基于协同过滤技术的电商品牌推荐系统实证研究

孟宪坤

浙江华坤道威数据科技有限公司 浙江杭州 310000

摘要：进入移动互联网时代，网络购物成为大众消费的重要渠道。网购平台中有着为数众多的品牌商品，如何设计并进行有效的品牌推荐，对提高用户购买率及解决新品牌冷启动问题十分重要。本文根据历史购买数据来计算用户对各品牌的购买力分值，通过购买力分值和品牌画像结合的方式，利用混合协同过滤推荐算法，将购买力预测分值最高的品牌推荐给用户。根据实验结果，本文算法在准确性上的表现优于基础协同过滤算法。

关键词：推荐系统；电商品牌；协同过滤

中图分类号：F724.2 **文献识别码**：A **文章编号**：1673-5889（2021）26-0010-03

DOI:10.14097/j.cnki.5392/2021.26.002

一、引言

近十年来，随着移动互联网的快速发展，网络购物已经从新兴消费模式发展成为人们的日常生活方式之一。目前，大型网购平台如淘宝、京东等已拥有较为庞大而稳定的用户群体，与此同时，传统的线下品牌、海外品牌纷纷入驻网购平台，开启了与纯线上销售的品牌同台竞争的局面。随着越来越多品牌的云集，通过构建推荐系统，将消费者历史行为数据和品牌画像数据结合，来为消费者筛选和推荐潜在购买度高的品牌产品，是为消费者提供个性化服务、扩展消费需求和提升满意度的有效方式，能为网购平台和品牌运营方带来切实效益。

为了满足不同环境的推荐系统要求，推荐算法可分为协同推荐方式、基于内容的推荐方式和基于知识的推荐方式。其中，协同过滤推荐方式是应用最为广泛、最具可操作性的方式之一，而协同过滤推荐又可细分为基于记忆的方式、基于模型的方式以及混合协同过滤方式。由于基础的协同过滤算法在数据稀疏特别是用户和产品冷启动时，推荐系统的性能受到较大的局限，在实际应用中常常需要根据情况对算法进行改进。本文采用用户－品牌评分和品牌画像混合的协同过滤算法，通过用户在各品牌上的历史购买行为，采用熵值法计算用户对各品牌的购买力分值，根据品牌与品牌之间共同购买用户数量来确定分值相似度和画像相似度的权重，并根据综合相似度来对相似品牌进行购买力评分预测，对预测分值最高的 Top-N 品牌进行推荐。

二、混合协同过滤算法设计

（一）算法设计思路

本文将用户的历史购买数据和品牌特征数据作为算法的输入，将预测分值较高的品牌清单作为输出，整个推荐算法的过程可分为如下几个步骤：

1. 构建用户－品牌评分矩阵。通过处理用户对各品牌产品的相关购买特征指标数据，得到取值为 1 ～ 5 的购买力分值指标，从而构建用户－品牌评分矩阵。

2. 构建品牌画像矩阵。通过用户的历史购买数据，计算各品牌产品的平均售价、复购周期、人气指数等，结合品牌自身的特征数据如品牌归属国家等指标，构建品牌画像矩阵。

3. 评分相似度计算。根据购买力评分矩阵计算出评分相似度矩阵。

4. 画像相似度计算。根据品牌画像矩阵计算出画像相似度矩阵。

5. 加权相似度计算。根据评分相似度矩阵和画像相似度矩阵计算出加权相似度矩阵。

6. 推荐清单输出：根据最终得到的相似度，找到与目标用户 u 评分较高的品牌所相似的 K 个品牌，根据下式计算得到相应的预测评分，对分值最高的 Top-N 各品牌进行推荐。

$$r_{uc}=\frac{\sum_{d\in N(c)} sim(c,d)\times r_{ud}}{\sum_{d\in N(c)} sim(c,d)}$$

其中，N(c) 代表了品牌 c 相似的且用户 u 已评分过的品牌的集合；sim(c,d) 表示品牌 c 和 d 的综合相似度，r_{ud} 表示用户 u 对品牌 c 的评分。

（二）购买力分值

对各个品牌，统计用户在该品牌中总购买次数、商品数量、支付总额、付款耗时等相关指标。采用熵值法得出各品牌的以上指标对应的权重，并据此计算出综合的购买力分数。根据购买力分数的分布，将购买力分数离散化为取值范围在 1 分 ～ 5 的购买力分值。具体步骤如下：

