文章编号：2095-2163(2024)02-0190-05 中图分类号：TP391.1 文献标志码：A

基于 LDA 主题模型的协同过滤推荐算法

张 宇，吴 静
（浙江理工大学 计算机科学与技术学院，杭州 310018）

摘 要：传统的协同过滤推荐算法直接根据用户对物品的评分进行推荐，忽略了评论文本中隐含的重要信息，当用户对物品的评论较少时，由于数据的稀疏性会造成推荐效果的不准确和单一。本文提出了一种基于 LDA 主题模型的协同过滤推荐算法 LDA-CF（Latent Dirichlet Allocation model-LDA-Collaborative Filtering），在传统的协同过滤算法基础上，通过 LDA 模型对评论文本中的主题进行分类，从各个主题层面挖掘用户的情感偏好，计算用户之间的相似度，进而向目标用户推荐商品。对京东平台牙膏的评论数据集的实验结果表明，该算法不仅可以缓解由于评分数据较少造成的稀疏性问题，推荐的精确度也有所提高。

关键词：协同过滤；推荐算法；LDA；评论文本

Collaborative filtering recommendation algorithm based on LDA topic model

ZHANG Yu, WU Jing
(School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Traditional collaborative filtering recommendation algorithms tend to recommend items directly according to users' scores, ignoring the important information implied in the comment text. Moreover, when users have few comments on items, the sparsity of the data will lead to the inaccuracy and singleness of the recommendation effect. Therefore, this paper proposes a collaborative filtering recommendation algorithm based on LDA topic model. Based on the traditional collaborative filtering algorithm, the algorithm classifies the topics in the review text through the LDA model, mines the emotional preferences of users from each topic level, calculates the similarity between users, and then recommends products to target users. The experimental results based on the review data set of toothpaste on JD platform show that the algorithm can not only alleviate the sparsity problem caused by few score data, but also improve the recommendation accuracy compared with the traditional collaborative filtering algorithm.

Key words: collaborative filtering; recommendation algorithm; LDA; comment text

0 引 言

随着互联网的飞速发展，海量数据的产生使得用户找到自己需要的内容十分艰难。为解决这一难题，提出了推荐算法[1]。常见的推荐算法分为基于内容的推荐算法、基于协同过滤的推荐算法以及混合推荐算法[2]。传统的推荐算法往往只分析用户对商品的共同评分，根据用户的历史行为记录进行推荐，如果用户之间共同评分的商品较少，推荐的性能就会受到数据稀疏性的影响，从而影响推荐的精度[3]。

传统的推荐算法是通过对商品的评分计算用户之间的相似度，而评分只能反映用户对商品的总体满意程度，不能准确反映用户对商品各个属性的满意度。以牙膏评论为例，用户 A 和 B 对同一款牙膏满意度打分均为 4 分，用户 A 认为该款牙膏的使用效果好而价格昂贵；用户 B 认为这款牙膏的价格合适而使用效果不明显。用户 A 和用户 B 对同一款牙膏的评分相同，但是他们喜爱偏好却不同。由此可见，从用户对商品的文本评论中可以挖掘出更有价值的信息，对这些文本评论加以利用，可以准确分析出用户的喜爱偏好，从而提高推荐的准确度。

本文对评论文本采用 LDA 主题模型进行分析，深刻挖掘用户对商品各个方面的喜爱程度，每个商品都包含各个方面的主题，如：牙膏包括价格、使用效果、味道等，通过对各主题的分析来预测用户对商品的总体评分，从而找到与目标用户最相似的用户进行推荐。

1 相关工作

1.1 LDA 主题模型

LDA 主题模型由 Blei 等于 2003 年提出，LDA 模

作者简介：吴 静(1997-)，女，硕士研究生，主要研究方向：自然语言处理。
通讯作者：张 宇(1982-)，女，博士，副教授，主要研究方向：数据挖掘、自然语言处理。Email：yzh@ zstu.edu.cn

