基于改进的 TF-IDF 标签权重算法的电商用户画像构建

白雨珂[1] 卢胜男[1]
BAI Yuke LU Shengnan

摘 要 在电商环境中，用户画像构建是为了更好地理解和满足用户需求而进行的重要任务。传统的 TF-IDF 标签权重计算方法无法很好地对标签权重进行调整，为了解决这一问题，提出基于 TF-IDF 算法的改进方法，旨在提高用户画像的准确性和个性化程度。融合相关系数矩阵，对相关性强的标签进行适当降权操作。不同类型的行为对标签信息产生不同的权重，并且标签的权重可能会随着时间的推移而衰减。因此，采用拟合记忆遗忘曲线模拟得到的兴趣遗忘曲线，对用户画像权重进行调优操作。实验结果表明，使用所提出的改进的 TF-IDF 算法构建用户画像的效果得到显著的提升。

关键词 电商；相关系数；标签权重；用户画像；TF-IDF 算法

doi：10.3969/j.issn.1672-9528.2024.08.011

0 引言

在当今数字化时代，随着电商平台交易量的不断增长，个性化推荐和精准营销逐渐成为各大电商企业竞争的关键要素。而实现个性化推荐的关键在于准确把握用户的兴趣和偏好，构建准确的用户画像则成为实现个性化推荐的基础。在这一背景下，基于 TF-IDF[1]（term frequency–inverse document frequency）标签权重算法在电商用户画像构建中发挥着重要作用。

TF-IDF 算法自提出以来，凭借着较高的准确率和召回率被广泛使用，但该算法仍然有很大的改进空间。张雷 [2] 提出 word2vec 结合 TF-IDF 权重计算方法进行个性化产品推荐。Shang 等人 [3] 提出一种标签与用户资源相结合的二分图

1. 西安石油大学 陕西西安 710065

标上优于均值插补与中位数插补；Logistic 回归法可有效修复类别型数据，对比 Kendall's W 系数，证明其在专家共识度上优于众数插补与中位数插补。两种方法针对不同数据类型，均可以有效描述专家群体一致性，MICE 对打分数据的插补能有效体现专家自身偏好性，实现对项目评估中缺失数据的修复工作，提升项目评估网的数据质量。

参考文献：

[1] 向南，豆亚杰，姜江，等 . 基于专家信任网络的不完全信息武器选择决策 [J]. 系统工程理论与实践 ,2021,41(3):759-770.
[2] 王晓静，王学丽 . 抽样调查中的不完全数据处理研究 [J]. 陕西科技大学学报 (自然科学版),2009,27(2):141-144.
[3] 岳勇，田考聪 . 数据缺失及其填补方法综述 [J]. 预防医学情报杂志 ,2005(6):683-685.
[4] 刘凤芹 . 基于链式方程的收入变量缺失值的多重插补 [J]. 统计研究 ,2009,26(1):71-77.
[5] 李锦绣 . 基于 Logistic 回归模型和支持向量机（SVM）模型的多分类研究 [D]. 武汉 : 华中师范大学 ,2014.
[6] 刘展，金勇进，韩显男 . 基于倾向得分匹配的缺失数据插补方法 [J]. 数学的实践与认识 ,2016,46(12):193-201.
[7] 刘燕 . 基于 Logistic 回归的近邻择优插补法 [D]. 天津 : 天津财经大学 ,2013.
[8]Sample Size Charts of Spearman and Kendall Coefficients[J]. Journal of biometrics & biostatistics,2021,12(2):1-7.
[9] YU W Z, WANG Y B. An efficient methodology for hardware Trojan detection based on canonical correlation analysis[J]. Microelectronics journal, 2021,115:105162.1-105162.7.

【作者简介】

李梓祎（2001—），女，贵州铜仁人，本科，助理工程师，研究方向：自动化技术、信息处理。

吴昕阳（1988—），女，山东淄博人，博士，副教授，研究方向：系统建模与综合评估。

（收稿日期：2024-05-13）

模型，利用标签信息计算图中边的权重，实现个性化推荐，提高了推荐的准确性。姚海英[4]提出了一种改进的 TF-IDF 算法，结合卡方统计方法和类内信息熵，用于计算特征项的权重。这种改进方法相较于传统方法在权重计算方面表现出显著的优势。高军等人[5]提出了一种改进的 TF-IDF 算法，采用基于日志关联的相关性权重计算方法。这一方法利用 MapReduce 编程框架，根据用户的历史检索记录动态地调整检索词的权重，以提升用户与系统的交互能力。肖慧莲等人[6]运用网络爬虫获取在线评论，通过词频统计和 K-Means 聚类得到顾客满意度评价体系的指标，使用 TF-IDF 计算各项指标的权重，以评估生鲜产品的总体满意度。