1. 构建各品牌购买特征指标。总购买次数即六个月内用户 u 在品牌 c 中总共购买成功的次数，宝贝数量即相同统计周期内该用户在品牌 c 中所购买的产品总数量，支付总额即相同统计周期内该用户在品牌 c 中所支付的总金额，付款耗时即该用户从下单到付款之间所经历的时间。

2. 熵值法得到购买力分数。对任意品牌 c 的指标 j 进行归一化处理，对处理后的指标计算用户 u 的占比，然后对所有用户计算该指标 j 的熵值，最后对各项指标根据其熵值估计指标对应的权重并根据权重与占比的乘积和计算得出用户 u 在品

牌 c 的综合分数。其中，在归一化步骤中，我们将总购买次数、商品数量、支付总额越多视为对品牌 c 的购买力的作用越正向从而进行正向归一化，而将付款耗时越久视为对购买力的作用越负向从而进行负向归一化。

3. 购买力分数离散化：根据品牌 c 的综合购买力分数在所有用户上的分布，确定出该分布的五分位数，并根据下式对购买力分数进行离散化处理。

$$r_{uc}=\begin{cases}1, if\ rank<q(0.2)\\ 2, if\ \text{rank}\gg q(0.2)\ and\ rank<q(0.4)\\ 3, if\ \text{rank}\gg q(0.4)\ and\ rank<q(0.6)\\ 4, if\ \text{rank}\gg q(0.6)\ and\ rank<q(0.8)\\ 5, if\ rank>q(0.8)\end{cases}$$

其中，r_{uc} 代表用户 u 对品牌 c 的购买力分值，rank 为购买力分数，q 代表分位数。

三、综合相似度设计

本文对基于项目的协同过滤算法进行改进，采用加权相似度来综合品牌评分和品牌内容的信息。本算法将品牌的用户评分相似度和品牌的画像相似度进行结合，用综合后的相似度来挑选品牌的近邻。两类相似度的加权结合方式如下式：

$$Sim(c,d)=\alpha\times Sim_{portrait}(c,d)+(1-\alpha)\times Sim_{score}(c,d)$$

其中，$Sim_{portrait}(c,d)$ 表示品牌 c 和 d 的画像相似度，$Sim_{score}(c,d)$ 表示评分的相似度；α 和（1-α）表示了两类相似度对综合相似度计算的贡献权重。

考虑到在一个用户量较为稳定的品牌网购系统中，新品牌的冷启动问题对推荐系统的影响较大，本文在设置权重时选择随用户数量平滑变化的函数。具体来说，新品牌推出时，在没有用户购买记录的情况下，品牌之间的相似度以品牌画像相似度作为主要参考；而品牌成熟时，有较多的用户购买记录，则品牌之间的相似度较多的以评分相似度为主要参考。因此，权重系数 α 的设置如下：

$$\alpha=\frac{1}{1+\log(1+|U|)}$$

其中，|U| 表示共同购买过这两个品牌的用户群体的数量。当新品牌推出时，|U| 取值为 0，根据权重系数值可知品牌之间的相似度全部由画像相似度决定；而当用户数量非常大时，评分相似度在很大程度上决定了综合相似度。

（一）品牌购买力相似度

若每个用户对品牌 c 和品牌 d 的购买力分值都相同或相近，则这两个品牌的评分相似度较高，可视为近邻（相似）品牌。为了准确量化不同品牌之间评分向量的距离（相似度），推荐系统常采用向量的余弦值或皮尔逊相关系数。本文采用余弦值来计算品牌的评分相似度，公式如下：

$$Sim_{score}(c,d)=\cos(r_c,r_d)=\frac{r_c\cdot r_d}{\|r_c\|\cdot\|r_d\|}$$

其中，r_c 和 r_d 表示品牌 c 和品牌 d 的用户购买力分值向量。

（二）品牌画像相似度

在大数据时代背景下，用户信息不断扩展，将用户的具体信息抽象成各类标签，利用用户标签来研究用户的方法被称为用户画像。同理，利用电商品牌画像来抽象出一个品牌的综合信息，能够更清晰的研究和了解该品牌。本文选择品牌产品均价、复购周期、人气指数及品牌归属国这四个指标来综合品牌信息，生产品牌画像数据矩阵，并对各指标定义其相似值度量函数。根据各指标的相似度可得出品牌画像的相似度，如下式：

$$Sim_{portrait}(c,d)=\alpha\times Sim_{price}(c,d)+\beta\times Sim_{\text{period}}(c,d)$$
$$+\gamma\times Sim_{popularity}(c,d)+\omega\times Sim_{\text{country}}(c,d)$$

