分类号 F426.471　　密级 公开
U D C　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于在线评论情感分析的农产品个性化推荐研究

研 究 生 姓 名：李佳儒
指导老师姓名、职称：王玉珍 教授
学 科、专 业 名 称：管理科学与工程
研 究 方 向：电子商务

提 交 日 期：2021 年 5 月 20 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：李佳陈 签字日期：2021.5.20

导师签名：王玉刚 签字日期：2021.5.20

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：李佳陈 签字日期：2021.5.20

导师签名：王玉刚 签字日期：2021.5.20

Research on Personalized Recommendation of Agricultural Products Based on Sentiment Analysis of Online Reviews

Candidate: Li Jiaru

Supervisor: Wang Yuzhen

摘 要

近年来，随着电子商务的快速发展，网上购买农产品作为购物的方式之一，深受消费者喜爱。然而，由于电商平台的农产品种类繁多，用户需要花费更多的时间和精力去寻找自己喜欢的产品，降低了购物效率，影响了购物体验。因此，对于电商平台来说，依据消费者的喜好，针对性的推荐就显得非常重要。传统的农产品推荐大多基于用户对所购买产品的评分推断用户兴趣，进而进行推荐，却忽略了产品在线评论对推荐效果的影响。而在线评论蕴含着大量用户对产品特征个人喜好方面的信息，对个性化推荐极为重要。鉴于此，本文通过对国内知名电商平台的在线评论进行分析，以甘肃省特色农产品为研究对象，构建了基于情感分析的个性化推荐模型。该模型在情感分析基础上，利用矩阵分解、改进巴氏系数相似度以及混合协同过滤算法对传统协同过滤算法进行了优化，进而提升推荐的准确性。概括起来，本文的主要工作如下：

（1）拓展了农产品领域的情感词典，并基于情感词典计算用户评论的情感值，得出情感评分矩阵。在现有情感词典的基础上，利用情感倾向点互信息 SO-PMI 算法和 LDA 主题模型对农产品领域的情感词典进行拓展，并加入网络词词典进行完善，以此计算出用户评论的情感值，从而得出情感评分矩阵。

（2）构建了融合矩阵分解和改进巴氏系数的混合推荐算法。首先，利用基于偏置的矩阵分解 BMF 算法对评分矩阵进行缺失值填充；其次，对传统巴氏系数相似度度量方法进行改进；再次，对基于用户和基于项目的两种协同过滤推荐算法进行融合；最后，将 BMF 算法、改进巴氏系数以及混合协同过滤算法融合得到混合推荐算法，并在数据集上验证了算法的准确性。

（3）构建了基于情感分析的农产品个性化推荐模型。基于词典的情感分析得到情感评分矩阵，在此基础上，融合混合推荐算法，建立基于情感分析的个性化推荐模型，并与 UCF、ICF、HCF 三种推荐模型对比，验证模型的有效性。

（4）设计了农产品个性化推荐系统。将基于情感分析的个性化推荐模型应用在推荐系统中，从而为用户提供个性化的农产品推荐。

关键词：情感分析 个性化推荐 BMF 算法 改进巴氏系数 混合协同过滤算法

Abstract

In recent years, with the rapid development of e-commerce, buying agricultural products online has been deeply loved by consumers as a way of shopping. However, due to the wide variety of agricultural products on e-commerce platforms, users need to spend more time and energy to find their favorite products, which affects the shopping experience. Therefore, for e-commerce platforms, targeted recommendations are very important according to consumers' preferences. Traditional agricultural product recommendation is mostly based on the user’s rating of the purchased product to infer user interest, thereby making recommendations, but ignores the impact of product online reviews on the recommendation effect. Online reviews contain a large number of users' personal preferences for product features, which are extremely important for personalized recommendations. In view of this, this article analyzes the online reviews of well-known domestic e-commerce platforms, takes Gansu Province's characteristic agricultural products as the research object, and constructs a personalized recommendation model based on sentiment analysis. The model is based on sentiment analysis, using matrix decomposition, improved Bhattacharyya coefficient similarity and hybrid collaborative filtering algorithm to improve the traditional collaborative filtering algorithm, and then improve the accuracy of recommendation. In summary, the

main work of this article is as follows:

(1) The sentiment dictionary in the field of agricultural products is expanded, and the sentiment score matrix of user reviews is calculated based on the sentiment dictionary. On the basis of the existing sentiment dictionary, the sentiment dictionary in the field of agricultural products is expanded by using the sentiment point mutual information algorithm and the LDA topic model, and the network word dictionary is added to improve it, so as to calculate the sentiment value of the user's comment, and then obtain sentiment score matrix.

(2) A hybrid recommendation algorithm combining matrix decomposition and improved Bhattacharyya coefficient is constructed. Firstly, use the bias-based matrix factorization algorithm BMF to fill in the missing values of the score matrix; secondly, improve the traditional Bhattacharyya coefficient similarity measurement method; thirdly, integrate the two collaborative filtering recommendation algorithms based on user and item ; Finally, the BMF algorithm, improved Bhattacharyya coefficient and hybrid collaborative filtering algorithm are combined to obtain a hybrid recommendation algorithm, and the accuracy of the algorithm is verified on the MovieLens data set.

(3) A personalized recommendation model of agricultural products based on sentiment analysis is constructed. Based on the sentiment analysis of the dictionary, the sentiment score matrix is obtained. On this

basis, a hybrid recommendation algorithm is combined to establish a personalized recommendation model based on sentiment analysis, and the superiority of the model is verified by comparison with UCF, ICF, and HCF three recommendation models.

（4）A personalized recommendation system for agricultural products is designed. The personalized recommendation model based on sentiment analysis is used in the recommendation system to provide users with personalized recommendation of agricultural products.

Keywords: Sentiment analysis；Personalized recommendation；BMF algorithm；Improved Bhattacharyya coefficient；Hybrid collaborative filtering algorithm

目 录

1 绪论

1.1 研究背景及意义

1.1.1 研究背景

近年来，随着电子商务的快速发展，农产品的销售也逐渐从线下扩展到线上，上网购买农产品深受消费者的喜爱，农产品电子商务的交易规模不断扩大。根据中国食品（农产品）安全电商研究院发布的报告显示，2019 年阿里农产品上行达到 2000 亿元，拼多多 1364 亿元，京东达到 1000 亿元，苏宁达到 500 亿元，中国邮政 30 亿元。从 2019 年的数据来看，国内各大农产品电商平台的竞争愈发激烈。伴随着农产品电商平台交易规模的增大，其农产品的种类越来越繁多，用户需要花费更多的时间和精力去寻找自己喜欢的产品，影响了购物体验。因此，电商平台的推荐技术应运而生。而目前电商平台对农产品的推荐普遍基于用户的浏览记录和对所购买产品的评分推断用户兴趣，从而进行推荐。但是，不同的用户之间对相同产品的浏览记录以及同样评分未必代表他们对产品的兴趣相同，会影响推荐的准确性。因此，为了提升用户的购物体验，同时提高电商平台的服务质量，进行个性化推荐显得极为重要。

要进行个性化推荐，用户兴趣偏好的挖掘是关键。随着电子商务和社会媒体的迅猛发展，用户往往通过在线评论的方式来表达个人对产品交易过程的情绪与观点。其他用户在购买产品时大多会参考评论信息，其购买行为以及后续对产品的体验和评价都会直接或者间接地受到评论的影响。这些评论文本蕴含着大量用户的兴趣偏好，对其进行深入挖掘与分析，有利于平台的个性化推荐。然而，在线评论文本由于并非销量、价格等结构化数据，无法直接进行加工处理，如何有效、准确地整合、处理这些非结构化的文本数据并从中提取出用户对产品的情感偏好是一个挑战。文本情感分析是分析用户对于产品及其属性所表达的观点、情感、评价、态度和情绪的方法，其结果不仅可以帮助人们做出购买决策，还可以帮助企业了解用户对产品的真实感受和体验。所以，基于情感分析方法挖掘用户在每一个产品上的情感态度，对产品的个性化推荐起到一定的促进作用。因此，本文首先对国内知名电商平台的在线评论进行研究，构建了基于情感分析的个性化推荐模型，即基于词典的情感分析方法计算用户评

论的情感值，得到情感评分矩阵；在此基础上，融合 BMF 算法、改进巴氏系数相似度以及混合协同过滤算法对传统协同过滤算法进行优化，从而达到较好的推荐效果。

1.1.2 研究意义

推荐技术是解决网络信息过载问题的有效技术之一，通过分析用户的历史行为数据，发现用户的兴趣爱好，并根据用户的兴趣变化调整信息服务的方式和内容，将用户感兴趣的信息推荐给用户。缺点是现有的电商平台推荐主要是利用用户的历史浏览数据以及星级评分进行推荐，反映不出用户对产品各个特征的喜好，故推荐结果并不能满足用户的个性化需求。而用户的在线评论蕴含个人对产品交易过程的情绪与观点，对其进行深入挖掘与分析，有利于平台的个性化推荐。文本情感分析是提取用户对产品偏好的方法之一，所以利用情感分析方法对评论文本进行分析，提取出用户对产品的偏好，并与推荐算法结合，从而提高推荐的准确度。因此，基于情感分析的个性化推荐不仅可以提高推荐的针对性与准确性，方便消费者迅速、便捷地找到自己需要的产品，同时对平台服务质量的提升具有重要的理论意义和现实意义。

（1）理论意义

从研究理论上来说，协同过滤模型作为常用的推荐模型之一，通过评分矩阵计算用户或项目邻居的偏好进行推荐。而评分矩阵主要是基于产品评分数据和用户历史浏览记录的量化，忽略了用户在线评论的深入挖掘与分析。因此，本文以在线评论文本为研究对象，基于情感分析得到情感评分矩阵，并结合推荐算法进行农产品推荐，一定程度上提高了推荐的准确性。此外，将两种协同过滤推荐模型进行融合，并结合 BMF 算法与改进巴氏系数相似性度量方法，对传统协同过滤模型进行优化，不仅丰富了推荐模型与方法，而且对农产品电商推荐研究来说也是一次创新。

（2）现实意义

从实际应用上来说，电商平台的农产品种类越来越多，在线评论也随之增加，使得用户在购买时不仅需要浏览大量的产品描述信息，还要参考其他买家的评论，耗费了大量的时间和精力，影响了购物体验。因此，利用用户评论深入挖掘与分析其中的情感偏好，提供针对性的推荐服务对平台来说极为必要。

综上所述，农产品的个性化推荐研究是很有必要的，而且在对农产品进行个性化推荐时，考虑用户评论的情感因素不仅提高了产品推荐的针对性与准确性，而且拓展了情感分析在农产品推荐领域的研究。

1.2 国内外研究现状

1.2.1 国外研究现状

（1）情感分析的研究现状

情感分析又叫做观点挖掘，是自然语言处理的基本任务之一，通过对评论文本进行情感分析得到用户对服务、商品等实体的观点或者情感[1]。在 2001 年，Das 和 Chen[2]以及 Tong[3]在对股票交易消息中投资者和在线讨论用户的情感进行分析时，正式提出了“情感分析”这个词。随后，国外逐渐开始大量涌现情感分析领域的研究工作，主要集中于特征提取和文本分类两个方面。

在特征提取的研究上，如 Thet T. T 和 Na J. C 等人[4]建立了电影领域的特征词典，并且针对情感词进行相应的标注，该方法可以确定电影评论语句中用户针对电影不同方面所表达出来的情感极性和程度；Abbasi 和 France 等人[5]提出了一种基于规则的多元变量文本特征抽取的算法，该算法在考虑语义信息的同时也考虑了句法关系带来的影响；Wang H 和 Yin 等人[6,7]在利用信息增益 IG、互信息 MI 等方式的特征选择基础上，构建向量空间模型，对文本分析中产品意见挖掘具有指导意义。

在文本分类的研究上，如 P. Turney[8]将分类算法应用于推荐领域，把评论文本的分类标签设置为推荐和不推荐两类，通过分类结果实现推荐；Aminu Muhammad[9]基于 Twitter 的评论数据，考虑了情感词的上下文信息，提出一种新的情感词典分析方法--SmartSA，实验结果表明，该方法显著提高了 Twitter 评论数据的情感分类效果；Li Ji 和 Lowe Dan 等人[10]从单词级和句子级两方面研究了基于词典的情感分析方法。在单词级情感分析上，使用高频情感词与评论词作对照，在句子级情感分析上，计算出用户对产品各特征的情感极性进行分类；Chih-Fong Tsai 等人[11]提出基于文本情感分析的自动编制评论摘要方法，通过构造文本评论分类器来分析句子的极性，结果表明，该方法的性能优于其他方法；Onur Can Sert 等人[12]融合情感分析和主题建模两种方法对新闻和社交媒体文本评论进行分类，用来预测股票市场的走势，结果表明，该方法的预测