收稿日期：2023-02-17

型是一种主题概率生成模型，构建“文档-主题-词”3层的贝叶斯结构，文档是词汇的集合，每篇文档都会有一个或者多个主题，每个主题会以一定概率选择某个词，词会以一定的概率生成某个主题[4-7]。近年来，很多研究者将 LDA 模型和推荐系统相结合，Liu 等[8]充分利用评论文本信息，通过 LDA 模型观察由丰富文档组成的本地上下文，这些文档可能直接或间接地影响目标文档的主题分布；Zhou 等[9]提出评级 LDA 模型，认为用户行为不是独立的，还受相似用户的影响，相似用户给出的评分高，则目标用户也有可能喜爱该商品；Huang 等[10]对 LDA 主题模型扩展了文本特征的数量，基于支持向量机（SVM）、随机森林（RF）等算法构建文本分类器，并通过十倍交叉和混淆矩阵验证情感分类方法的有效性。

1.2　推荐系统

推荐系统的探索源于 20 世纪 90 年代初，综合了诸多领域的知识，如信息检索、预测结果、数据存储以及市场分析等[11-13]。推荐系统是非常有用的工具，随着用户、服务和在线数据的规模迅速扩大，可以在用户购买产品之前提供适当的建议[14]。一个高效的有价值的推荐系统，要解决在推荐过程中的推荐精准度问题、冷启动问题以及大规模的计算与存储等问题[15]，Lee 等[16]研究发现，在推荐算法中引入社会关系和历史行为等，可以更好地为推荐系统服务，由于社交网络中存在同伴影响或共同兴趣等隐性因素，具有相似隐性因素的用户更有可能成为目标推荐的相似对象。随着研究的不断深入，陆续产生了基于关联规则挖掘的推荐系统、基于贝叶斯分类的推荐系统、个性化推荐服务等[17]。

2　推荐模型

2.1　数据预处理

获取用户对商品的评论文本，按照如下步骤对评论文本进行预处理：首先，对评论文本去除同一用户短时期内的重复商品评论；其次，由于较短的评论文本所包含的信息较少，为保证推荐结果的准确性，去除评论文本字数少于 5 个的商品评论；最后，去除评论文本中完全没有用或者没有意义的词，如助词、拟声词、虚词等，使用 jieba 分词进行中文分词，得到数据集。

2.2　构建 LDA 主题模型

将一个评论文本视为 LDA 模型中的一个文档进行分析，所有的评论文本视为文档集合。

LDA 模型的生成过程：从狄利克雷分布 α 中取样生成文档 p 的主题分布 θ_p[18]；从主题的多项式分布 θ_p 中取样，生成文档 p 第 q 个词的主题 $Z_{p,q}$；从狄利克雷分布 β 中取样，生成主题 $Z_{p,q}$ 对应的词语分布 $\varphi_{p,q}$；从词语的多项式分布 $\varphi_{p,q}$ 中采样，最终生成词语 $W_{p,q}$。其中，参数 α 和参数 β 根据 *Gibbs* 采样方法进行参数估计[19]，最佳主题数目 K 的值根据困惑度的大小来确定，困惑度越小，主题数目 K 越合适。

2.3　商品评分计算

在主题-词汇矩阵中，主题下只有部分词汇具有情感极性，采用知网情感词典判断每个主题下包含的词汇是否为正面情感词、负面情感词和中性词[20-21]，若主题每包含一个正面情感词，则将该主题的情感得分加 1；每包含一个负面情感词，则将该主题的情感得分减 1；中性词不计入情感得分，依次类推直至统计完所有词汇，得到每个主题对应的情感得分。

在文档-主题矩阵中，每个文档所包含各个主题的概率不同，每个文档的评分也不同，用户 u_i 对商品 x_h 的评分 s'_{u_i,x_h}，计算公式(1)：

$$s'_{u_i,x_h}=\sum_{k=0}^{K} g^k_{u_i,x_h}\times p^k_{u_i,x_h} \tag{1}$$

其中，$1\leqslant i\leqslant N$，N 表示用户总数；$1\leqslant h\leqslant H$，H 表示商品总数；$k(k=0,1,2,3,\cdots,K)$ 表示文档-主题矩阵中文档的第 k 个主题；K 表示文档所包含的主题数目；$g^k_{u_i,x_h}$ 表示用户 u_i 对商品 x_h 的评论文本在第 k 个主题上的情感得分；$p^k_{u_i,x_h}$ 表示用户 u_i 对商品 x_h 的评论文本包含主题 k 的概率。

