10003	农民宇航员	1.441562
6	10003	奔跑吧：奔跑团探班 鹿晗热巴	1.455324
7	10003	人间至味是清欢	1.481004
8	10003	极限挑战 第三季：沙溢被虐惨	1.484697
9	10003	了不起的菲丽西	1.501427
10	10003	神奇大块头	1.505466
11	10003	飞向太空2002	1.505466
12	10003	超凡蜘蛛侠	1.54529
13	10003	秒速5厘米	1.561701
14	10003	超级飞侠二	1.567236
15	10003	神奇阿呦	1.589905
16	10003	火星情报局 第二季	1.613592
17	10003	醉玲珑	1.658611
18	10003	超凡蜘蛛侠2	1.658683
19	10003	楚乔传	1.665395

图 2 传统协同过滤推荐结果

	用户号	产品名称	推荐指数
2	10003	醉玲珑	5.497354
3	10003	警察故事	5.501277
4	10003	骇客时空	5.502695
5	10003	开心魔法	5.503151
6	10003	逃出克隆岛	5.50558
7	10003	了不起的孩子 第二季：陈妍	5.5169
8	10003	超凡蜘蛛侠2	5.517449
9	10003	跨界喜剧王 第二季：王博文	5.520195
10	10003	超人：挣脱束缚	5.540483
11	10003	憨豆先生动画版 第二季	5.541057
12	10003	猪猪侠之五灵守卫者	5.544708
13	10003	神奇阿呦	5.545701
14	10003	超凡蜘蛛侠	5.547954
15	10003	蝙蝠侠：黑暗骑士崛起	5.556991
16	10003	飞向太空2002	5.589861
17	10003	秒速5厘米	5.589912
18	10003	金箍棒传奇2：沙僧的逆袭	5.593421
19	10003	农民宇航员	5.593717

图 3 融合时间与兴趣相似度的协同过滤模型推荐结果

为了比较两种模型之间用户兴趣变化的大小，设计了 3 组实验，将它们的推荐效果与传统的基于物品的协同过滤算法进行比对。

实验 1：不同 β 值的融合比例对模型性能的影响。

为了获得式 4 中最优的 β 值，表 1 给出了融合时间与兴趣相似度的协同过滤模型在不同情况下 β 值由 0.3 依次增长到 0.8 时，模型的准确率、召回率和推荐覆盖率。

表 1 不同 β 值的融合比例对准确率、召回率、推荐覆盖率的影响 %

融合比例	准确率	召回率	推荐覆盖率
0.3	0.419 2	0.566 3	60.462 6
0.4	0.261 6	0.364 1	59.867 3
0.5	0.428 1	0.566 3	61.683 3
0.6	0.265 5	0.364 1	58.071 6
0.7	0.256 4	0.364 1	55.475 7
0.8	0.362 5	0.485 4	56.243 1

通过实验发现 β 值的选取对模型效果有较大影响,得到 $\beta = 0.6$ 时模型在准确率、召回率、推荐覆盖率上性能较优。这是由于它结合了两种加权方法的优点,不仅能突出近期用户观看电视产品信息的重要性,又避免了早期所观看的电视产品信息被忽略的问题,从而更准确地反映了用户的兴趣变化趋势。

实验 2: 分析不同 k 值的推荐数量对模型性能的影响。

表 2 给出了融合时间与兴趣相似度的协同过滤模型在不同推荐数量 k 值下的推荐准确率、召回率、推荐覆盖率。其中推荐数量 k 分别取值为 5,10,20,30,40。

表 2　不同 k 值的推荐数量对准确率、召回率和推荐覆盖率的影响　%

k	准确率	召回率	推荐覆盖率
5	0.426 8	0.283 2	44.125 8
10	0.304 9	0.404 5	62.989 5
20	0.228 7	0.606 8	78.543 8
30	0.223 6	0.890 0	85.659 1
40	0.266 8	1.415 9	89.851 1