其中，权重系数和为 1，即 α+β+γ+ω=1。对各指标的相似度度量函数则通过以下的进一步分析确定。

1. 品牌产品均价。在统计周期（六个月）内各品牌所售出的产品的平均价格。了解品牌的平均售价，可以了解品牌的价格策略和目标人群定位，在计算用户对品牌的购买力预测分值时有一定作用。例如，用户 u 具有购买高价格品牌产品的偏好，则与这些品牌相似价格定位的品牌 c 也在该用户的购买能力范围内。对于不同品牌之间均价相似值采用如下度量函数：

$$Sim_{price}(c,d)=\frac{1}{\log(|price^c-price^d|+1)}$$

其中，$price^c$ 和 $price^d$ 表示品牌 c 和品牌 d 的产品均价。

2. 复购周期。用户购买同一品牌的产品所经历的平均间隔天数。通过复购周期的不同，可以了解到品牌的产品类别定位。通常，定位在食品类的品牌产品具有较短的复购周期，而彩妆护肤类的品牌产品的平均复购周期则在两个月以上，而耐用品如家电家具等则具有较长的复购周期。一个喜欢购买较短周期的品牌产品的用户和一个只倾向购买耐用品的用户对同一品牌 c 的购买力预测分值会有较大差异。对复购周期可定义如下相似值度量函数：

$$Sim_{\text{period}}(c,d)=\frac{1}{|level^c-level^d|+1}$$

其中，level 表示复购周期的长短等级划分，可根据复购周期（天数）period 通过下式确定：

$$\text{level}=\begin{cases}1, & if\ period<60\\ 2, & if\ period\gg 60\ and\ period<120\\ 3, & if\ period\gg 120\end{cases}$$

3. 人气指数。品牌的日均活跃买家数量。不同的人气指数可以区分品牌的热度，对更喜欢小众品牌的用户，可以向其推荐具有相似热度的品牌。不同品牌的人气指数的相似值可用如下度量函数：

$$Sim_{popularity}(c,d)=\frac{1}{|level^c-level^d|+1}$$

其中，level 表示人气指数的多寡等级划分，可根据人气指数 popularity 通过下式确定：

$$\text{level}=\begin{cases}1, & if\ popularity<100\\ 2, & if\ popularity\gg 100\ and\ popularity<500\\ 3, & if\ popularity\gg 500\end{cases}$$

4. 品牌归属国。创立该品牌的公司所在的国家。由于本文针对的电商系统主要销售海外产品，目标用户对不同国家的品牌偏好度不同，可以通过品牌归属国来有效区分品牌的相似性。品牌归属国的相似值度量函数如下：

$$Sim_{\text{country}}(c,d)=\begin{cases}1, & if\ country^c=country^d\\ 0, & if\ country^c\neq country^d\end{cases}$$

四、实验结果

推荐系统通常从准确性和多样性来对算法性能进行评价，而准确性是算法在实验阶段的最重要指标。对评分预测类的推荐算法而言，准确性是对算法预测分值和实际分值之间的误差的表述，可以采用平均绝对误差法（Mean Absolute Error, MAE）、均方根误差法（Root Mean Squared Error, RMSE）

等方法来衡量算法的准确性，误差越小代表算法的准确性能越好。在评分取值为[1,2,3,4,5]的推荐系统中，对所有预测的预测分值与实际分值的绝对误差取平均可得到MAE值，其计算公式如下：

$$\mathbf{MAE}=\frac{\sum_{i=1}^{N}|r_i-\overline{r_i}|}{N}$$

其中，r_i和$\overline{r_i}$分别表示了实际购买力分值和预测购买了分值，N表示了总共进行的预测次数。

推荐系统中共涉及358个品牌和3000用户，将数据随机分为训练集和测试集，其中80%数据归为训练集，20%数据归为测试集，并将品牌画像相似度中的权重系数设置为[0.4,0.2,0.3,0.1]。实验结果如图所示，即通过MAE来将本文算法与基础的基于用户的协同过滤算法、基于项目的协同过滤算法进行比较。