通过最大最小标准化公式使得用户对商品的评分在[1,5]之间，标准化后的用户 u_i 对商品 x_h 的评分 s_{u_i,x_h}，计算公式(2)：

$$s_{u_i,x_h}=\frac{s'_{u_i,x_h}-s'_{u_i,\min}}{s'_{u_i,\max}-s'_{u_i,\min}}\times 5 \tag{2}$$

其中，$s'_{u_i,\min}$ 表示用户 u_i 对已购买商品的评分最小值，$s'_{u_i,\max}$ 表示用户 u_i 对已购买商品的评分最大值。

计算结果四舍五入取整，不足 1 分的记为 1 分，情感值范围为 1-5 分，5 分为完全满意，1 分为完全不满意，分数越高表示用户对商品越满意，空缺值表示用户未对该商品进行打分，最后得到用户对商品的评分矩阵。

根据用户对所有商品的评分，计算用户 u_i 对全部已评价商品评分的均值 $\bar{s}_{u_i}$，计算公式(3)：

$$\bar{s}_{u_i}=\frac{1}{H}\sum_{h=1}^{H} s_{u_i,x_h} \tag{3}$$

2.4 用户相似度计算

使用余弦相似度方法计算两个用户之间的相似度。$sim(u_i,u_j)$ 表示用户 u_i 和用户 u_j 的相似度,计算公式(4):

$$sim\ (u_i,u_j)=\frac{\sum_{h=1}^{H} s_{u_i,x_h}\times s_{u_j,x_h}}{\sqrt{\sum_{h=1}^{H}(s_{u_i,x_h})^2}\times\sqrt{\sum_{h=1}^{H}(s_{u_j,x_h})^2}} \tag{4}$$

其中,s_{u_i,x_h} 表示用户 u_i 对商品 x_h 的评分,s_{u_j,x_h} 表示用户 u_j 对商品 x_h 的评分。

按照相似度数值排序,数值越大说明两个用户之间的相似度越高,按照相似度大小对所有用户降序排列,即越靠前的用户同目标用户之间的相似度越高,得到目标用户的相似用户群。

2.5 生成推荐商品

根据 Top - N 策略从相似用户群中选取排名前 N 个用户作为最近邻用户,根据最近邻用户预测目标用户对待推荐商品的评分,根据预测结果,生成向目标用户推荐的商品。g_{u_o,x_h} 表示目标用户 u_o 对待推荐商品 x_h 的预测评分,公式(5):

$$g_{u_o,x_h}=\frac{\sum_{u_m\in M}(sim(u_o,u_m)\times(s_{u_m,x_h}-\bar{s}_{u_m}))}{\sum_{u_m\in M}sim(u_o,u_m)} \tag{5}$$

其中,M 表示相似用户群中对待推荐商品 x_h 进行了评分的所有用户的集合;$sim(u_o,u_m)$ 表示目标用户 u_o 和用户 u_m 的相似度;s_{u_m,x_h} 表示用户 u_m 对待推荐商品 x_h 的评分;$\bar{s}_{u_m}$ 表示用户 u_m 对其全部已评价商品评分的均值。

3 实验与分析

3.1 实验数据集

本文实验使用的初始数据集为京东 11 个排名较高的牙膏品牌的评论文本,数据集共包括 5 089 个用户的 103 850 条商品评论。对数据进行预处理,可用数据集数包括 4 870 个用户的 94 815 条商品评论,每条评论文本平均字数为 56 个,按 8∶2 的比例将数据集随机地分为训练集和测试集。

3.2 对比实验与评估指标

为验证推荐方法的准确性,将 LDA-CF 算法与以下两种传统的推荐算法进行比较:

(1) CB(Content-Based Recommendations CB):基于内容的推荐算法。

(2) CF(Collaborative Filtering):传统的协同过滤推荐算法。

本文采用的评价指标:

(1)平均绝对值误差(*MAE*):反映推荐算法预测评分与实际评分的相似程度,公式(6):

$$MAE=\frac{\sum_{(i,j)\in E^U}|u_{i,j}-u'_{i,j}|}{|E^U|} \tag{6}$$

其中,$u_{i,j}$ 表示用户 i 对商品 j 的实际评分;$u'_{i,j}$ 表示用户 i 对商品 j 的预测评分;$|E^U|$ 表示预测评分总数。

(2) $F1$ - Score:综合了分类模型的精确率 P 和召回率 R,公式(7) ~ 公式(9):

$$P=\frac{TP}{TP+FP} \tag{7}$$

$$R=\frac{TN}{TN+FN} \tag{8}$$

$$F1-Score=\frac{2PR}{P+R} \tag{9}$$

其中,TP 表示将文本正向样本预测为正向样本;FN 表示将文本正向样本预测为负向样本;TN 表示将文本负向样本预测为负向样本;FP 表示将文本负向样本预测为正向样本。

3.3 实验结果与分析

采用 LDA 主题模型对用户的评论数据进行分析,通过 Gibbs 采样的方法对参数进行估计,设置 Dirichlet 先验参数 $\alpha=50/T$ 和 Dirichlet 先验参数 $\beta=0.01$,依据困惑度的方法来确立最佳主题数 K,通过 模型训练,得出文档 - 主题矩阵和主题 - 词汇矩阵。主题数目 K 和困惑度的关系如图 1 所示,可以看出当主题数 $K=7$ 时, LDA 主题模型的困惑度最小。

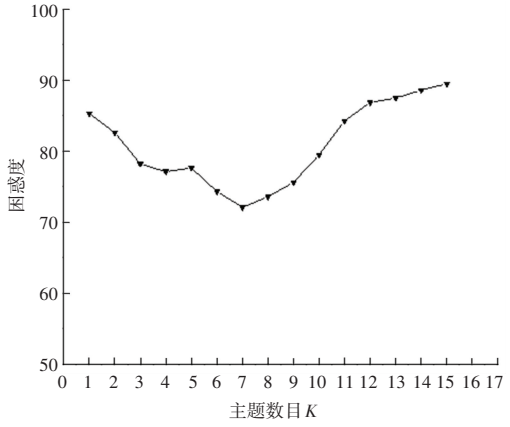

图 1 主题数目 *K* 和困惑度的关系

Fig. 1 Relationship between the number of topics *K* and the degree

of confusion

LDA 主题模型得到的部分用户评论数据的主题-词汇概率分布见表 1。

表 1　部分用户评论数据的主题-词汇概率分布

Table 1　Topic-word probability distribution of partial user review data

主题	词汇	概率
主题 1	不错,快递,物流速度快,方便快捷	0.428,0.040,0.031,0.027
主题 2	品牌,信赖,牌子,老字号	0.354,0.121,0.087,0.401
主题 3	回购,质量,实惠,物超所值	0.272,0.120,0.085,0.046
主题 4	包装,精美,物流完好,完好无损	0.149,0.090,0.074,0.067
主题 5	第一次,别人,朋友,好用	0.186,0.103,0.081,0.063
主题 6	味道,薄荷,清新,好闻	0.243,0.076,0.071,0.061
主题 7	活动,凑单,赠品,划算	0.223,0.103,0.090,0.082

LDA 主题模型得到的部分用户评论数据的文档-主题概率分布见表 2。

表 2　部分用户评论数据的文档-主题概率分布

Table 2　Document-topic probability distribution of partial user review data

文档	主题	概率
文档 1	[7,3,5,6,4,2,1]	0.355,0.270,0.196,0.072,0.006,0.005,0.004
文档 2	[5,7,1,6,3,4,2]	0.287,0.137,0.074,0.063,0.005,0.004,0.003
文档 3	[3,7,6,5,4,2,1]	0.201,0.153,0.009,0.009,0.006,0.005,0.003
文档 4	[4,2,7,6,5,3,1]	0.535,0.419,0.006,0.005,0.004,0.002,0.001
文档 5	[6,2,1,5,4,3,7]	0.304,0.296,0.170,0.087,0.066,0.056,0.005