通过实验发现,推荐数量 k 的取值对模型有较大影响,且推荐覆盖率与召回率随着 k 的增大而提高,而准确率则随着 k 的增大而降低。得到 $k = 10$ 时模型在准确率、召回率、推荐覆盖率上性能要好于指定的推荐数量 $k = 20$ 的情况。这说明向潜在用户推荐过多的电视产品可能无法真实反映用户的当前兴趣,推荐数量过少则会使推荐出的电视产品具有较大的随机性,难于实现个性化推荐。

实验 3: 模型性能评估。

评估传统协同过滤模型和融合时间与兴趣相似度的推荐模型的运行时间、准确率。实验结果如表 3 所示。

表 3　两种协同过滤模型的评测指标对比 ($\beta = 0.6$, $k = 10$)

模型	运行时间/s	准确率/%
传统协同过滤模型	358.85	0.183 5
融合时间与兴趣相似度的协同过滤模型	564.36	0.304 9

观察表 3 发现,融合时间与兴趣相似度的协同过滤模型在准确率上优于传统协同过滤模型。这是因为融合时间和兴趣相似度模型注重了用户兴趣随时间的变化,避免了用户早期观看的电视节目数据被忽略的问题。但是融合时间与兴趣相似度的协同过滤模型运行花费的时间稍高于传统的协同过滤模型。其原因是融合时间与兴趣相似度的协同过滤模型在计算电视产品间相似度矩阵和初步推荐矩阵时花费的时间比传统的协同过滤模型长。

4　结束语

文中提出了一种基于时间权重与兴趣相似度的改进协同过滤模型。首先提出传统协同过滤在实际应用时的问题,其次针对时间与兴趣相似度问题提出了改进型模型,最后将模型使用 Python 语言实现,并应用在电视产品推荐中,证明了改进后模型的命中率比传统模型的更高。

参考文献:

[1] 杨嘉博.新媒体时代电视发展的困境及对策研究[D].长春: 吉林大学,2017.

[2] 齐晨阳.新媒体背景下电视综艺节目营销策略研究[D].杭州: 浙江传媒学院,2017.

[3] 张良均.Python 数据分析与挖掘实战[M].北京: 机械工业出版社,2016.

[4] 马宏伟,张光卫,李　鹏.协同过滤推荐算法综述[J].小型微型计算机系统,2009,30(7):1282-1288.

[5] 郭宁宁,王宝亮,侯永宏,等.融合社交网络特征的协同过滤推荐算法[J].计算机科学与探索,2018,12(2):208-217.

[6] 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.

[7] 邢春晓,高凤荣,战思南,等.适应用户兴趣变化的协同过滤推荐算法[J].计算机研究与发展,2007,44(2):296-301.

[8] 刘青文.基于协同过滤的推荐算法研究[D].合肥: 中国科学技术大学,2013.

[9] 朱扬勇,孙　婧.推荐系统研究进展[J].计算机科学与探索,2015,9(5):513-525.

[10] 范　波,程久军.用户间多相似度协同过滤推荐算法[J].计算机科学,2012,39(1):23-26.

[11] 张晓琳,付英姿,褚培肖.杰卡德相似系数在推荐系统中的应用[J].计算机技术与发展,2015,25(4):158-161.

[12] 陈华友,盛昭瀚,刘春林.基于向量夹角余弦的组合预测模型的性质研究[J].管理科学学报,2006,9(2):1-8.

[13] 沈睿芳,郭立甫,时希杰.数据挖掘中的数据预处理模型与算法研究[J].计算机系统应用,2005(7):44-46.

[14] 徐　蕾,杨　成,姜春晓,等.协同过滤推荐系统中的用户博弈[J].计算机学报,2016,39(6):1176-1189.

[15] 许必宵,陈升波,韩重阳,等.改进的数据预处理算法及其应用[J].计算机技术与发展,2015,25(12):143-146.

[16] 吴海霞,何　苑,路　璐.个性化推荐系统评测指标与实验方法研究[J].晋中学院学报,2015,32(3):77-81.

[17] 朱郁筱,吕琳媛.推荐系统评价指标综述[J].电子科技大学学报,2012,41(2):163-175.