准确度有了大幅提高，等等。

由此可见，国外学者们对文本情感分析的研究主要集中于情感分类和文本的特征提取等方面，且应用领域较广。然而，对于评论文本数据收集的难度较大，而且国外学者研究的情感分析方法大都基于英文文本数据，而中文和英文语种的不同，导致在词性标注、词汇粒度、句法结构等方面的处理都各不相同。因此，适合中文文本情感分析方法的选择极为重要。

（2）推荐技术的研究现状

推荐技术是解决信息过载问题的常用技术之一。然而，直到 20 世纪 90 年代，推荐技术才被单独提取出来成为一个独立的部分[13]，此后的推荐技术发展迅速。个性化推荐最先出现在亚马逊网站中[14]，使用的是基于协同过滤的推荐算法[15]。随着互联网技术的快速发展，网上充斥着海量的信息，学术界对于推荐技术的研究愈加重视，取得了较为显著的成果，主要集中于推荐系统的应用和推荐模型的优化两个方面。

在推荐系统应用的研究上，如 Goldbery 提出的 Tapestry 系统[16]是利用协同过滤模型构建的推荐系统，主要用于电子文档和邮件分类过滤；Ahn 和 Brusilovsky 等人[17]构建了个性化新闻访问系统 YourNews，基于用户的历史浏览记录，抽取用户感兴趣的新闻信息，实现新闻主题的推荐；Celma 以及 Ram ́ez 等人[18,19]提出的 FOAF 音乐推荐系统，主要是基于音乐的描述内容进行个性化推荐；GroupLens[20]和 Ringo[21]推荐系统在评估用户 u 对某个商品兴趣程度时的推荐流程是，首先寻找到曾经购买和评价过该商品的用户群体，使用适合的相似性度量方法计算出该用户群体中用户的兴趣程度，其次对比目标用户 u 与该用户群体的兴趣程度的相似性，最后系统将 N 个相似程度高的用户为 u 进行推荐。

在推荐模型的优化研究上，如 Guang Xing Lye 等[22]基于提出的知识-欲望-意图模型，通过分析用户观点和周围环境的数据，构建了一种带有个性化推荐框架的 SIoT 体系结构，实验精度和召回率结果表明，该方法比传统方法能获得高达 28%左右的 F 值；Yijia Zhang 等[23]针对推荐项目推荐的“冷启动问题”，提出了联合个性化马尔可夫链（JPMC）模型，对用户偏好进行动态建模。并在三个真实数据集上的实验结果表明，该模型显著优于其他模型；Xueping Su 等

[24]基于多类支持向量机对人脸表情进行分类，利用混合 RCNN 计算用户的多兴趣值，提出融合多兴趣值的个性化推荐方法，实验结果表明，与其他算法相比，该方法有明显的改进，等等。

由此可见，国外的学者对推荐技术的研究较早，主要集中于推荐系统在电子邮件、新闻、音乐等领域的应用和推荐模型的应用等方面。然而所研究对象只考虑了用户历史浏览行为或历史交易数据，很少考虑用户的在线评论对推荐效果的影响，推荐的准确性有待提高。

1.2.2 国内研究现状

（1）情感分析研究现状

近年来，随着自然语言处理技术的成熟，情感分析成了文本分析领域研究的热点，目前国内情感分析的文本以中文为主，主要的研究成果集中于文本的分类和情感分析应用两方面。其中，文本的分类有情感词典、机器学习和深度学习三种方法。基于情感词典的文本分类，如周咏梅等人[25]基于知网 HowNet 和 SentiWordNet 的中文情感词典，通过词语义原计算各词语之间的情感极性强度，实验结果表明，该方法可以较好地分析出文本中的情感信息；朱嫣岚等[26]根据语义将篇章进行不同程度的分解，使用知网 HowNet 中文情感词典提出了基于语义相似度和基于语义相关场两种词汇语义倾向性计算的方法，实验结果表明，词汇的细粒度分析可以提高文本分类的准确性。基于机器学习的文本分类，如姚天昉和娄德成[27]通过分析句子中主题词与修饰词之间的关系，提出了计算词语上下文相关词语情感倾向性的算法，结果表明，该算法可以提高情感分类的准确性；龚安等[28]对文本语法规则进行改进，并将一元词、句法、依存词语搭配等特征融合，利用支持向量机分类器进行评论文本的识别与分类，结果提高了文本分类的性能和精度；赵刚等人[29]以餐饮领域的在线评论文本为研究对象，通过扩充的情感词典与朴素贝叶斯算法进行融合，结果表明，客户的情感倾向分类准确性较高；赵志滨等人[30]通过抽取中文产品评论中的特征维度信息，提出维度权值计算方法，该方法对特征维度的细粒度挖掘以及维度情感分析能有效反映市场反馈和用户偏好信息；曾子明等人[31]以微博文本为研究对象，将 LDA 主题模型和 Ada Boost 分类模型进行融合，从而提高情感分类的准确度。基于深度学习的文本分类，如吴鹏等人[32]利用词嵌入技术增加文本的情

感特征，基于 LSTM 模型对生成的词向量进行训练，以此来对负面倾向文本进行情感分类，结果表明该模型取得较好的情感识别效果；冯兴杰等人[33]将卷积神经网络和注意力模型相结合进行文本分析，结果表明，该方法相较于传统的机器学习方法与情感词典方法准确率、召回率均有明显提高；刘思琴等[34]考虑了情感词的上下文关系，提出了融合 BERT 词向量模型、BLSTM 和注意力机制的文本分类模型，测试结果显示，准确率有一定程度的提高；胡德敏等[35]针对多语言文本提出了融合 BLSTM 和注意力机制的细粒度情感分析模型，实验结果表明，该模型的分类效果较好。

此外，情感分析应用领域较广，主要为舆情监测、销量预测以及产品推荐的应用。情感分析在舆情监测的应用，如安璐等[36]基于情感词典分析方法，采用 word2vec 技术提取微博主题的词向量，对不同主题下的评论文本进行细粒度情感分类，以此发现突发事件下网络舆情的演化规律；崔彦琛等[37]以微博突发事件的衍生舆情评论为研究对象，采用词集合并法、SO-PMI、PMI-IR 等方法拓展了舆情情感词典，并结合时间序列分析法对不同阶段的舆情进行实证分析，结果与实际情感值拟合程度较好，证明了该方法的科学性；何天翔等[38]基于情感分析对微博评论数据进行时间分片，然后使用 DTM 模型进行舆情演化分析。实验表明，该模型有助于更好地拟合网络舆情的发展趋势。情感分析在销量预测的应用，如李宏媛等[39]拓展完善了服装领域的情感词典，使用支持向量机对其销量进行预测，与回归模型的对比结果表明，结合情感因素的销量预测模型误差较小；孟园等[40]提出情感指数和 ARMA 融合的产品销量预测模型，实验结果表明，加入情感因素的销量预测模型提高了预测精度。情感分析在产品推荐的应用，如杨春晓等[41]以卷烟产品的网上评论为例，基于拓展的烟草领域词典按照产品、时间、地区等维度分别计算出情感指数，依据情感指数提出相应的对策和建议；王梓萌等[42]总结出物流服务满意度的六个影响因素，通过对评论文本的标注，提出威尔逊置信区间的方法实现基于服务满意度的各产品排序，从而得出推荐列表；梁霞等[43]对在线评论中的产品属性进行提取，将情感值表示成评价的概率分布，并通过随机占优准则对产品进行排序，从而帮助消费者做出产品选择决策；涂海丽等[44]构建了游客情感分析模型，计算出游客关于旅游各要素的情感极性值，结果能直观显示出游客对旅游目的地的总体情感倾向，

为游客选择旅游地提供参考，等等。

可以看出，国内情感分析的研究任务主要是对评论文本进行分类，包括情感词典、机器学习以及深度学习三种方法，从篇章、句子、词语等细粒度方式下分析来提高分类的准确度。在情感分析的应用方面，包括舆情监测、销量预测和产品推荐等。而在使用情感分析进行产品推荐方面，学者们大多是从评论中提取出产品的特征，基于情感分析计算出各特征的情感倾向，然后使用权重加权等方法进行排序，从而达到为用户推荐的效果。但是，这种方法只能给大部分用户在选择前提供大致的参考，不符合他们的个性化需求，推荐的准确度有待提高。

（2）推荐技术的研究现状

近年来，随着电子商务的迅猛发展，淘宝、美团、京东等知名电商将推荐系统应用到自己的平台中，推荐技术的广泛运用引发了国内学者的研究热潮，主要集中于基于内容、关联规则、协同过滤、深度学习以及混合推荐的研究方法。

基于内容的推荐，如骆亮[45]基于内容推荐和余弦相似度算法设计了政府决策辅助系统，结果表明，根据用户偏好和检索词特征进行推荐，具有较好的准确性和针对性；崔春生等[46]借助可拓学的方法对基于内容的推荐算法中相似度计算进行改进，通过实例分析，验证了该方法的可靠性和有效性。基于关联规则的推荐，如高晟[47]提出基于关联规则和贝叶斯结合的个性化图书推荐模型，结果表明，该模型能够提高图书推荐的准确度；陈双双等[48]考虑了用户偏好和产品标签，提出了一种基于关联规则的标签推荐方法，结果表明，该方法有效地提高推荐的准确度。

基于协同过滤的推荐，如袁泉等[49]采用知识图谱补全缺失的数据信息，然后使用协同过滤算法对相似度进行计算，最后应用于矩阵分解的推荐中，对比实验结果表明，该算法在召回率、准确率等指标上都有所提升；李昆仑等[50]提出了一种改进数据填充和相似度的协同过滤推荐方法，将用户特征、评分、时间戳三个因素的加权综合来计算相似度，从而解决数据稀疏性问题和冷启动问题；崔国琪等[51]针对热门项目的流行度对推荐效果的影响，统计出热门项目与活跃用户的惩罚因子来调整项目之间的相似性，并使用协同过滤算法进行推荐，

结果显示，加入惩罚因子的协同过滤推荐算法在保持算法准确率的同时，可在一定程度上降低流行度对推荐结果的影响；郑修猛等[52]基于情感分析将用户的评论文本转化为用户对该项目的评分，构建情感评分矩阵，并结合推荐算法进行推荐，该方法有效解决推荐系统评分矩阵的稀疏性问题；卢竹兵等[53]通过对在线评论文本进行情感分析，计算各用户情感值的相似性建立用户间的信任关系，并与原始评分的相似度结合，提出基于用户情感信任关系的推荐策略，实验表明，改进的引入情感分析信任模型的协同过滤推荐能够有效地降低平均绝对误差值；钱春琳等[54]结合情感分析和协同过滤设计了个性化推荐算法，并对协同过滤搜索邻居集上进行改进，实验结果表明，该方法能够有效改进推荐结果的准确率；彭敏等[55]从评论文本中提取出 K 个用户关注的产品属性，通过情感分析计算各个属性的情感偏好，从而构建基于情感分析的构建推荐，结果表明，该模型有效改善了数据稀疏性的问题，同时提高了推荐系统的精度。基于深度学习的推荐，如熊旭东等[56]构建融合多阶、多层次的图卷积模型，对交互节点进行嵌入表示，提出了一种基于二分图卷积学习的推荐算法，实验表明，该算法在 HR 和 NDCG 衡量指标上相较于其他算法较好；杨丹等[57]提出生成对抗网络推荐算法，实现实时动态的个性化推荐，结果表明，该算法具有较好的推荐准确度。

基于混合算法的推荐，如何婧等[58]为解决缺失数据对推荐准确度的影响，提出一种融合矩阵分解和 XGBoost 的推荐算法，实验结果显示，该方法的推荐准确度优于其他方法；杨兴雨等[59]根据随机森林模型对聚类后的用户-项目矩阵进行评分预测，实验结果表明，该方法的推荐效率高于基于近邻关系的协同过滤算法；沈晶磊等[60]将用户浏览帖子问题转化为分类模型，提出了基于随机森林的推荐系统，实验结果证明了系统的有效性并提升了效率；陆君之[61]从五个维度构建电影的特征向量，通过随机森林回归算法构建电影评分预测模型，结果表明，该模型可以有效地预测出电影的评分；滕传志等[62]通过随机森林对数据分类，再利用马尔科夫链去除冗余信息，提出基于随机森林-马尔可夫链相结合的方法，结果表明，该方法具有较高的准确率和召回率，等等。

