基于评论关键词的电子商务平台推荐系统设计

袁剑锋 刘佳 严磊

传统的协同过滤推荐算法存在的数据稀疏性导致物品间相似性计算不准确、推荐准确度不高的问题。针对上述问题，提出一种基于评论关键词的协同过滤推荐算法，采用权重排序自学习的方法提取商品评论中的关键词，将TopN关键词的重合度作为重要指标引入到商品相似度的计算中，并在此基础上实现协同过滤商品推荐算法。在阿里巴巴天池大数据竞赛公开数据集上进行实验。结果显示当N取7时，算法表现最好，准确率相比于传统的itemCF提高了28.2%，召回率提高了23.6%，推荐精度在很大程度上得到了提高。

**关键词：电子商务，推荐，关键词，协同过滤**

**1 引言**

近年来，随着Web技术的飞速发展、大数据时代的到来，各类信息呈现爆炸式增长。用户在面对海量复杂信息和数据进行信息检索和过滤时，常常会被淹没在信息资源的海洋中[1]。当人们在互联网上享受信息带来的便利的同时，也会面临信息过载的难题，且已非常严峻，迅速而又正确地定位所需信息变得非常困难，用户可能会迷失在大量冗余的信息中，最终导致用户流失[2]。用户可能会更热衷于根据自身需求定制个性化的信息服务，如何根据用户的购物习惯为其精准高效的推荐合适的商品，将是信息决策支持领域中的重要研究课题之一。

**2.相关算法**

推荐算法的研究在国外起步较早，不少研究者从不同的角度对推荐算法进行了不同分类。从信息技术的角度分类，可依据推荐结果的算法和生成机制分为协同过滤﹑基于内容推荐。协同过滤算法是目前业内公认的较为成熟的推荐算法。Goldberg等[3]首次提出协同过滤算法并将其应用于Tapes-try 电子邮件过滤系统。该算法主要根据用户之前的喜好以及与其兴趣相近的用户的选择来推荐物品。上述最近邻法的弊端在于需要大量计算，因此Linden等[4]提出基于物品的协同过滤算法。该算法通过寻找相似的商品代替寻找用户最近邻方法，在线计算花费与用户数量和物品数量无关，可以在海量数据上实时生成高质量推荐。协同过滤算法的优点较为明显﹐工程上实现简单、效果好﹑模型通用性强。但当电子商务系统规模扩大，用户、项目数据急剧增加时，数据稀疏问题便暴露。此外，它也存在较为严重的冷启动问题。Shambour 等[5]在传统的基于用户的协同过滤推荐算法中融入了项目评分信任度的思想，同时摒弃了传统的相似度计算方法﹐实验证明该算法能缓解数据稀疏性问题。薛美琦[6]提出了一种隐式反馈协同过滤推荐系统模型，对推荐系统中的冷启动问题提出改进方案。利用负采样技术解决训练模型复杂性问题，提高了模型的训练效率。该推荐系统只可以解决冷启动用户问题，但是对于冷启动商品仍然束手无策。江水[7]基于协同过滤技术推荐系统的探究一文中阐述了协同过滤技术的概念内容、前提假设、数据采集与过滤模式、数据结构、评价标准和算法分类，并着重分析对比了基于存储、基于模型和关联规则等三种推荐系统算法的不同实现方式及其优缺点。

**3. 用户网购流程研究**

对不同年龄段，不同职业的用户进行深入的调查，总结出网购流程如图1所示。



图1 网购流程

图1可看出用户在购物时，商品的评价信息往往对用户是否购买本商品具有决定性的影响。因此商品的评价信息中隐藏着大量的信息，从商品的评价中提取关键因素，做商品推荐系统具有较好的实际意义。当前有很多类似的研究，但没有专门针对评价关键词进行推荐的系统。

**4.基于关键词的推荐算法设计**

本文在协同过滤算法中评价矩阵稀疏的背景下，从用户对商品的评价信息中获取商品的内容数据，提取出评论的关键词，通过商品评论关键词之间的相似性度量商品之间的关联性，最终生成推荐结果，使得商品推荐结果更加符合用户的兴趣，从而达到提升推荐效果的目的，算法流程如图2所示。