现阶段比较缺乏对标签权重计算的研究，而传统的 TF-IDF 标签权重算法又存在一定的不足。因此，本文旨在探讨如何改进 TF-IDF 标签权重算法，以构建更准确、更全面的电商用户画像。同时，进一步探讨改进后的 TF-IDF 标签权重算法在电商个性化推荐和营销领域的应用前景，为电商企业提供更有效的用户服务和营销策略。

1 用户标签权重计算

1.1 TF-IDF 标签权重计算

TF-IDF 算法是一种统计方法，它可以评估单词或短语相对文档集合中其他词语的重要程度。在此实验中，每个用户身上同一标签出现的次数越多，该标签对用户的重要性就越大；该标签在所有用户的所有标签中出现的次数越多，标签的重要性就越低。所以，本文通过电商环境下 TF-IDF 标签权重算法对用户标签权重进行计算。使用 TF-IDF 方法来表示标签 T（Tag）和用户 P（User）之间的关系，其中，$w(P,T)$ 表示用户 P 被 T 标记的次数，$TF(P,T)$ 表示所有标签中标签数量的比例。

$$TF(P,T)=\frac{w(P,T)}{\sum_{T_i=tags} w(P,T_i)} \tag{1}$$

该比率反映了用户 P 与标签 T 的关联程度，比率越大，用户 P 与标签 T 的关系越强。

逆向文件频率（inverse document frequency，IDF）表示标签 T 的稀缺程度，即对于所有用户，T 标签出现在所有用户创建的标签集中的概率。若标记用户 P 的 T 标签出现在标签集中的概率较低，说明标记用户 P 的 T 标签本身较稀缺，则用户 P 和 T 标签之间的关系更紧密。

$$IDF(P,T)=\log\left(\frac{\sum_{P_j=users}\sum_{T_i=tags} w(P_j,T_i)}{\sum_{P_j=tags} w(P_j,T)}\right) \tag{2}$$

式中：$\sum_{P_j=users}\sum_{T_i=tags} w(P_i,T_i)$ 为全部用户的全部标签之和，$\sum_{P_j=tags} w(P_j,T)$ 为所有标签对用户 P 的标记次数。

最后，用户 P 和标签 T 的关系系数即为 $TF(P,T)$ 与 $IDF(P,T)$ 的乘积：

$$TF-IDF=TF(P,T)\times IDF(P,T) \tag{3}$$

1.2 融合相关系数矩阵的权重计算

在电商环境中，标签之间的相关性可能对标签权重产生重要影响，但这一点在传统的 TF-IDF 算法中未被充分考虑。因此，在计算标签权重时，会因为相关性而放大标签的权重，可能导致推荐准确度不够高。为此，本文提出，在使用 TF-IDF 算法计算标签权重后，应使用相关系数矩阵来识别具有强相关性的标签并进行相应的权值处理。这种方法能够在标签权重计算的基础上考虑标签之间的相关性[7]，以提高推荐系统在电商环境中的准确性和效果。

在相关系数矩阵中，每个单元格的值表示被相应行和列标签标记的用户数。最初，对角线上的值为 0，即相关系数矩阵为对称矩阵。如图 1 所示，被 B 和 C 标签同时标记的为用户 1，被标签 A 和 C 同时标记的为用户 2，被标签 A 和 B 同时标记的为用户 3。

0 1 1
1 0 1
1 1 0

图 1 相关系数矩阵

第一行第二列的数字 1 表示被标签 A 和 B 同时标记的人数为 1，其他行和列的值可以依次检索。在相关系数矩阵中，以标签 A 和标签 B 为例，被标签 A 和 B 标记的用户数与同时被两个标签标记的用户总数的之比为标签 A 和 B 的相关系数。计算公式为：

$$r_{A,B}=\frac{sum(A,B)}{\sum_{i,j\in 标签集合} sum(i,j)} \tag{4}$$

本文采用相关系数作为衡量标签相关性的指标，一般相关系数越大，相关性越强。它的取值范围通常在 -1 ～ 1 之间，并设定了一个临界值为 0.4。当两个标签的相关系数小于 0.4 时，认为两个标签之间的相关性不强；反之，认为相关性较强。对于较强相关的标签，通过乘以一个较小的权重来减弱其影响，然后用一个较大的权重减去这个结果，以减少较小权重标签对较大权重标签的影响。同样的方法也适用于减少较大权重标签对较小权重标签的影响[8]。如果两个标签相关系数结果小于 0，就丢弃标签。否则，更新标签权重以反映新结果。

2 用户画像构建

2.1 数据预处理

构建用户画像首要的工作就是数据采集。数据的完整性和准确性会直接影响用户画像的准确性。一般用户数据会分为静态数据和动态数据两种。

本实验采用的数据集为某电商平台的实际交易记录。每笔交易记录都包含客户 ID、交易时间、交易金额和交易附言等四个字段，时间跨度为五年，以上数据均已做脱敏处理。数据集各字段说明如表 1 所示。