可见，国内对推荐技术的研究成果主要集中在基于内容、关联规则、协同过滤、深度学习以及混合推荐等研究方法上。也有部分学者将推荐问题转化为

分类问题，运用随机森林、XGBoost 等方法进行推荐。其中，基于协同过滤的推荐方法使用较多，针对该方法存在的数据稀疏性问题，加入情感因素、时间因素、知识图谱等填充和完善评分矩阵进行算法的改进。而在基于情感分析的协同过滤推荐上，学者们主要从热门商品的惩罚因子、不同商品的情感评分、同类商品的属性评分等方面构建情感评分矩阵，进行个性化推荐。而基于情感分析得到的评分矩阵受到用户共同评分的影响，在评分数据极为稀疏没有足够的共同评分项情况下效果不佳以及忽略了用户的整体偏好，影响了推荐的针对性与准确性。

1.2.3 文献评述

通过对国内外研究成果进行分析可以发现，情感分析的研究成果主要集中在特征提取、文本分类、情感分析应用等几个方面。但目前存在以下问题：（1）英文文本在词性标注、词汇粒度、句法结构等方面的文本特征处理与中文文本存在极大不同，英文文本的情感分析特征提取方法不完全适用于中文文本。（2）目前国内的情感分析大部分在情感分析方法的研究上，而情感分析在推荐领域的应用比较少。鉴于此，本文基于中文在线评论文本，在现有的中文情感词典基础上，拓展适用于农产品领域的情感词典，并且加入网络词词典进行词典的完善，通过对电商平台上用户的在线评论进行情感分析，将结果与推荐算法结合，构建基于在线评论情感分析的农产品个性化推荐模型，并通过与其他三种推荐模型的对比，验证模型的可行性。此外，推荐技术的研究成果主要集中在推荐系统的应用、推荐模型的优化、推荐方法的研究等几个方面。但目前存在以下问题：（1）现有的推荐方法忽略了在线评论对推荐的影响。（2）受到评分矩阵稀疏性的影响以及忽略了用户的整体偏好，推荐方法的准确度较低。

因此，本文提出基于情感分析的农产品个性化推荐模型，即对电商平台农产品在线评论进行情感分析，得到情感评分矩阵。在此基础上，使用 BMF 矩阵分解算法对评分矩阵的缺失值进行填充，然后使用改进的巴氏系数计算用户的相似度得到邻居集，最后将基于用户和基于物品的两种协同过滤方法融合，进行目标用户的产品推荐，从而提高推荐的准确率。

1.3 研究内容与框架

1.3.1 研究内容

本文以提高农产品推荐的准确性为目标，构建基于情感分析的农产品个性化推荐模型，并与 UCF、ICF、HCF 三种推荐模型对比，验证模型的优越性。概括起来，本文的主要研究内容如下：

（1）基于情感词典计算用户评论的情感值，得出情感评分矩阵。在现有情感词典的基础上，利用情感倾向点互信息算法和 LDA 主题模型对农产品领域的情感词典进行拓展，并加入网络词词典进行完善，以此计算出用户评论的情感值，从而得出情感评分矩阵。

（2）构建了融合矩阵分解和改进巴氏系数的混合推荐算法。将 BMF 算法、改进巴氏系数以及混合协同过滤算法融合得到混合推荐算法，并在数据集上验证了算法的有效性。

（3）构建了基于情感分析的农产品个性化推荐模型。基于词典的情感分析得到情感评分矩阵，并融合混合推荐算法，建立基于情感分析的个性化推荐模型。通过与其他三种推荐模型的对比，验证模型的可行性。

（4）设计了农产品个性化推荐系统。将基于情感分析的个性化推荐模型应用在推荐系统中，从而为用户提供个性化的农产品推荐。

1.3.2 研究框架

全文共分为七章，文章框架及各章节的内容简介如下：

第一章：绪论。介绍了基于情感分析的个性化推荐方法的研究背景以及研究意义，分析了情感分析和推荐方法的研究现状，同时阐述了本文的主要研究内容及创新点。

第二章：相关理论研究。介绍了情感分析的相关理论及推荐算法的基本思想和原理。

第三章：基于词典的在线评论情感分析。在现有情感词典基础上拓展了农产品领域的情感词典，并加入网络热词词典对词典进行完善。通过文本预处理，基于构建的情感词典计算评论文本的情感值，得到情感评分矩阵。

第四章：融合矩阵分解和改进巴氏系数的混合推荐算法。本章构建了混合推荐算法，并在标准数据集上，对该算法进行训练与检验。

第五章：基于情感分析的农产品个性化推荐模型的构建。以甘肃省农产品电商的在线评论文本为研究对象，将情感分析与混合推荐算法结合，建立了基

于情感分析的农产品个性化推荐模型，并对该模型进行训练和检验。

第六章：基于情感分析的农产品个性化推荐系统设计。将基于情感分析的农产品个性化推荐模型应用在推荐系统中，从而为用户提供个性化的农产品推荐。

第七章：总结与展望。总结论文的主要工作，在此基础上提出文章存在的不足之处，并对未来该领域的研究进行展望。

1.4 论文的创新点

本文的创新点主要包括以下两个方面：

第一，农产品领域情感词典的拓展。为了准确地计算出用户在评论文本中农产品特征的情感值，本文在通用情感词典基础上拓展了农产品领域的情感词典，并加入了近 3 年的网络热词词典进行词典的完善。

第二，改进了传统巴氏系数度量方法。在传统的巴氏系数相似性度量方法中加入了调和平均权值因子和用户的整体偏好特征，并将其应用到推荐算法中，取得较好的推荐效果。

1.5 本章小结

本章主要介绍了论文的研究背景及研究意义，系统分析了国内外学者对于情感分析及推荐方法的研究现状，在此基础上，介绍了本文的研究方法、研究内容、研究框架以及创新点。

2 相关理论研究

个性化推荐需要一定的理论基础作为支撑，本章对研究用到的相关理论与方法进行介绍，包括基于情感词典、基于机器学习的情感分析方法，以及基于内容、基于关联规则以及协同过滤的基本思想和建立步骤等。

2.1 情感分析相关理论研究

情感分析(Sentiment Analysis,简称 SA)，是指利用自然语言处理和文本挖掘技术，对带有情感色彩的主观性文本进行分类、观点挖掘的过程[63]。目前，文本情感分析在舆情分析、信息检索、自然语言处理、文本观点挖掘等多个领域的应用逐渐增多，受到许多学者的关注。

通常情况下，情感分析研究的任务类型，可分为情感特征提取、情感极性分类、情感检索与归纳等[64]。其中，情感极性分类又称情感倾向性分析，是指对给定的主观性文本，即用户对产品或服务的主观性描述，进行正向、负向以及中性的判断。情感极性分类的关键是根据文本特征，包括词语的词性、词频、情感词以及副词等，识别出文本信息的情感极性。目前，情感分析中文本极性分类主要的研究方法分别为基于词典的方法和基于机器学习的方法。

2.1.1 基于词典的情感分析方法

情感词典是文本情感分析的一种无监督的方法，即利用构建的情感词典，并对其进行极性和强度标注，从而实现文本情感分类。因研究领域的不同，情感词典中正向和负向的情感词之间差异较大，所以需要在现有的基础情感词典分析、总结基础上，进行情感词典的扩充和标注，形成适应所研究领域的情感词典，进而有效地进行文本情感分类。目前在实际应用中，主要从通用情感词、程度副词、否定词以及领域情感词四个方面来对情感词典进行构建。不同于需要大量训练数据集的机器学习算法，基于情感词典的文本情感分类方法采用权值累加的统计方法进行分类，即给不同情感强度的程度副词、否定词和情感词等赋予不同权值，然后加权求和。最后确定阈值来判断文本情感倾向。一般情况下，计算结果为正表示该文本为正面倾向，计算结果为负表示该文本为负面倾向，计算结果为零表示该文本为中性评论。

基于词典的文本情感分类方法属于粗粒度的倾向性分类算法，优点是不受

标注的训练集的影响，所以对短文本的分类相对有效。缺点是构建的词典往往只针对某个特定的研究领域，对于跨领域文本的分类效果不佳。而且使用该方法的关键点在于情感词典的覆盖度，需要过多的人工标注来扩充词条信息，这大大增加了人工成本。

2.1.2 基于机器学习的情感分析方法

机器学习是文本情感分析的一种有监督的方法，即将文本当成分类问题进行研究。基于机器学习的情感分析工作流程，首先，标注大部分语料作为训练集；其次，对训练集中的语料进行分词、词性标注等预处理；再次，使用 TF-IDF、Word2Vec 等词向量模型对预处理后的训练集进行文本特征工程，选择一种分类方法对分类器进行训练；最后，将训练好的分类器对测试集进行文本分类。

基于机器学习的文本情感分类方法，不受所研究领域文本的影响，应用范围比较广，而且供选择的分类方法也较多。但是，因文本中存在复杂的情感表达和大量的情感歧义词，造成文本分类的准确率较低，所以文本的预处理工作和特征提取直接影响情感分类任务的性能。

为了准确得到整个文本最终的情感值，通过对以上两种方法的总结分析，本文采用逻辑回归（Logistic Regression）机器学习算法对文本进行情感倾向分类，在此基础上，利用在文本检索领域常用的 LDA 主题模型对文本中产品的特征与情感词进行提取，作为情感词典的补充，最后使用情感词典计算情感值。基于逻辑回归与 LDA 主题模型的原理介绍如下：

（1）逻辑回归

逻辑回归（Logistic Regression）算法[65]主要解决分类问题，用来解决事件发生的可能性，最早用于二分类问题，在原有算法基础上进行改进，可以用逻辑回归解决多分类问题。以二分类为例，逻辑回归的思想是：通过非线性函数将连续值转换为离散的二值问题，即分类输出结果 y 不是连续的值，其输出只能为 0 或 1，分别代表两种不同类别。假设训练样本为{x,y}，x 是 m 维的样本特征向量，y 为输出结果，且 y∈{0,1}。则逻辑回归模型如公式（2.1）所示。

$$h_\theta(\mathrm{x}) = \frac{1}{1+\mathrm{e}^{-\theta^T x}} \tag{2.1}$$

其中，θ为输入变量对应的系数，表示权重，$h_\theta(x)$为 sigmoid 函数，其结果为输出值。通过该函数将输入的数据压缩到 0 和 1 二值范围内。若响应变量

$\theta^T x$大于阈值 0，则被判为正向类 1（如公式（2.2）所示）；$\theta^T x$小于阈值 0，则被判为负向类 0（如公式（2.3）所示）。逻辑回归模型通常采用极大似然估计法计算对应输入变量的系数，即

$$P(y=1|x)=h_\theta(x) \tag{2.2}$$

$$P(y=0|x)=1-h_\theta(x) \tag{2.3}$$

由公式（2.2）和公式（2.3）可得

$$P(y|x)=\big(h_\theta(x)\big)^y\big(1-h_\theta(x)\big)^{1-y} \tag{2.4}$$

则最大似然函数为

$$L(\theta)=\prod_{i=1}^{m}P(y_i|x_i)=\prod_{i=1}^{m}(h_\theta(x_i))^{y_i}\,(1-h_\theta(x_i))^{1-y_i} \tag{2.5}$$

对公式（2.5）两边取对数，然后两边求导，应用牛顿法进行迭代求解得到对应输入变量的系数。将求得的系数带入公式（2.2）中即可建立逻辑回归分类模型，求$L(\theta)$的最小值是该算法的最终目标。

（2）LDA 主题模型

LDA（Latent Dirichlet Allocation）主题模型，也称为三层贝叶斯概率模型[66]。因其在文本训练时不需要对训练集过多的标注，仅设定需确定的文档集和指定的主题数目，所以 LDA 主题模型是用于对文档主题提取的一种无监督机器学习模型。

由于对文本抽取的主题及特征词数量不限，在对 LDA 模型的参数估计时，通常采用 Gibbs 抽样对其进行参数估计。首先，基于 Gibbs 抽样的参数估计需要计算每一个主题下每一个词的主题概率，其计算规则如公式（2.6）所示。

$$p(\omega|\alpha,\beta)=\int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)\,p(\omega_n|z_n,\beta)\right)d\theta \tag{2.6}$$