图2 基于关键词的推荐算法流程

**5.关键词提取技术**

关键词是能够表达文档中心内容的词语，常用于计算机系统标引论文内容特征、信息检索、系统汇集以供读者检阅。关键词提取是文本挖掘领域的一个分支，是文本检索、文档比较、摘要生成、文档分类和聚类等文本挖掘研究的基础性工作。

从算法的角度来看，关键词提取算法主要有两类[8]：无监督关键词提取方法和有监督关键词提取方法。无监督关键词提取方法不需要人工标注的语料，利用某些方法发现文本中比较重要的词作为关键词，进行关键词提取。该方法是先抽取出候选词，然后对各个候选词进行打分，然后输出topK个分值最高的候选词作为关键词。根据打分的策略不同，有不同的算法，例如TF-IDF，TextRank，LDA等算法。有监督关键词提取方法将关键词抽取过程视为二分类问题，先提取出候选词，然后对于每个候选词划定标签，要么是关键词，要么不是关键词，然后训练关键词抽取分类器。当新来一篇文档时，提取出所有的候选词，然后利用训练好的关键词提取分类器，对各个候选词进行分类，最终将标签为关键词的候选词作为关键词。

无监督方法不需要人工标注训练集合的过程，因此更加快捷，但由于无法有效综合利用多种信息对候选关键词排序，所以效果无法与有监督方法媲美；而有监督方法可以通过训练学习调节多种信息对于判断关键词的影响程度，因此效果更优。

**5.1数据预处理**

数据预处理的主要目的是将一些与实际不符或不具有参考价值的评价筛选出去。

1. 去除停用词：比如一些没有实际意义的语气词“哈哈”，“呵呵”等，还有一些纯标点符号的评价。
2. 长度过短的评价：若长度过短，则关键词提取的准确率较低影响下一步商品相似度的计算，最终影响推荐结果的准确性，因此要将其剔除。
3. 不具有参考价值的内容：例如表情符号组成的评价，以及为了刷排名大量重复的评价。

**5.2分词处理**

分词的主要工作是将原始样本转换为关键词提取的样本，过程如图5所示。



图3 分词处理过程

**5.3 提取关键词**

经过分词预处理，得到分词后的词语集，本文综合考虑词频，区域位置，词距，3个特征项，提出了候选词权重计算函数如下：

 (1)

其中表示候选词的权重为3个调节因子。

### （1）词频

词频特征项表示词语在文本中出现的次数，是对词语是否关键的一种常见测度[6]。本文采用公式对词频进行计算，为候选词出现的次数。 使用非线性函数计算词频基于两点考虑：其一，词语的词频和类正比，随着词语出现次数的增大逐渐向1 收敛，即词出现越多，成为关键词的可能性越大，符合词频价值。其二，随着词语出现次数持续增加，词频对词语成为关键词的贡献越来越小，即成为关键词的可能性不会随着次数的持续增加而无限增长，这与语言实际情况相符。

### （2）区域位置

经研究发现[7]，词语出现在首段中比出现在其他段落更能反映文档主题，摘要中的词语比正文中的词语对文档主题贡献大，而标题中的词语对文献主题的贡献最大。因此，对于区域位置的计算，设为根据分词位置标记的位置值，本文将文本位置区别为3种情况：标题中设置为10.0；摘要或首个段落设为5.0；其他段落设置为3.0。若词语在各个位置都出现，则取其最高值。在分词处理时，就可标记词语的位置值，然后使用规约化公式（2）进行处理，归约化是使数值在(0，1]范围内，能够消除指标之间的量纲影响，解决数据指标之间的可比性。位置值经过数据归约化处理后，该指标处于同一数量级，适合进行综合对比评价。

 (2)