表 1 客户交易记录字段说明

英文名称	中文名称	备注
user_id	客户 ID	客户的唯一标识
payment	交易金额	正为支出，负为收入
describe	交易附言	对此项交易的文字描述
unix_time	交易时间	Unix 时间戳

由于原始数据存在一些质量问题，为了便于后续的数据分析，需要进行数据预处理。利用 Pandas 对交易数据进行预处理。首先对原始数据进行统计分析以及异常值和缺失值处理，将处理好的数据进行格式转换后过滤重复数据，生成处理后的交易数据。

2.2 用户交易行为分析

用户行为分析不仅限于基本的用户信息，还需要通过以下三个部分对用户交易行为进行分析。

第一，时间维度的分析。交易主要集中在 2016 年 7 月至 2017 年 12 月，所以选取该时间段的交易记录进行重点分析得知，大部分交易时间集中在 0 点左右，凌晨 1 ～ 7 点交易数量较少，其余时间段交易数量分布较为均衡。

第二，交易属性的分析。对交易金额和次数进行分析得知，客户交易次数为 0 ～ 7000 不等。根据客户的交易记录构建指标，如果交易数量不足，就将这类用户归类为休眠客户。不同用户的平均交易金额差异较大，分布主要集中在 0 ～ 10 000 之间，呈长尾分布。

第三，文本数据的分析。对交易附言使用 jieba 库进行分词、去停用词、关键词抽取和词性标注等文本预处理。通过对数据的分析，可以使用时域分析方法来计算关键字权重。

2.3 标签体系构建

用户画像的本质是用户信息标签化，通过给每个用户贴上不同的标签，可以描绘出他们的特定特征，进而理解他们的购物偏好[9]。根据用户交易行为分析提取关键词作为标签，构建标签体系，得到用户标签集合。标签主要分为四大类。

事实类标签：从用户交易记录中进行统计和计算，可以得到网购首单时间、网购尾单时间、月均消费频度、网购订单平均金额等 40 个事实类标签。

规则类标签：在事实类标签的基础上，对用户的某项指标进行计算或归类，判断用户是否为休眠客户、是否有高端消费等，得到 9 个规则类标签。

预测类标签：借助用户价值模型 RFM，将客户价值等级分为高价值用户和低价值用户，得到 1 个预测类标签。

文本类标签：从用户交易附言文本中提取的关键词作为 100 个文本类标签，如表 2 所示。

表 2 文本类标签

文本特征	对应人群
停彩、大乐透、双色球、福利彩票、彩票、竞彩、追号	爱好购买彩票的人群
儿童、卡通、可爱、女童、零食	家庭中有孩子的人群
男士、衬衫、透气、真皮、足球、运动、小米、汽车、电子、汽车	广大的男士人群
修身、专柜、保湿、女士、女装、显瘦、打底、蕾丝、欧美、韩版	爱美的女士人群
婴儿、孕妇、宝宝	家庭中有孕妇 / 婴儿的人群
分期、贷款、金融服务	有借贷分期需求的人群
基金、投资、证券	有投资需求的人群

将得到的用户标签从交易属性、消费偏好、行为特征三大维度结合结构化数据和非结构化文本数据构建用户标签体系。用户标签体系共包含 150 个标签，框架如图 2 所示。

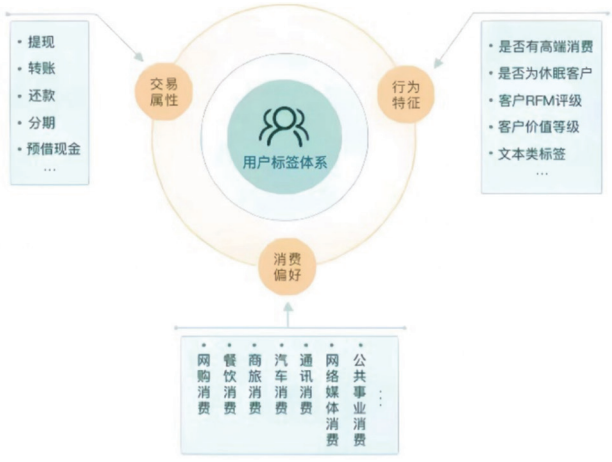

图 2 用户标签体系框架

用户画像构建的常用方式为标签云，输入标签及对应的权重，其中标签的大小或颜色深浅表示其权重大小，从而形成用户的个性化画像。

2.4 电商环境下标签权重的优化

考虑到用户的兴趣和偏好会随着时间的推移而逐渐减少，就像人类的记忆逐渐减弱一样[9]，因此，通过拟合记忆遗忘曲线来模拟用户的兴趣遗忘曲线，用于优化用户画像标签的权重，具体步骤如下。