其中，ω表示文本中的单词，α表示评论-主题分布的 Dirichlet 超参数，β表示主题-单词分布的 Dirichlet 超参数，θ表示评论-主题分布，z 表示评论中单词的主题。

确定了每个词的主题，参数估计即计算词对应列表下主题序列的条件概率。其计算规则如公式（2.7）所示。

$$p(z_i = k|\overrightarrow{z_{\neg\iota}}, \vec{\omega}) = \frac{p(\vec{\omega}, \vec{z})}{p(\vec{\omega}, \overrightarrow{z_{\neg\iota}})} \propto \frac{n_{k,\neg i}^{\omega} + \beta_{\omega}}{\sum_{t=1}^{V} n_{k,\neg i}^{\omega} + \beta_{\omega}} \left(n_{m,\neg i}^{t} + \alpha_k\right) \tag{2.7}$$

其中，$\neg i$表示不包括第 i 项，z_i表示第 i 个词对应的主题变量，n_k^{ω}表示第 k 个主题中对应的词ω出现的次数，α_k表示主题 k 的 Dirichlet 先验，β_{ω}表示词ω的 Dirichlet 先验。

获得每个词的主题概率后，计算所需参数。其计算规则如公式（2.8）与公式（2.9）所示。

$$\Phi_{k,\omega} = \frac{n_k^{\omega} + \beta_{\omega}}{\sum_{t=1}^{V} n_k^{\omega} + \beta_{\omega}} \tag{2.8}$$

$$\theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k} \tag{2.9}$$

其中，$\Phi_{k,\omega}$表示文本主题 k 中词ω的概率，$\theta_{m,k}$表示文本 m 中主题 k 的概率，n_m^k表示文本 m 出现主题 k 的次数。

2.2 推荐技术相关理论研究

随着互联网技术的日益成熟与发展，海量的数据信息为商家和企业实现高质量的服务以及帮助消费者快速、便捷地找到需要的产品带来了机遇与挑战，如何从中挖掘和发现消费者感兴趣的信息使得商家和企业提供针对性的服务，进而提升消费者的购物体验成为一个有意义的研究课题。推荐技术作为一种有效的识别用户偏好的方法，被许多的网站使用，如 Amazon、Netflix、京东等，以高效和令人满意的方式为消费者推荐需要的产品信息，从而提高用户体验。目前，在个性化推荐领域，运用最多的推荐算法有基于内容的推荐算法、基于关联规则的推荐算法以及协同过滤推荐算法。

2.2.1 基于内容的推荐算法

基于内容的推荐（Content Based Recommendation），是早期在推荐系统领域应用最广的推荐算法。该算法的主要思想是通过对用户关注或购买后的历史项目特征进行挖掘分析，比较用户偏好特征和项目特征的相似性，按顺序生成项目列表进而推荐给目标用户。基于内容的推荐算法整个流程如图 2.1 所示，具体可分为三个步骤：

步骤 1：项目特征建模。通过读取项目内容，抽取其特征，并对其进行向量化表示。

步骤 2：用户特征偏好学习。利用每个用户的历史数据中关于项目的特征数据，学习和发现用户的兴趣偏好特征。

步骤 3：项目 TOP-N 推荐。对用户偏好特征和项目特征进行相似性度量，将相似性大的前 N 个项目为目标用户进行推荐，构成推荐列表。

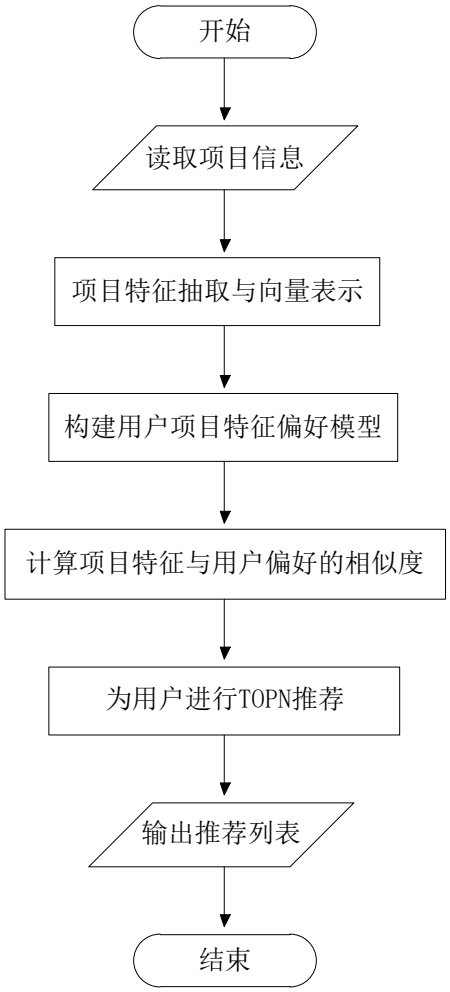


图 2.1 基于内容的推荐算法流程

相比于其他推荐算法，基于内容的推荐算法主要的优点如下：

第一，该算法基于项目特征进行推荐，不受其他用户兴趣偏好的影响，也不存在过度推荐的问题。

第二，不过分依赖用户历史评分数据，不存在针对新项目的冷启动问题。

另外，该算法存在如下缺点：

第一，存在用户冷启动问题。由于没有新用户的历史数据，无法提取其对项目的偏好特征，因此不能为新用户提供推荐服务。

第二，难以发现用户潜在兴趣。由于该算法只是对用户的历史数据进行特

征偏好学习并对其进行相应地推荐，推荐结果仅符合用户过去的兴趣偏好，无法发现用户当前或未来的潜在偏好。

2.2.2 基于关联规则的推荐算法

基于关联规则的推荐（Association Rules Recommendation），是较常见的推荐算法之一，在零售业销售中应用广泛。其核心思想是：如果用户在购买了 A 产品后同时购买了 B 产品，则产品 A 和产品 B 具备关联性，即购买过 A 产品的用户存在更大可能购买 B 产品。基于关联规则的推荐算法整个流程如图 2.2 所示，具体可分为三个步骤：

步骤 1：数据预处理。对数据中的用户和项目分别计数，设定阈值，对不活跃的用户和冷门项目进行过滤。

步骤 2：计算关联指标，并进行剪枝。计算两两项目的支持度（support）和置信度（confidence），根据最低支持度和置信度进行规则剪枝。

步骤 3：项目 TOP-N 推荐。查找项目的所有规则，并按照置信度降序排列，将与项目最相关的前 N 个项目进行推荐，构成推荐列表。

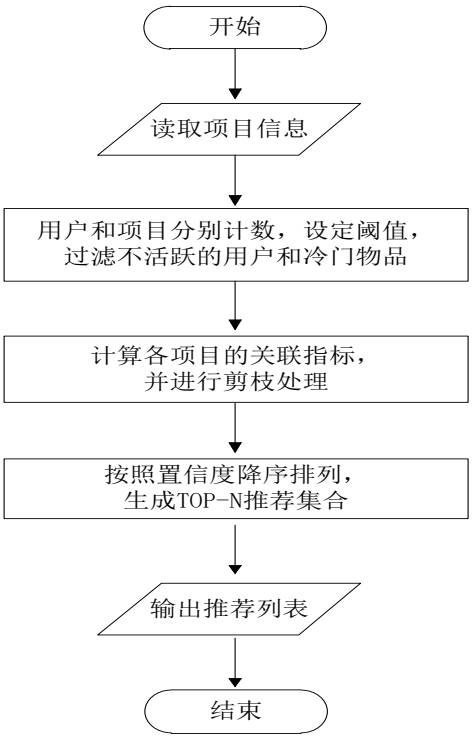


图 2.2 基于关联规则的推荐算法流程图

基于关联规则的推荐优点是：能从大量的用户购买数据中挖掘出产品之间的关联信息，并且帮助用户发现潜在的喜好产品，提高了推荐算法的“惊喜度”，

从而指导商家制定“捆绑销售”策略来满足用户需求，同时也提高了产品的销量。但是，缺点也很明显。第一，各项目之间的关联指标计算量较大，耗时长。第二，受用户欢迎的热门项目会被过度推荐，项目推荐的覆盖范围有限。

2.2.3 协同过滤推荐算法

协同过滤推荐算法（Collaborative Filtering Recommendation），是一种在推荐领域应用最广泛的推荐算法，其核心思想是：根据用户对项目的关注或购买的历史数据，挖掘发现与其兴趣爱好相似的其他用户或品味相投的其他项目，据此进行相应推荐。目前主要有三类协同过滤算法：即基于用户的协同过滤算法、基于项目的协同过滤算法、基于模型的协同过滤算法[67]。

（1）基于用户的协同过滤算法

基于用户的协同过滤算法（User Based Collaborative Filtering），是根据用户对项目的历史交互数据，发现用户的偏好习惯，寻找与其相似的“邻居”群体。基于用户的协同过滤算法的推荐机制如图 2.3 所示，具体可分为三个步骤。

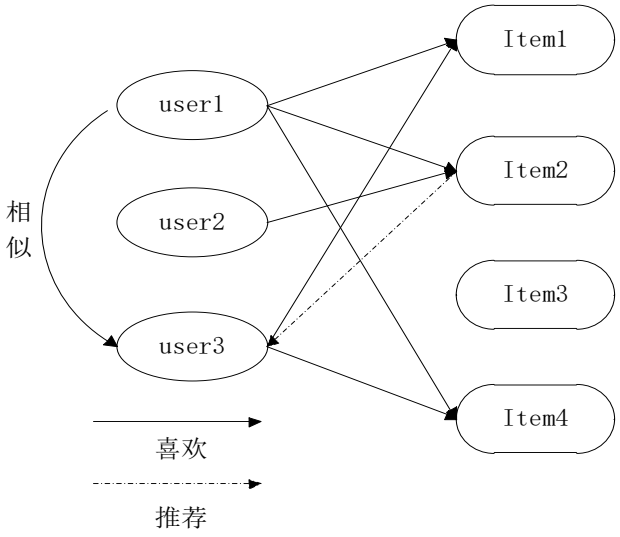


图 2.3 基于用户的协同过滤推荐机制

步骤 1：用户对项目的偏好建模。预处理“用户-项目”的评分信息，将评分信息存储在数据透视表中，如表 2.1 所示。其中，$user_i$表示第 i 个用户，$item_j$表示第 j 个项目，v_{mn}表示第 m 个用户对第 n 个项目的评分值。

表 2.1 “用户-项目”评分矩阵

	$item_1$	$item_2$	$item_3$	…	$item_j$
$user_1$	v_{11}	v_{12}	v_{13}	…	v_{1j}

$user_2$	v_{21}	v_{22}	v_{23}	…	v_{2j}
$user_3$	v_{31}	v_{32}	v_{33}	…	v_{3j}
…	…	…	…	v_{mn}	…
$user_i$	v_{i1}	v_{i2}	v_{i3}	…	v_{ij}

步骤 2：用户相似性计算，查找用户的邻居集。即依据“用户-项目”评分矩阵，计算用户之间的相似性。在“用户-项目”评分矩阵中，如果用户没有对某项目进行评分，则该项目维度对应的值为 0。相似性计算的方法主要有以下几种：

1）余弦相似度

余弦相似度（Cosine Similarity），假设有用户 u 和用户 v 两个用户，在 n 维项目空间上的评分向量为 $\vec{u}$ 和 $\vec{v}$，其中 $\vec{u} = (\mathrm{u}_1, \mathrm{u}_2, \mathrm{u}_3 \ldots, \mathrm{u}_n)$，$\vec{v} = (\mathrm{v}_1, \mathrm{v}_2, \mathrm{v}_3 \ldots, \mathrm{v}_n)$，则两个用户 u 和 v 的相似度为 sim(u,v)，R_{ui}和R_{vi}分别代表用户 u 和用户 v 对项目 i 的评分，其计算规则如公式（2.10）所示。

$$\mathrm{sim}(\mathrm{u},\mathrm{v}) = \cos(\vec{u},\vec{v}) = \frac{\vec{u} * \vec{v}}{|\vec{u}| * |\vec{v}|} = \frac{\sum_{i=1}^{n} \mathrm{R}_{ui} * R_{vi}}{\sqrt{\sum_{i=1}^{n} R_{ui}^2} * \sqrt{\sum_{i=1}^{n} R_{vi}^2}} \tag{2.10}$$

余弦相似度的缺点在于过分依赖共同评分集合，而且在实际中不同用户的打分尺度不一致造成评分差异显著，其也能表现出高相似性，此时余弦相似性就不能准确地度量用户间的相似性。