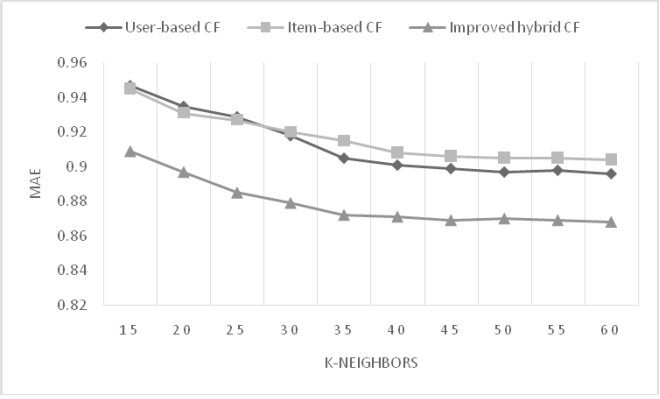

图　不同算法的准确性表现

实验结果表明，本文所设计的算法在推荐准确性上有较好的表现，在各近邻水平上的MAE值均低于基础协同过滤算法。改进的算法能够综合用户对品牌的购买力分值信息和品牌自身的相似信息，并根据数据稀疏度不同来调整两类信息之间的权重，既能应对新品牌数据稀疏的情况，也可以有效应用品牌之间的评分相似度。因此与基础算法相比，本文所设计的混合协同过滤算法具有较优的推荐准确性。

五、结论

本文通过用户的历史购买信息来计算用户对品牌的购买力分值，并设计了混合协同过滤算法，采用购买力分值和品牌画像加权的相似度来度量品牌的相似性，从而根据相似品牌的购买力分值来进行分值预测和品牌推荐。本文所设计的算法在准确性上表现优于基础协同过滤算法，后续可以考虑从综合用户画像信息、扩充品牌画像指标、启发式调整权重系数以及调整画像指标的相似值度量函数等方面进行进一步的算法改进和调优。

参考文献：

[1]牛温佳.用户网络行为画像[M].北京：电子工业出版社,2016.

[2]Su X,Khoshgoftaar TM.A survey of collaborative filtering techniques[J].Advances in Artifical Intelligence,2009.

[3]Cacheda F,Carneiro V,Fernández D,et al.Comparison of collaborative filtering algorithms：Limitations of current techniques and proposals for scalable,high-performance recommender systems[J].ACM Transactions on the Web,2011,5(1):1-33.

[4]石佩生,何军,舒莉,尹皓,冯俊凯.基于用户属性与评分相似因子的推荐算法研究[J].计算机科学与应用,2018,8(01):1-8.

[5]Deshpande M,Karypis G.Item-based top-N recommendation algorithms[J].ACM Transactions on Information Systems,2004,22(1):143-177.

[6]Martinez L,Perez LG,Barranco MJ.Incomplete preference relations to smooth out the cold-start in collaborative Recommender Systems[A]//Fuzzy Information Processing Society,Nafips Meeting of the North American[C].IEEE,2009.

[7]刘宇,陈俊挺,闵华清,等.协同标签系统中基于标签组合效应的推荐算法[J].华南理工大学学报(自然科学版),2013,41(09):65-70.

[8]Vekariya V,Kulkarni GR.2012 International Conference on Communication Systems and Network Technologies-Hybrid Recommender Systems：Content-Boosted Collaborative Filtering for Improved Recommendations[A]//IEEE 2012 International Conference on Communication Systems and Network Technologies(CSNT)-Rajkot,Gujarat,India(2012.05.11-2012.05.13)[C].2012:649-653.

[9]Gupta J,Gadge J.A framework for a recommendation system based on collaborative filtering and demographics[A]//2014 International Conference on Circuits,Systems[C].IEEE,2014.

[10]程文娟,刘云海.融合项目因素的用户部分特征协同过滤算法[J].计算机科学与应用,2018,8(11):1689-1695.

[11]邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法[J].软件学报,2003,14(9):1621-1628.

[12]Breese JS,Heckerman D,Kadie C.Empirical Analysis of Predictive Algorithms for Collaborative Filtering[M].Burlington,Massachusetts：Morgan Kaufmann,1998:43-52.

[13]项亮.推荐系统实践[M].北京：人民邮电出版社，2012.

作者简介：

孟宪坤，供职于浙江华坤道威数据科技有限公司，董事长，中国互联网协会理事。