（1）计算时间间隔：对于标签，遍历与用户标签匹配的所有订单，获取用户交易的最后一个订单的时间，与当前时间作差，以月为单位，若不到一个月则计算为 0。

（2）调整标签权重：将用户标签的融合相关系数矩阵计算出的权重乘以遗忘曲线获得的新的遗忘系数，得到标签的新权重。

$$W_n = W_0 \cdot \left(\frac{0.75}{1+0.42t} + \frac{0.25}{1+0.0003t} \right) \quad (5)$$

式中：W_n 表示某一用户某一标签的新的权重；W_0 表示之前的权重；t 表示间隔时间，单位是月。

3 实验结果与分析

3.1 实验环境

本实验在 Windows 10 操作系统上进行，使用 Python3.9.2、PyCharm Community Edition2021.3.3，通过 Python 的 wordCloud 库函数和 scipy.misc 来生成词云。

3.2 实验结果分析

将文本数据使用分词处理后，即可实现词云的生成。最终通过 WordCloud 生成的用户画像标签可视化效果如图 3 所示。

图 3 用户标签词云

为了验证改进后算法的性能提升，与传统的 TF-IDF 算法的实验结果进行对比，各项评价指标如表 3 所示。

表 3 实验各项评价指标

	P	R	F_1-Score
TF-IDF	0.843 6	0.823 8	0.833 6
优化后 TF-IDF	0.865 4	0.892 7	0.878 8

由表 3 的实验结果可以看出，后者的准确率和召回率略高于前者，侧面反映了使用融合相关系数矩阵的 TF-IDF 算法构建用户画像的效果得到较显著的提升。

4 结语

本研究旨在探讨在电商环境下通过 TF-IDF 算法改进的用户画像构建方法，并验证其在个性化推荐方面的有效性。本文提出了一种改进的 TF-IDF 算法来构建用户画像。通过融合相关系数矩阵，对相关性强的标签进行了适当的降权操作，同时考虑了不同行为类型产生的标签信息具有不同的权重以及标签权重可能随时间衰减的情况。通过拟合记忆遗忘曲线建模的兴趣遗忘曲线，优化了用户画像的权重。本研究的结果表明，TF-IDF 算法在用户画像构建中具有良好的适用性和实用性，能够有效提升电商平台的竞争力和用户体验。未来的研究可以进一步探索其他算法和技术的应用，不断完善和优化用户画像构建方法，为电商行业的发展和用户需求的满足做出更大的贡献。

参考文献：

[1] 郑霖，徐德华．基于改进 TF-IDF 算法的文本分类研究 [J]. 计算机与现代化 ,2014(9):6-9+14.

[2] 张雷．基于 word2vec 和 TF-IDF 算法实现酒店评论的个性化推送 [J]. 电脑与信息技术 ,2017,25(6):8-11.

[3]SHANG M, ZHANG Z.Diffusion-based recommendation in collaborative tagging systems[J].Chinese physics letters, 2009, 26(11):250-253.

[4] 姚海英．中文文本分类中卡方统计特征选择方法和 TF-IDF 权重计算方法的研究 [D]. 长春：吉林大学 ,2016.

[5] 高军，黄献策．基于 Hadoop 平台的相关性权重算法设计与实现 [J]. 计算机工程 ,2019,45(3):26-31.

[6] 肖慧莲，徐锐．基于文本挖掘的生鲜电商顾客满意度研究 [J]. 科技和产业 ,2022,22(1):288-294.

[7] 王洋，丁志刚，郑树泉，等．一种用户画像系统的设计与实现 [J]. 计算机应用与软件 ,2018(3):8-14.

[8] 李传亮．相关系数的意义 [J]. 西南石油大学学报 (自然科学版),2010,32(6):74.

[9]ZHU Z, LI D, LIANG J, et al.A dynamic personalized news recommendation system based on BAP user profiling method[J]. IEEE access,2018,6:2169-3536.

[10]HERMANN E.Memory: a contribution to experimental psychology[J].Annals of neurosciences, 2013,20(4):200408.

[11] 刘海鸥，孙晶晶，苏妍嫄，等．国内外用户画像研究综述 [J]. 情报理论与实践 ,2018,41(11):155-160.

[12] 袁航．基于用户画像的商品推荐系统设计与实现 [D]. 武汉：华中科技大学 ,2021.

[13] 陈煜东．基于用户画像的商品推荐研究 [D]. 南昌：东华理工大学 , 2020.

【作者简介】

白雨珂（2000—），女，陕西西安人，硕士研究生，研究方向：机器学习、大数据技术与应用。

卢胜男（1982—），女，江苏徐州人，博士，副教授，硕士生导师，研究方向：图像处理与机器学习、大数据技术应用。

（收稿日期：2024-05-16）