2）皮尔逊相关系数

皮尔逊相关系数（Pearson coefficient），是一种度量两个用户之间的关联程度的方法。假设用I_{uv}表示第 u 个用户和第 v 个用户的评分集合交集，用户 u 和 v 的相似度为$sim_{pc}(\mathrm{u},\mathrm{v})$，其计算规则如公式（2.11）所示。

$$sim_{pc}(\mathrm{u},\mathrm{v}) = \frac{\sum_{i \in I_{uv}} [(\mathrm{R}_{ui} - \bar{R}_u) * (\mathrm{R}_{vi} - \bar{R}_v)]}{\sqrt{\sum_{i \in I_{uv}} (\mathrm{R}_{ui} - \bar{R}_u)^2} * \sqrt{\sum_{i \in I_{uv}} (\mathrm{R}_{vi} - \bar{R}_v)^2}} \tag{2.11}$$

其中，R_{ui}和R_{vi}分别表示用户 u 和用户 v 对项目 i 的评分，$\bar{R}_u$和$\bar{R}_v$分别表示用户 u 对项目评分的均值和用户 v 对项目评分的均值。皮尔逊相关系数在一定程度上考虑了用户的评分差异，但该度量方法依赖用户评分的数目，并不能很好的度量用户间的相似性。

3）巴氏系数

巴氏系数（Bhattacharyya coefficient），是通过概率分布的最大限度表示两个数据的相关性，由 Bidyut Kr. Patra 等人[68]将其引入协同过滤推荐中。假设在 n 维项目空间上的评分数据中，m 为所有分值（通常分值范围在 1-5 分之间），用户 u 和 v 的相似度为$sim_{bc}(\mathrm{u},\mathrm{v})$，其计算规则如公式（2.12）所示。

$$sim_{bc}(\mathrm{u},\mathrm{v})=\sum_{h=1}^{m}\sqrt{\widehat{P_{uh}}*\widehat{P_{vh}}}*\frac{\sum_{i\in I_{uv}}[(\mathrm{R}_{ui}-\bar{R}_u)*(\mathrm{R}_{vi}-\bar{R}_v)]}{\sqrt{\sum_{i\in I_{uv}}(\mathrm{R}_{ui}-\bar{R}_u)^2}*\sqrt{\sum_{i\in I_{uv}}(\mathrm{R}_{vi}-\bar{R}_v)^2}} \tag{2.12}$$

其中，用户 u 对所有项目的评分个数为u'，用户 v 对所有项目的评分个数为v'，两个用户评分值均等于 h 的概率密度分布为$\widehat{P_{uh}}$和$\widehat{P_{vh}}$，则$\widehat{P_{uh}}=\frac{\#h}{u'}$且$\widehat{P_{vh}}=\frac{\#h}{v'}$，$\#h$表示用户 u 和用户 v 评分值为 h 的个数。从公式（2.12）可以看出，该算法是用户 u 和用户 v 两组评分数据的全局相似和局部相似的加和运算。全局相似是对用户的所有评分项进行相似性度量，局部相似是对两用户的共同评分项进行相似性度量。当两用户 u 和 v 评分值没有交集时，两个用户的巴氏系数等于局部相似性。

巴氏系数协同过滤算法（BCF）是一种新的在稀疏数据中寻找用户之间相似性的方法，通过利用一对用户的所有评分项进行相似性度量，解决了在极为稀疏的数据集中，传统相似性度量方法过分依赖用户共同评分项所造成的预测准确性不高问题。

最终根据以上方法计算得出的相似性矩阵，选出相似值较大的前 N 个用户，作为近邻集。相似性矩阵如表 2.2 所示。其中，$user_i$表示第 i 个用户，p_{mn}表示第 m 个用户和第 n 个用户之间的相似度。

表 2.2 用户相似度矩阵

	$user_1$	$user_2$	$user_3$	…	$user_i$
$user_1$	p_{11}	p_{12}	p_{13}	…	p_{1i}
$user_2$	p_{21}	p_{22}	p_{23}	…	p_{2i}
$user_3$	p_{31}	p_{32}	p_{33}	…	p_{3i}
…	…	…	…	p_{mn}	…
$user_i$	p_{i1}	p_{i2}	p_{i3}	…	p_{ii}

步骤 3：预测用户评分，生成推荐列表。待预测的用户 u 对项目 i 的评分为P_{ui}，其计算规则如公式（2.13）所示。

$$P_{ui} = \overline{R_u} + \frac{\sum_{n \in S} sim(u,v) * (R_{vi} - \overline{R_v})}{\sum_{n \in S} |sim(u,v)|} \tag{2.13}$$

其中，S 表示与用户 u 最相似的邻居集合，$sim(u,v)$表示用户 u 和用户 v 之间的相似度，$\overline{R_u}$和$\overline{R_v}$分别表示用户 u 对所有项目的平均值和用户 v 对所有项目的平均值。

（2）基于项目的协同过滤算法

基于项目的协同过滤算法（Item Based Collaborative Filtering），是根据用户对项目的历史交互数据，计算项目之间的相似度，寻找与其相似的项目集合，向目标用户推荐与其购买项目相似的其他项目。基于项目的协同过滤算法的推荐机制如图 2.4 所示。由图可知，user1 喜好 Item1、Item3，user2 喜好 Item1、Item2、Item3，user3 喜好 Item1，可得 Item1 和 Item3 相似，所以将 Item3 推荐给同样喜欢 Item1 的 user3。

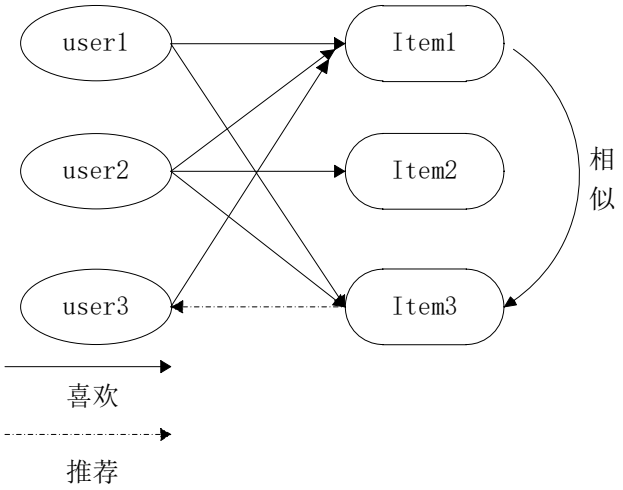


图 2.4 基于项目的协同过滤推荐机制

（3）基于模型的协同过滤算法

基于模型的协同过滤推荐算法，是在协同过滤思想的基础上，使用机器学习和统计学理论等方法对高维度的用户-项目评分矩阵进行模型的训练，并为目标用户进行推荐，一定程度上提升了推荐质量。常见的模型可分为：基于概率的模型、基于矩阵分解模型和聚类模型等。

1）基于概率的模型

基于概率的模型，核心思想是：从评分矩阵中训练出函数映射关系，而这个映射关系被称为分类模型，故基于概率的模型推荐可以看成分类问题，代表

的算法是朴素贝叶斯方法。该方法构建的用户偏好模型，可以很好地减弱噪声点的影响，能够在短时间内学习出比较通用的模型。但是，随着用户-项目评分矩阵的维度越来越大，用户和项目的数量不断增大，该算法后期的维护和更新变得越来越困难。

2）基于矩阵分解模型

基于矩阵分解的模型，核心思想是：通过降低评分矩阵的维度来填充矩阵中的缺失值，从而缓解稀疏性问题。计算规则如公式（2.14）所示。

$$R_{m\times k} = P_{m\times n} * Q_{n\times k} = \bar{R}_{m\times k} \tag{2.14}$$

其中，$P_{m\times n}$代表用户的因子矩阵，$Q_{n\times k}$代表项目的因子矩阵，$R_{m\times k}$代表用户和项目的对应矩阵，$\bar{R}_{m\times k}$代表预测后的用户和项目的对应矩阵。该方法通过降维的方式简化数据，获取矩阵的关键信息，提高了推荐结果。但是，推荐结果可解释性差且会丢失部分信息。

3）聚类模型

聚类模型，在推荐中的核心思想是：通过计算该用户或项目与各簇中心的距离，与各簇距离最近即为该用户或项目的所属类别。因事先确定了所求用户或项目的所属类别，所以在协同过滤推荐中，不要通过全局搜索查找近邻集，只需在其所属类别中寻找，因此在一定程度上节省了时间并提升了效率。基于聚类的协同过滤算法提高了算法的实时性和扩展性，但是无法反映用户兴趣的多样性。

2.3 模型性能的评价

关于机器学习分类的评价指标，本文使用精准度(accuracy)对文本分类模型进行评价，如公式（2.15）所示。

$$\text{accuracy} = \frac{F+N}{T} \tag{2.15}$$

其中，F 表示分类结果为正向的文本个数，N 表示分类结果为负向的文本个数，T 则表示总的文本数量。

推荐系统的评价指标可以综合客观的衡量推荐系统的性能，可分为预测准确度和 TOP-N 推荐两类推荐指标[69]。其中，预测准确度是测试预测评分与实际评分之间的误差，可分为平均绝对误差（MAE）和均方根误差（RMSE）两种。

（1）平均绝对误差

平均绝对误差的计算规则如公式（2.16）所示。

$$MAE = \frac{\sum_{u,i \in T} |r_{ui} - \hat{r}_{ui}|}{|T|} \tag{2.16}$$

其中，r_{ui}表示测试集中用户 u 对项目 i 的实际评分，而$\hat{r}_{ui}$是推荐算法给出的预测评分，T 表示测试集的所有用户或项目。

2）均方根误差

均方根误差的计算规则如公式（2.17）所示。

$$RMSE = \sqrt{\frac{\sum_{u,i \in T} (r_{ui} - \hat{r}_{ui})^2}{|T|}} \tag{2.17}$$

其中，r_{ui}表示测试集中用户 u 对项目 i 的实际评分，而$\hat{r}_{ui}$是推荐算法给出的预测评分，T 表示测试集的所有用户或项目。

另外，TOP-N 推荐是提供满足目标用户需求的推荐列表，包括准确率 Precision、召回率 Recall、F1 值等评价指标。

2.4 本章小结

本章主要介绍了论文中所涉及到的相关理论，首先，详细分析了两种情感分析方法，分别介绍了基于情感词典与基于机器学习的情感分析方法原理。其次，分别介绍了基于内容、基于关联规则以及协同过滤三种推荐方法，其中重点分析了协同过滤三种主要方法的原理，包括基于用户的协同过滤、基于项目的协同过滤以及基于模型的协同过滤。最后，对分类模型与推荐模型的评价指标进行解释说明。

3 基于词典的在线评论情感分析

情感词典，包括通用词典、副词词典、领域词词典以及网络词典等，通过对各种词典中的词语进行标注，实现文本信息的量化，从而识别该文本的情感倾向。因此，构建全面、准确的情感词典是情感分析的关键。本文从文本分类、领域词典拓展以及 LDA 情感词抽取三方面对情感词典进行完善，基于情感词典计算用户评论的情感值。

3.1 基本流程

基于词典的情感分析基本流程如图 3.1 所示。本章基于词典的情感分析可分为 6 个步骤。

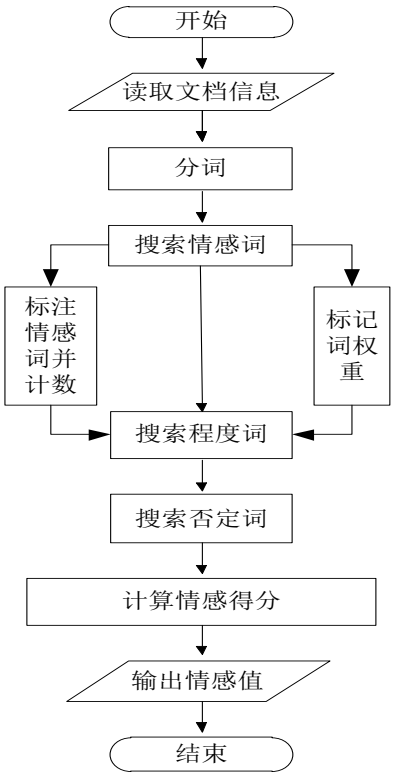


图 3.1 情感词典基本流程

步骤 1：文本的预处理。首先对全部文本进行清洗和筛选等规范化处理，然后对其进行分词、词性标注、去除停用词等处理工作。

步骤 2：文本分类。首先使用词向量模型将文本的特征词转换成向量，然后基于逻辑回归进行特征训练，实现文本分类。

步骤 3：构建基础情感词典。对现有的情感词典人工进行扩充标注，形成基础情感词典。

步骤 4：拓展领域情感词典。提取文本中的候选词，将其与基础情感词典逐个扫描，并使用情感倾向点互信息算法 SO-PMI 拓展领域情感词典。此外，基于 LDA 主题模型对产品特征与情感词进行抽取，作为领域情感词典的补充。