LAD-CF 算法先通过 LDA 主题模型得出用户评论的文档-主题概率分布和主题-词汇概率分布，得出用户的评分。LDA-CF 算法与传统的协同过滤算法和基于内容的推荐算法推荐效果如图 2 所示，可见相比其他两种方法，LDA-CF 算法的 *MAE* 值较小，在评分预测的准确性优于其他两种算法。

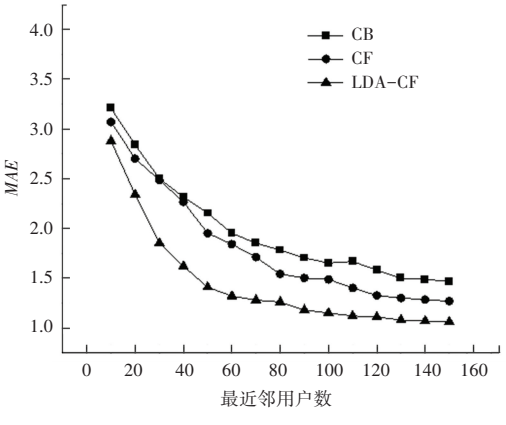

图 2　不同模型的 MAE 值对比

Fig. 2　Comparison of MAE values of different models

利用余弦相似度的方法得出目标用户的待推荐商品，将本文的 LDA-CF 算法与基于内容的推荐算法(CB)和传统的协同过滤推荐算法(CF)进行对比实验，实验结果见表 3，可见在同等条件和数据下，LDA-CF 在各项指标上均有所提高，推荐较为准确。

表 3　评论数据集评测结果对比

Table 3　Comparison of evaluation results of review data sets

模型	准确率	召回率	*F*1 - *Score* 值
CB	0.867	0.847	0.857
CF	0.875	0.889	0.882
LDA-CF	0.895	0.914	0.904

4　结束语

本文提出一种基于 LDA 主题模型的协同过滤推荐算法，使用 LDA 模型获取评论文本的信息，并将文本主题与评论文本相融合，利用协同过滤算法得出用户对商品的评分。与传统的推荐算法相比，本文提出的 LDA-CF 算法能够充分利用评论文本包含的信息，更加深刻地分析挖掘用户在商品各个主题下的喜爱偏好，从而提高了推荐的准确性。在后续的研究中，将会尝试提高文本主题提取的精度，从而更加精准的分析出评论文本包含的主题信息，进一步提高推荐算法的精准度。

参考文献

[1] NASSARN, JAFARA, RAHHAL Y. A novel deep multi-criteria collaborative filtering model for recommendation system [J].