### （3）词距

词距是指文本中词语首次出现的位置和最后一次出现的位置之间的距离。随着文本长度的增加，局部关键词可能影响文本整体关键词。因此为了滤除局部关键词，本研究将词距引入关键词的抽取算法中。根据以往实验分析，采用如下公式来计算候选词的词距[8]：

 (3)

为该分词第一次出现位置序号，为该分词最后一次出现位置序号，为经分词后的文本中候选词的总数。

### （4）训练调节因子

在确定了计算3个特征项值的公式后，需要考虑如何确定调节因子，使其更加合理地反映各个因素对权重的贡献程度。对每种类型的商品分别进行训练，能得出适合本类商品的调节因子。本研究采用机器学习最小均方误差(LMS)训练法训练公式的调节因子[9]，确定调节因子的学习过程如图6所示。学习完成后，便得到合适的调节因子。



图4 确定调节因子的学习过程

### （5）关键词列表输出

确定了后，使用公式(1)进行计算而得出候选词的权重，然后根据关键词权重大小对词语进行排序，最终得到物品-评价关键词表如表1所示，用于下一步商品相似度计算。

表1 关键词提取示例

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 商品ID | 商品名称 | Top5关键词 | | | | |
| 1 | Ayuko 拼色套头毛衣 | 洋气 | 厚实 | 宽松 | 柔软 | 划算 |
| 2 | 小葱良裁 针织开衫 | 好看 | 舒服 | 偏薄 | 舒服 | 有点贵 |
| 3 | 日系插肩袖棒球服 | 偏厚 | 宽松 | 温柔风 | 厚实 | 显胖 |
| 4 | Boolibyoo 复古大码学院风菱形格纹针织裙 | 显胖 | 学院风 | 弹力大 | 暖和 | 面料 |
| 5 | EM八月阔腿牛仔裤 | 宽腿裤 | 实惠 | 复古 | 腿长 | 百搭 |

**6.基于评论关键词的推荐算法**

基于物品的协同过滤算法（ItemCF）指的是根据用户喜欢的物品，推荐这些物品的相似物品[9]。此算法计算物品的相似度不是分析物品的内容特征，而是根据物品-用户的评分矩阵，认为喜欢物品A的用户都会喜欢物品B。

**6.1基于关键词的协同过滤算法**

基于物品的协同过滤算法也分为两步：

1. 计算物品之间的相似度。

一般使用如下公式计算物品间的相似度，

 (4)

N(i)表示喜欢物品i的用户列表，N(j)表示喜欢物品j的用户列表。考虑到该公式会造成任何物品都和热门物品有很大的相似性。为避免推荐出热门物品可以用下面公式：

 (5)

其中

1. 计算用户对物品的潜在喜爱程度。

 (6)

其中S(j，k)是和物品j相似的K个物品的集合。代表是物品i和j的相似程度。一般适用隐式反馈原则，即如果用户对物品i有行为，则的值为1，反之则的值为0。

根据上文处理过的数据集，计算物品之间的相似性，计算出物品之间的相似度之后，基于物品的协同过滤推荐算法将根据用户的高分物品，寻找其高度相似的物品，从而得到 TOP-N推荐结果集[10-11]，预测兴趣度公式如： (7)

其中N(u)是用户喜欢的物品合集，S（j，k）是和物品j最相似的k个物品集合。其根本意义就是越和用户的历史喜欢物品相似的，越有可能是用户潜在的喜欢的物品。但引入关键词后，在传统兴趣度公式的基础上，引入物品评论相同关键词个数作为参数，重新构建相似度公式：

 (8)

其中，是两个商品排名前N的关键词中重复的个数。

**6.2算法评估指标**

准确度P和召回率R是评价一个推荐算法性能的重要指标[12]。

 (9)

 (10)

构建混淆矩阵将样例数据划分为真正类样本、负类样本、预测正类、预测负类四种情况如表2所示。

表2 混淆矩阵

|  |  |  |
| --- | --- | --- |
|  | 正类样本 | 负类样本 |
| 预测正类 | TP | FP |
| 预测负类 | FN | TN |

结合购物网站实际场景，准确度和召回率又可以用下面公式表示。

 (11)