步骤 5：加入其他词典。收集程度词词典以及近 3 年流行的网络热词词典，进行人工标注，对情感词典进行完善。

步骤 6：计算文本情感得分。

3.2 情感词典的构建与拓展

3.2.1 情感词典的构建

本章所构建的情感词典主要由基础情感词典、程度词词典、网络词典以及停用词词典组成。

（1）基础情感词典

基础情感词典主要分为积极和消极两类具有明显情感极性的词语构成。目前，现有的基础情感词典包括知网 HowNet、台湾大学 NTUSD、大连理工大学--情感词汇本体、清华大学 TSING 等。本文通过对以上词典通过人工识别与合并去重的处理方式，将情感极性相同的词语归为相关的极性词典，结果如表 3.1 所示。

表 3.1 基础情感词

情感极性	举例	词数量/个
积极	喜欢、高雅、干净	5567
消极	浮夸、忧虑、苦涩	4370

（2）程度词词典

程度词词典包括副词词典和否定词词典。其中，副词会对文本的情感强度有影响，但不会改变其情感极性。例如，“我买的水果味道不正宗”和“我买的水果味道是最不正宗的一次”，两条文本虽然都是表达对水果味道的不喜欢，而且情感倾向均是消极的，但是表达的情感强度差别很大，后一句比前一句情感消极的程度明显更加强烈。可以看出，副词对文本情感分析有一定的影响，需构建副词词典。本文参照知网 HowNet 副词词典，根据副词的不同程度赋予不同的权重。此外，否定词对文本的情感极性影响很大。例如，“我喜欢这家店的水果”、“我不喜欢这家店的水果”以及“我不可能不喜欢这家店的水果”，前两

条文本因存在否定词“不”，其情感极性很明显，前一条为积极的情感，后一条为消极的情感。而第三条文本中因否定词数量的增多，其情感极性也会出现相应的变化。因此，否定词在文本情感分析中需加入进去。最终，构建的程度词词典结果如表 3.2 所示。

表 3.2 程度词词典

程度级别	举例	词数量/个	权值大小
most	万分、最为	69	2.00
very	格外、特别	42	1.50
more	比较、愈发	37	1.25
ish	稍微、略微	29	0.50
insufficiently	不多、一点儿	12	0.25
inverse	没有，不是	18	-1.00

（3）网络词典

在线评论文本因是主观性评论，表达过于口语化，所以充斥着大量的网络用语。这些网络用语大都是已有情感词典中不存在的，因此构建网络情感词典是有必要的。本文的网络词典主要通过微博、博客等收集并标注具有明显情感倾向，且为近 3 年流行的词语，结果如表 3.3 所示，以此作为情感词典的补充。

表 3.3 网络词典

情感极性	举例	词数量/个
积极	打 call、omg、skr	36
消极	diss、太难了、扎心	41

（4）停用词词典

停用词词典，是指在文本中出现的频率极高，但是对分析文本的情感倾向没有或只有极小意义的词语。例如，“的”、“了”、“上面”等助词、介词。关于停用词词典本文采用现有的百度停用词词典，在对文本进行处理时，将处理结果与其进行比对，过滤掉停用词，以保留与文本语境、情感倾向相关的词。

3.2.2 情感词典的拓展

基于词典的文本情感分析，构建全面、准确的情感词典是关键。然而分析不同领域的评论文本，其情感词典之间差异较大。所以，在对特定领域的文本

进行情感分析时，需根据实际需求对情感词典进行拓展，完善情感词典，从而提高情感极性分析的准确性。为此，本文通过情感倾向点互信息算法（SO-PMI）对领域情感词典进行拓展，同时基于 LDA 主题模型对文本的情感词进行抽取，作为领域词典的补充，从而有效地计算出文本中的情感分值。基于 SO-PMI 算法的基本流程如图 3.2 所示，具体可分为 5 个步骤。

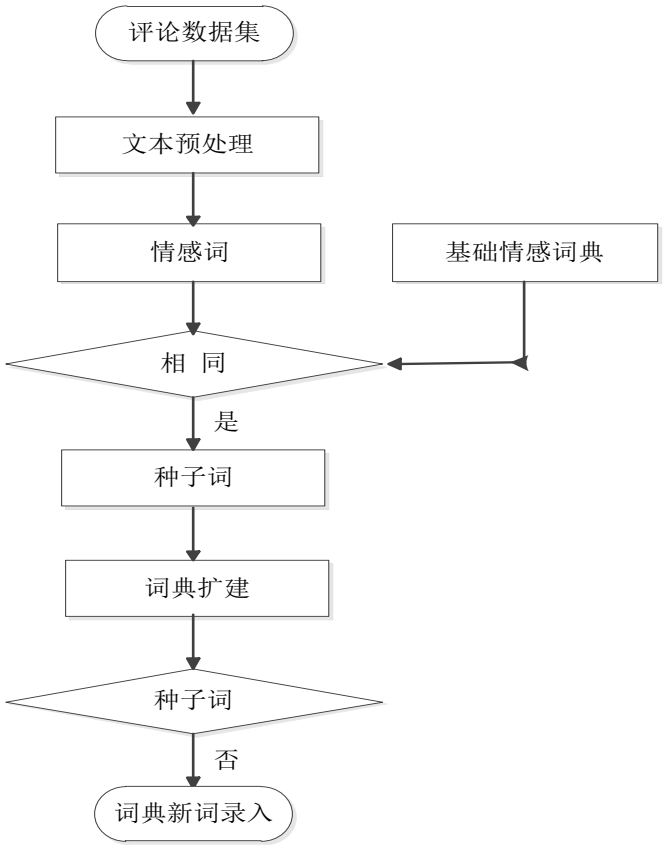


图 3.2 SO-PMI 算法的基本流程

步骤 1：将知网 HowNet 情感词典、清华大学褒贬义词典、大连理工情感词汇本体词库进行合并去重，得到基础情感词典。

步骤 2：对评论文本进行分句、分词和词性标注等预处理，收集情感词。

步骤 3：将基础情感词典与在文本中提取出的情感词进行比对，取相同的情感词作为种子词。

步骤 4：使用情感倾向点互信息（SO-PMI）算法（如公式 3.2 所示）进行情感词典的扩建。SO-PMI 算法是在点间互信息（PMI）算法（如公式 3.1 所示）的基础上以正向情感种子词（Pwords）和负向情感种子词（Nwords）为基准词，把文本中的某一个词语 w1 与 Pwords 的点间互信息减去 w1 与 Nwords 的点间互信息[70]，得到一个差值，通过差值大小判断 w1 的情感倾向。如 SO-PMI（w1）>0，

则 w1 为正向情感词，否则为负向情感词。

$$PMI(w1,w2) = log_2\left(\frac{P(w1\&w2)}{P(w1)P(w2)}\right) \tag{3.1}$$

$$SO\text{-}PMI(w1) = \sum_{pw\in Pwords} PMI(w1,pw) - \sum_{nw\in Nwords} PMI(w1,nw) \tag{3.2}$$

步骤 5：词典新词录入。

LDA 主题模型的主要思想：文档集是若干主题的混合分布，每一个主题均对应特定的特征词分布。因此，隐含主题可以看成词项的概率分布，每篇文档可表示为这些隐含主题的概率分布，最后使用概率的产生式提取出潜在主题。该模型的基本流程如图 3.3 所示，具体可分为 6 个步骤。

步骤 1：模型逐个读取文档集 W 中的每一个文档 m，其中 m 服从泊松分布。

步骤 2：依照概率结果，对文档集 W 中的每一篇文档 m 生成“Doc-topic”分布 θ。

步骤 3：对每一篇文档 m 中的各个主题形成“Topic-word”分布 Φ。

步骤 4：从文档 m 的 θ 分布中选择一个主题 z，再从 z 的 Φ 分布中选择一个词项 ω.

步骤 5：使用 Gibbs 进行参数估计。

步骤 6：输出结果。

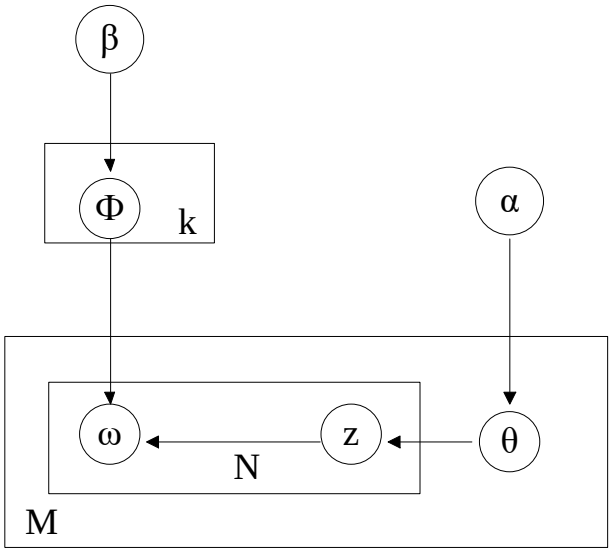


图 3.3 LDA 主题模型基本流程

3.3 文本预处理

文本预处理是情感分析的必要前提，其是对用户评论文本中存在的部分低价值甚至没有价值的信息进行清洗与过滤，从而提升情感分析的准确性。文本

预处理主要包括文本规范化、分词、词性标注、去除停用词、文本分类等。

（1）文本规范化

文本规范化，是对评论文本中的冗余字符、语法错误、重复文本等口语化的表达进行人工规范化处理。所以，进行情感分析，首先要对文本进行规范化处理，包括删除冗余字符、修正语法错误、删除每个用户的重复性文本等，最终得到规范的文本数据便于后续分析与处理。

（2）分词

对中文文本进行情感分析，首先需要进行分词，即将文本分成一个个有意义的词。目前常用的中文文本分词工具有 NLPIR-ICTCLAS 汉语分词系统、LTP 平台、Jieba 分词等。其中，Jieba 分词是用 python 语言编写的第三方分词开源库，其特点是支持三种分词模式：精确模式、全模式和搜索引擎模式，是目前受欢迎的分词工具之一。因此，本文采用 Jieba 分词对评论文本进行分词处理。

（3）词性标注

词性标注（Part-Of-Speech tagging, POS tagging），是在分词结果的基础上为所有词语确定一种最合适的词性，其结果的正确与否将会直接影响到后续的情感分析结果。本文采用 Jieba 分词对分词结果进行词性标注。

（4）去除停用词

在对中文文本分词后，会存在如“了”、“的”、“和”等对文本极性影响极小甚至没有影响的词语，为了简化文本处理的复杂性，节省存储空间，提高处理效率，需要将其设为停用词，并逐一去除。本文使用的是百度停用词表，将分词结果与其进行比对，删除停用词，保留对情感分析具有最大意义的词语。

（5）文本分类

目前，由于网络平台中评论文本的语料逐渐增多，网络新词也随之逐渐增加，情感词典因新词的缺失已经不能完全满足当下文本分类的研究需要。所以，基于情感词典的文本分类需要不断对情感词典进行新词扩充，而过多的人工标注扩充词条信息大大增加了人工成本。基于机器学习的文本分类算法不需要过多的人工成本投入，并且效率高，是近几年较为流行的情感分类方法。文本情感分类常用的机器学习算法有以下几种：随机森林（Random Forest）、支持向量机（SVM）、逻辑回归（Logistic Regression）等。其中，相比于其他分类算