Knowledge-Based Systems,2020(1):11-17.
[2] Prasad RVVSV. A Categorical Review of Recommender Systems [J]. International Journal of Distributed and Parallel Systems, 2012,3(5):108-119.
[3] YU Chengyuan, HUANG Linpeng. CluCF: A clustering CF algorithm to address data sparsity problem[J]. Service Oriented Computing and Applications,2017,11(1):33-45.
[4] SAFI'IE M A, UTAMI E, FATTA H A. Latent Dirichlet Allocation (LDA) model and kNN algorithm to classify research project selection[J]. IOP Conference Series: Materials Science and Engineering,2018,333(1):49-70.
[5] 王李冬,魏宝刚,袁杰. 基于概率主题模型的文档聚类[J]. 电子学报,2012,40(11):2346-2350.
[6] Venugopalan Manju, Gupta Deepa. An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis [J]. Knowledge - Based Systems,2022,5(1):108-668.
[7] ZHENG Wei, GE Bin, WANG Chishe. Building a TIN-LDA Model for Mining Microblog Users´ Interest[J]. IEEE Access, 2019,7(1):21795-21806.
[8] LIU Yang, XU Songhua. A local context-aware LDA model for topic modeling in a document network [J]. Journal of the Association for Information Science and Technology, 2017, 68 (6):1429-1448.
[9] ZHOU Xiuze, WU Shunxiang. Rating LDA model for collaborative filtering[J]. Knowledge-Based Systems,2016,110(5):135-143.
[10] HUANG Yanrong, WANG Rui, HUANG Bin, et al. Sentiment classification of crowdsourcing participants' reviews text based on LDA topic model[J]. IEEE ACCESS,2021,9(7):1921-1937.
[11] 黄勃,严非凡,张昊,等. 推荐系统研究进展与应用[J]. 武汉大学学报(理学版), 2021, 67(6): 503-516. DOI: 10.14188/j.1671-8836.2021.1001.
[12] CHEN Rui, HUA Qingyi, CHANG Yanshuo, et al. A survey of collaborative filtering - based recommender systems: from traditional methods to hybrid methods based on social networks [J]. IEEE Access,2018,6(10):1036-1055.
[13] Devdatta Godbole, Manish Narnaware. A survey on personalized service recommendation systems [J]. International Journal of Engineering Research and Technology,2016,5(2):616-620.
[14] Nachiket Sadashiv Bhosale, Sachin S Pande. A survey on recommendation system for big data applications[J]. Data Mining and Knowledge Engineering,2015,7(1):42-44.
[15] 罗婷予,Miguel Baptista Nunes. 从用户视角理解智能推荐系统[J]. 数字图书馆论坛,2019,3(10):30-36.
[16] LEE Yan-Li, ZHOU Tao, YANG Kexin, et al. Personalized recommender systems based on social relationships and historical behaviors[J]. Applied Mathematics and Computation, 2023, 43 (7):82-100.
[17] Gabroveanu Mihai. Recommendation system based on association rules for distributed E-learning management systems[J]. ACTA Universitatis Cibiniensis,2015,67(1):90-104.
[18] 万志成,郑静. 基于狄利克雷过程高斯混合模型的变分推断[J]. 杭州电子科技大学学报(自然科学版),2021,41(5):54-61.
[19] 凤维明,尹一通. 分布式采样理论综述[J]. 软件学报,2022,33 (10):3673-3699.
[20] LIU W Y, XIAO B S, WANG T , et al. Building Chinese sentiment lexicon based on howNet [J]. Advanced Materials Research,2011,1198(187-187):405-410.
[21] LIU L, LEI M, WANG H. Combining domain-specific sentiment lexicon with hownet for Chinese sentiment analysis[J].Journal of Computers,2013,8(4):878-883.

(上接第 189 页)

表 3 实验结果

Table 3 Experimental results

模型	精确度	召回率	*F*1 - *score*
本模型	90.54	92.87	91.69
BiLSTM	90.16	92.84	91.03
RNN	87.17	86.59	86.59
CNN	85.15	84.13	84.71
LSTM	88.46	88.07	86.06

4 结束语

本文针对影评提出了基于星级权重和双向长短期记忆网络的神经网络模型,能够解决单一特征无法充分利用文章上下文信息的问题,改善影评情感偏向不明显的情况,从而能提高了分类准确率。

参考文献

[1] 陈晓东. 基于情感词典的中文微博情感倾向分析研究[D]. 武汉:华中科技大学,2012.
[2] 尹宝才,王文通,王立春. 深度学习研究综述[J]. 北京工业大学学报,2015,41(1):48-59.
[3] 庞亮,兰艳艳,徐君,等. 深度文本匹配综述[J]. 计算机学报, 2017,40(4):985-1003.
[4] 唐明,朱磊,邹显春. 基于 Word2Vec 的一种文档向量表示[J]. 计算机科学,2016,43(6):214-217,269.
[5] 黄磊,杜昌顺. 基于递归神经网络的文本分类研究[J]. 北京化工大学学报(自然科学版),2017,44(1):98-104.
[6] TAI K S,SOCHER R,MANNING C D. Improved semantic representations from tree-structured long short-term memory networks[J]. arXiv preprint arXiv:1503.00075, 2015.
[7] BAZIOTIS C, PELEKIS N, DOULKERIDIS C. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis[C]//Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017: 747-754.