 (12)

其中R(u)是用户u推荐的物品集合，T(u)是用户u在测试集上的物品集合。准确度和召回率越高表示推荐系统越好。

**6.3算法效率分析**

本文采用数据源于阿里旗下天池大数据平台下数据集Women's E-Commerce Clothing Reviews，该数据集是女装电子商务评论数据集，包括23486行和10个特征变量[13]，将数据随机分为5份，其中4份为训练集和1份作为测试集。当N取值为5，6，7，8，9，10时与传统的ItemCF协同过滤算法相进行对比实验，结果如图5所示。

图5 准确率召回率对比图

由结果可知，本文提出的推荐模型相对于传统的基于商品的协同过滤算法具有较大的优势，在一定程度上提高了推荐的准确率。随着关键字的增加，模型推荐结果的准确率和召回率逐渐增加，但是当关键词数量增加到一定程度后，对算法的性能的影响不再增大。当选取7个关键字进行计算时，算法的性能最佳。

**7结束语**

本文对电子商务平台的推荐算法进行了深入的研究，在传统的协同过滤算法的基础上，设计了基于评论关键词的协同过滤推荐算法，在所选数据集上达到了很好的效果。但当评论数过少时，依然存在冷启动问题，未来可考虑将卷积神经网络融合在模型中，研究是否可以进一步解决协同过滤系统中冷启动商品的问题。

**参考文献**

[1] 赵俊逸等. 协同过滤推荐系统综述[J].信息安全学报. Vol.6 No. 5 September 2021:17-19

[2] 郑建兴等.基于评论文本情感注意力的推荐方法研究[J], 郑州大学学报(工学版). 2022年02期 第44-50+57页

[3] GOLDBERGD, NICHOLSDA, OKIBM, et al. Using collaborative filtering to weave an information TAPESTRY[J]. Communications of the ACM,1992,35(12):60-70

[4] VERBERT, MANOUSELIS N, OCHOA X, et al. Context Aware Recommender Systems for Learning: A Survey and Future Challenges[J].IEEE Transactions on Learning Technolo-gies,2012(4):318-335.

[5] SHAMBOUR Q, LU J. An effective recommender system by unifying user and item trust information for B2B applications[J]. Journal of Computer and System Sciences, 2015,81(7):1110- 1126.

[6] 薛美琦.基于协同过滤技术的物品选购推荐系统设计[D].电子科技大学.2019.03

[7] 江水. 基于协同过滤技术推荐系统的探究[J].计算机技术与发展. Vol.31 No.11 Nov. 2021:1-7

[8] 马慧芳.李苗等. 基于通配符模式与随机游走的关键词提取方法[J].计算机工程. 2020年07期: 78-83

[9] 朱磊.胡沁涵等.基于评分偏好和项目属性的协同过滤算法[J].计算机科学. 2020年04期 第67-73页

[10] Ricci F, Rokach L, Shapira B. Recommender systems: introduction and challenges[M].Recommender

systems handbook. Springer, Boston, MA, 2015: 1-34.

[11] 刘文佳,张骏.改进的协同过滤算法在电影推荐系统中的应用[J].现代商贸工业,2018,39(17):59-62.

[12] 朱育颉.刘虎沉.网上购物平台多推荐融合算法研究[J].计算机科学. 2021年S2期第232-235

[13] Women's E-Commerce Clothing Reviews 女士电子商务服装的评测[EB/OL]. <https://tianchi.aliyun.com/dataset/dataDetail?dataId=93847>

[7] 杨佳莉, 李直旭, 许佳捷, 赵朋朋, 赵雷, 周晓方. 一种自适应的混合协同过滤推荐算法[J]. 计算机工程, 2019, 45(7): 222-228.

YANG Jiali, LI Zhixu, XU Jiajie, ZHAO Pengpeng, ZHAO Lei, ZHOU Xiaofang. An Adaptive Hybrid Collaborative Filtering Recommendation Algorithm[J]. Computer Engineering, 2019, 45(7): 222-228.