法，逻辑回归因不需要缩放输入特征、不需要太大的计算量，且很容易调整特征等优点，是一种常用的机器学习分类算法。因此，本文采用逻辑回归进行文本分类。

基于逻辑回归在线评论情感分类模型的整个流程如图 3.4 所示，具体可分为 5 个步骤。

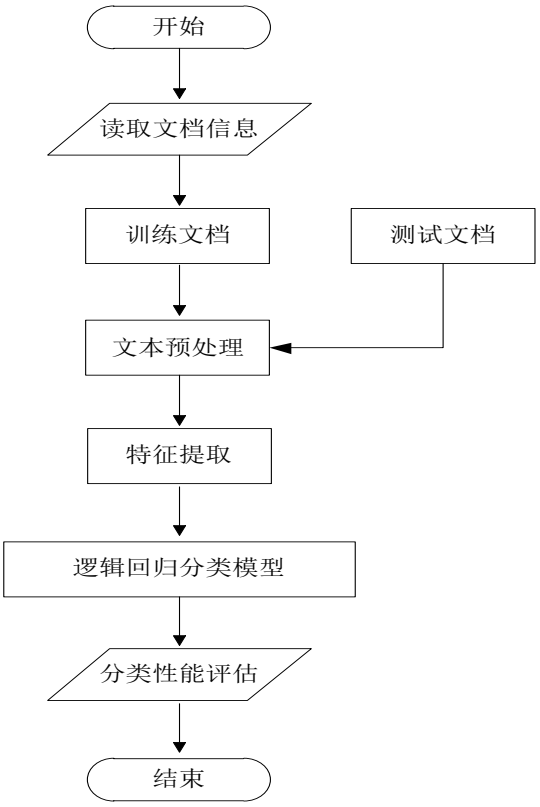


图 3.4 逻辑回归模型基本流程

步骤 1：数据集划分。将好评和差评文本 70%的数据作为训练集，剩下的 30%作为测试集，中评文本为待分类集。

步骤 2：文本的预处理。首先对全部文本进行清洗和筛选等规范化处理，然后对其进行分词、词性标注、去除停用词等处理工作。

步骤 3：特征工程。使用词向量模型将能描述文本的特征词转换成向量形式，便于模型的特征训练。

步骤 4：逻辑回归算法的训练与性能评估。

步骤 5：中评的文本分类。

3.4 基于词典的在线评论情感分析

3.4.1 计算方法

基于词典的文本情感分析方法，是把情感词语进行极性的归纳整理，并赋予适当的权重，当文本输入后与词典的所有词语进行匹配，通过权值加和的结果判断文本的情感倾向。

文本的情感倾向由其情感值的大小来反映，而情感值的计算包括情感词、程度副词、否定词、网络热词等词语的加和，通过综合计算情感词以及其他相关词语的权重，反映该文本的情感倾向。因此，本文使用基于词典的情感值计算函数[71]（如公式 3.3 所示），综合文本中情感词和修饰词的计算结果，进而反映其情感倾向。

$$F = \sum_{i=1}^{n}[f(x_i)\prod_{j=1}^{m}a_{ij}] \tag{3.3}$$

在公式（3.3）中，$f(x_i)$表示第 i 个情感词的极值，a_{ij}表示第 i 个情感词对应的修饰词的权值，m 和 n 分别代表第 i 个情感词的修饰词个数和文本中情感词的个数。

3.4.2 计算结果

（1）数据来源

本文所使用的文本数据均来自“京东-农资频道”网站的评论数据，以“兰州百合”为例，其评价的基本形式如图 3.5 所示。通过网络爬虫软件--八爪鱼采集器，以“兰州百合”、“天水苹果”、“苦水玫瑰”等 36 种具有地方特色的农产品主题词为检索词，按照用户名、评分、评论文本、类别、购买日期等信息对好、中和差三种类型的评论文本进行分别爬取，共采集到 15284 条评论。

图 3.5　甘肃省特色农产品评论数据

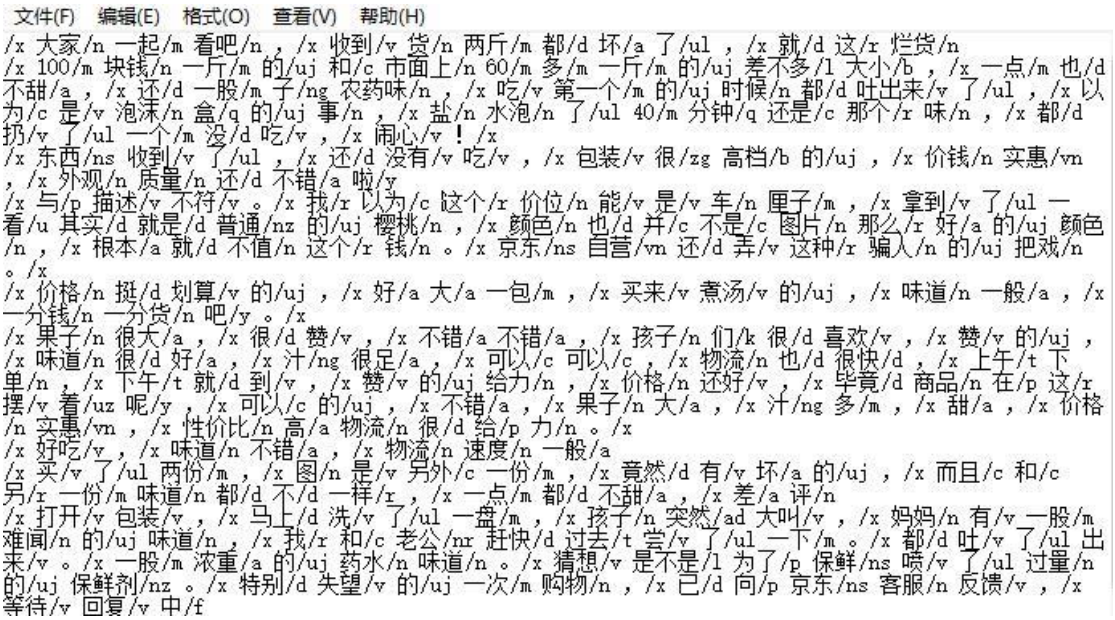
（2）文本预处理

为避免数据在表达上的随意性，需对其进行清洗与筛选，过程包括：第一，针对原始数据中不包含属性特征词（如口感、包装等）的文本进行删除；第二，

对未被识别（乱码）的符号进行删除；第三，对同一消费者评论的重复性文本进行删除。最终对预处理后的 12743 条有效数据，进行处理与分析，部分结果如图 3.6 所示。此外，通过结巴分词对文本进行分词和词性标注，部分结果如图 3.7 所示。

用户名	评分	评论	类别	购买日期
j***r	2	大家一起看吧，收到货两斤都坏了，就这烂货	天水樱桃 JJ级大果 2斤	2020/2/28 10:41
老歌乍听	2	100块钱一斤的和市面上60多一斤的差不多大小，一点也不	天水樱桃 巨无霸JJJ级	2020/2/22 12:46
追王凯老师的影迷	5	东西收到了，还没有吃，包装很高档的，价钱实惠，外观质	兰州鲜百合 1斤/500g(真	2020/2/14 18:05
G***a	2	与描述不符。我以为这个价位能是丰厘子，拿到了一看其实	天水樱桃 XL级 2斤装	2020/4/1 12:05
蓝色沸点55	3	价格挺划算的，好大一包，买来煮汤的，味道一般，一分钱	庆阳黄花菜 500g/袋	2020/1/16 1:09
孙玉备	5	果子很大，很赞，不错不错，孩子们很喜欢，赞的，味道很	静宁苹果 24枚75中果家	2020/1/10 11:45
午***8	3	好吃，味道不错，物流速度一般	民勤人参果 中果5斤	2020/1/9 21:29
j***d	2	买了两份，图是另外一份，竟然有坏的，而且和另一份味道	天水樱桃 巨无霸JJJ级	2020/1/4 17:39
j***o	2	打开包装，马上洗了一盘，孩子突然大叫，妈妈有一股难闻	天水樱桃 巨无霸JJJ级	2020/1/2 20:30
133*****552_p	5	价格优惠，颜色干净，很好	兰州灰豆 5斤装	2020/5/16 17:18
中国wzy	3	品质一般般	会宁黑枸杞 500g	2020/5/16 16:48
jd_448077700	1	质量实在差的难以想象，物流也慢的可以，顺丰发生鲜竟然	天水苹果 花牛5斤	2020/5/16 15:56
樊猫儿	4	搞活动时候购买了一箱，物流速度很快，到家打开一看包装	静宁苹果 24枚75中果家	2020/5/15 18:33
a***9	5	很耐泡，喝着口味清香，好喝，料很足，不错	岷县当归 250g	2020/5/12 16:43
j***7	1	质量差的很，软绵绵的，不新鲜，还有烂果子，差评	天水苹果 花牛5斤	2020/5/12 15:12
jd_good2016	4	这款大樱桃已多次购买了，酸甜爽口，非常好吃；这次买的	天水樱桃 巨无霸JJJ级	2020/5/11 21:22
再人人	5	个头真的超级超级大！我的天啊~而且水分很足，第一口吃	庆阳苹果 25枚中果家庭	2020/5/10 10:58
一撮岁月	5	性价比高口感好，会回购的味道好，质量好	兰州百合干礼盒240g*2	2020/5/9 13:09
大***把	3	货到了，口感挺好的，味道满意，包装一般	民勤人参果 中果8斤	2020/5/7 10:14

图 3.6 文本预处理结果


文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

/x 大家/n 一起/m 看吧/n ，/x 收到/v 货/n 两斤/m 都/d 坏/a 了/ul ，/x 就/d 这/r 烂货/n
/x 100/m 块钱/n 一斤/m 的/uj 和/c 市面上/n 60/m 多/m 一斤/m 的/uj 差不多/l 大小/b ，/x 一点/m 也/d 不甜/a ，/x 还/d 一股/m 子/ng 农药味/n ，/x 吃/v 第一个/m 的/uj 时候/n 都/d 吐出来/v 了/ul ，/x 以为/c 是/v 泡沫/n 盒/q 的/uj 事/n ，/x 盐/n 水泡/n 了/ul 40/m 分钟/q 还是/c 那个/r 味/n ，/x 都/d 扔/v 了/ul 一个/m 没/d 吃/v ，/x 闹心/v ！/x
/x 东西/ns 收到/v 了/ul ，/x 还/d 没有/v 吃/v ，/x 包装/v 很/zg 高档/b 的/uj ，/x 价钱/n 实惠/vn ，/x 外观/n 质量/n 还/d 不错/a 啦/y
/x 与/p 描述/v 不符/v 。/x 我/r 以为/c 这个/r 价位/n 能/v 是/v 车/n 厘子/m ，/x 拿到/v 了/ul 一看/u 其实/d 就是/d 普通/nz 的/uj 樱桃/n ，/x 颜色/n 也/d 并/c 不是/c 图片/n 那么/r 好/a 的/uj 颜色/n ，/x 根本/a 就/d 不值/n 这个/r 钱/n 。/x 京东/ns 自营/vn 还/d 弄/v 这种/r 骗人/n 的/uj 把戏/n 。/x
/x 价格/n 挺/d 划算/v 的/uj ，/x 好/a 大/a 一包/m ，/x 买来/v 煮汤/v 的/uj ，/x 味道/n 一般/a ，/x 一分钱/n 一分货/n 吧/y 。/x
/x 果子/n 很大/a ，/x 很/d 赞/v ，/x 不错/a 不错/a ，/x 孩子/n 们/k 很/d 喜欢/v ，/x 赞/v 的/uj ，/x 味道/n 很/d 好/a ，/x 汁/ng 很足/a ，/x 可以/c 可以/c ，/x 物流/n 也/d 很快/d ，/x 上午/t 下单/n ，/x 下午/t 就/d 到/v ，/x 赞/v 的/uj 给力/n ，/x 价格/n 还好/v ，/x 毕竟/d 商品/n 在/p 这/r 摆/v 着/uz 呢/y ，/x 可以/c 的/uj ，/x 不错/a ，/x 果子/n 大/a ，/x 汁/ng 多/m ，/x 甜/a ，/x 价格/n 实惠/vn ，/x 性价比/n 高/a 物流/n 很/d 给/p 力/n 。/x
/x 好吃/v ，/x 味道/n 不错/a ，/x 物流/n 速度/n 一般/a
/x 买/v 了/ul 两份/m ，/x 图/n 是/v 另外/c 一份/m ，/x 竟然/d 有/v 坏/a 的/uj ，/x 而且/c 和/c 另/r 一份/m 味道/n 都/d 不/d 一样/r ，/x 一点/m 都/d 不甜/a ，/x 差/a 评/n
/x 打开/v 包装/v ，/x 马上/d 洗/v 了/ul 一盘/m ，/x 孩子/n 突然/ad 大叫/v ，/x 妈妈/n 有/v 一股/m 难闻/n 的/uj 味道/n ，/x 我/r 和/c 老公/nr 赶快/d 过去/t 尝/v 了/ul 一下/m 。/x 都/d 吐/v 了/ul 出来/v 。/x 一股/m 浓重/a 的/uj 药水/n 味道/n 。/x 猜想/v 是不是/l 为了/p 保鲜/ns 喷/v 了/ul 过量/n 的/uj 保鲜剂/nz 。/x 特别/d 失望/v 的/uj 一次/m 购物/n ，/x 已/d 向/p 京东/ns 客服/n 反馈/v ，/x 等待/v 回复/v 中/f

图 3.7 文本分词及词性标注

预处理后的 12743 条评论文本中，好评有 7349 条，中评有 4268 条，差评有 1126 条，如图 3.8 所示。由图可知，中评文本的占比为 33%，比重较大。通过进一步分析发现，中评文本也包含正面和负面的情感倾向。因此，需要对中评文本进行分类，将正面、负面的情感倾向分别归类到对应的好评、差评中，作为评论的补充。本文采用逻辑回归算法进行文本情感分类，该算法的主要参数、性能评估以及部分分类结果如表 3.4、表 3.5 和图 3.9 所示。

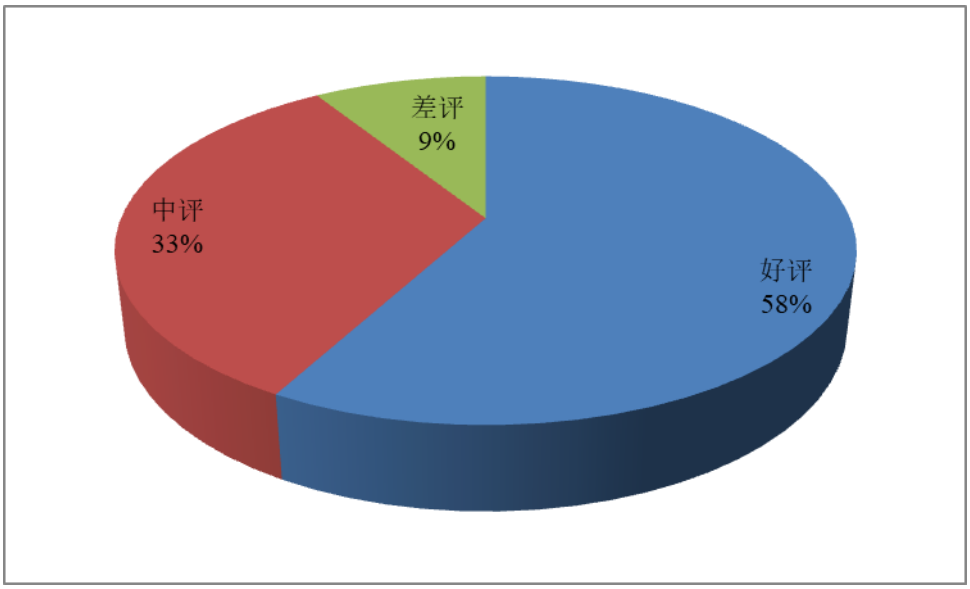


图 3.8 好、中、差评文本占比

表 3.4 主要参数

参数	参数值
Vector model	Word2Vector
Penalty	L2
Multi_Class	OVR
C	0.8
Max_iter	100

表 3.5 模型性能评估

状态	准确率	待分类集总数
交叉验证前	70.25%	4268
交叉验证后	71.72%	4268

文本分类的部分结果如图 3.9 所示。其中，分类结果代表情感极性，1、2 表示为差评的文本，4、5 表示为好评的文本，3 表示为中评的文本。经统计，中评文本分类后的好评、中评和差评文本的数量分别为 943、1965、1360，占比分别为 22%、46%、32%。因中评文本对情感极性分析无太大意义，故删除。最终将分类后的好评和差评文本补充到对应的评论文本中。

用户	评分	评论	分类结果
他***x	star3	味道太重了，喝了想吐，感觉煲汤放里面应该可以，不适合当茶喝	1
洁***8	star3	刚收到，现在这个季节本来有点上火长溃疡了，喝了好像更上火啊。	1
京***6	star3	颗粒小，虫眼还是有的，这是第一次体验这么差，如图所示，后悔了，白花冤枉钱。	1
T***者	star3	味道基本过得去	3
T***y	star3	由于研究需要，同时购买了，总的来看，还是甘肃产的黑枸杞效果更好。	5
大***哥	star3	这次的东西有点泛黄，受潮了，不知道是不是季节的问题，过期产品	3
j***我	star3	没什么效果。有点中药的味道，喝不不习惯	3
Dsz57	star3	味道马马虎虎，唉很无奈	3
1***m	star3	不好！好多小粒的，都泡不开！！！应该是最低级的枸杞卖给我了	1
淡***心	star3	为什么泡出来的水是绿色的？不应该是蓝色或者紫色的吗？怀疑是假货	1
j***i	star3	感觉一般，没有去网上找人查证，不如以前在市场买的	3
宴***楚	star3	黄芪收到了，很大一罐，但是真的好小啊！说实话，品质差	1
C***9	star3	这个味儿真不是一般人能喝的了，一小包泡一大壶，像中药，口感差	1
jd_138639fym	star3	口感有点勉强，表示没有办法	3
一***h	star3	这颜色有点不对	1
x***2	star3	口感有点勉强	3
j***u	star3	品相可以，种类齐全	5
淡***心	star3	一般般吧！过得去哪种感觉。	3
j***E	star3	这款黄芪我家已经买了几次了，物美价廉，好大一桶，京东的物流快，都是次日达，该款黄芪来自甘肃陇西	5

图 3.9　中评文本分类结果

（3）农产品领域词典拓展

1）基于 SO-PMI 算法的农产品领域词拓展

基于情感倾向点互信息（SO-PMI）算法，得到的农产品领域情感词典，其部分积极和消极情感词典输出结果如图 3.10 和 3.11 所示。

图 3.10　基于 SO-PMI 的领域积极词典拓展

图 3.11　基于 SO-PMI 的领域消极词典拓展

2）基于 LDA 主题模型的农产品领域词典补充

本文采用 TF-IDF 词向量模型，运用基于 Gibbs 采样的 LDA 模型对语料集

进行处理。关于先验超参数 α 和 β 的选取，参考其他相关文献[71,72]，最终确定 α=50/K，β=0.01。由于本文的数据类别已知，主题数 K 直接给出，便于实验验证。基于 Gibbs 采样的 LDA 模型部分主题-词见表 3.6，部分主题概率分布见表 3.7。

表 3.6 基于 Gibbs 采样的 LDA 模型主题-词

Topic1	Topic2	Topic3	Topic4	…
质量	价格	超市	感觉	…
推荐	购物	便宜	甜	…
味道	好吃	还好	营养	…
购买	新鲜	没想到	味道	…
习惯	朋友	老人	灰豆	…
…	…	…	…	…

表 3.7 部分文本主题概率分布

	Topic1	Topic2	Topic3	Topic4	…
1	0.00874938	0.03383897	0.04242369	0.00476928	…
2	0.02112516	0.03439895	0.04563891	0.03243255	…
3	0.02985551	0.03863336	0.05735442	0.08263235	…
4	0.03212478	0.03922115	0.05915656	0.11956655	…
5	0.03309149	0.04117295	0.07021045	0.05431875	…
…	…	…	…	…	…

由表 3.6 可知，每个主题词之间的区分是比较明显的。Topic1 中的主题词主要是回头客经常购买的产品，Topic2 主要是送朋友的产品，Topic3 主要是打折促销的产品，Topic4 主要是豆类产品等。为了便于下一步分析，本文对特征主题词进行汇总归类，结果见表 3.8。

表 3.8 农产品特征分类

主题类别	主题词
质量	质量，日期，品质，营养，效果
味道	味道，口味，香味，口感
价格	价格，价钱，性价比

外观	外观，颗粒，图片，汤色，颜色，个头
物流	物流，速度，送货，配送，包装
服务	服务，态度，活动，评价

根据拓展的领域情感词典，综合基础情感词典与网络词典。得到最终的情感极性词典如表 3.9 所示。

表 3.9 情感词典

情感极性	数量/个
积极情感词典	6724
消极情感词典	6598

（4）情感值计算

本文根据公式（3.3）的文本情感值计算方法，计算出农产品各特征的在线评论情感值，部分结果如表 3.10 所示。各特征的情感值的范围在 1-5 之间，当情感值大于 3 时，该特征的评论是积极的；当情感值小于 3 时，该特征的评论是消极的；当情感值等于 3 时，该特征的评论是中性的。基于用户-特征评分矩阵，需转化为用户-产品评分矩阵。处理原则包括：对同一种产品下的各特征评论情感值求平均，部分结果如表 3.11 所示。

表 3.10 用户-特征评分矩阵

userID	itemID	geres	rating
1	4	Taste	2.94
1	13	Quality	1.99
2	3	Logistics	2.94
2	13	Appearance	2.52
2	13	Price	1.79
2	6	Price	2.31
…	…	…	…

表 3.11 用户-产品评分矩阵

userID	itemID	rating
1	4	2.94
1	13	1.99

2	3	2.94
2	6	2.31
2	13	2.16
…	…	…

3.5 本章小结

本章对情感词典的基本流程进行介绍，之后对情感词典构建与拓展的方法进行了概述，并使用情感倾向点互信息算法和 LDA 主题模型拓展农产品领域情感词典，最后计算情感值得到农产品情感评分矩阵，为下一步的产品推荐做准备。

4 融合矩阵分解和改进巴氏系数的混合推荐算法

本文将 BMF 矩阵分解算法与改进巴氏系数的混合协同推荐算法相结合。利用 BMF 矩阵分解算法对评分矩阵进行缺失值填充，并基于改进巴氏系数的混合协同推荐算法预测目标用户的评分。因农产品推荐领域没有标准的数据集可以参考，故本章选择 MovieLens 标准数据集来验证算法的准确性和可靠性。

4.1 算法构建思想

基于协同过滤推荐算法是在用户-物品评分矩阵基础上进行推荐，评分矩阵稀疏会直接影响推荐效果。协同过滤推荐算法具体分为基于用户和基于物品的两种算法，其推荐核心是寻找相似用户邻居集或相似项目集进行推荐，所以相似性度量方法显得极为重要。传统的基于用户和基于物品的两种协同过滤算法在推荐过程中过于单一，需要将两种算法进行融合改进。因此，针对如上问题，本文基于偏置的矩阵分解算法解决评分矩阵数据稀疏问题，加入调和平均权值因子与用户偏好两个因素对巴氏系数相似性度量方法进行改进。最终构建融合矩阵分解和改进巴氏系数的混合推荐算法进行物品推荐。

4.2 算法构建过程

融合矩阵分解和改进巴氏系数的混合推荐算法是在评分矩阵基础上，使用基于偏置的矩阵分解算法（BMF）对评分矩阵进行部分填充，运用改进巴氏系数寻找最相似的邻居集和项目集，并融合基于用户和基于项目的两种协同过滤推荐算法，预测用户评分，其基本流程图如 4.1 所示。

（1）BMF 缺失评分值填补

BMF 算法考虑到用户和商品的偏置部分，并以概率的形式描述评分矩阵与用户、商品分别隐含的特征矩阵。因评分矩阵非常稀疏，BMF 算法通过降维的思想将评分矩阵分解为用户特征矩阵 U 和商品特征矩阵 V，从而对缺失值进行填补，具体如公式（4.1）所示。

$$\hat{R}_{ij} = \mu + b_i + b_j + U_i^T V_j \tag{4.1}$$

在公式（4.1）中，μ 表示评分矩阵所有评分的平均值，b_i和b_j分别表示基于用户的偏置和基于商品的偏置。然后，利用公式（4.2）构造损失函数，使已知评分与基于 BMF 的填补评分误差最小。

$$E=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M}(R_{ij}-u-b_i-b_j-U_i^TV_j)^2+\frac{\lambda_U}{2}\sum_{i=1}^{N}||U_i||_{Fro}^2+\frac{\lambda_V}{2}\sum_{j=1}^{M}||V_j||_{Fro}^2 \quad (4.2)$$

在公式（4.2）中，$\lambda_U=6^2/6_U^2$，$\lambda_V=6^2/6_V^2$，6_U^2和6_V^2分别为用户特征矩阵 U 和商品特征矩阵 V 的方差，$||U_i||_{Fro}^2$为用户特征矩阵 U 的 F 范数，$||V_j||_{Fro}^2$为商品特征矩阵 V 的 F 范数。利用随机梯度下降法求解 E。基于 BMF 算法计算出每个用户对某一商品的评分值，从而得到较完整的评分数据。

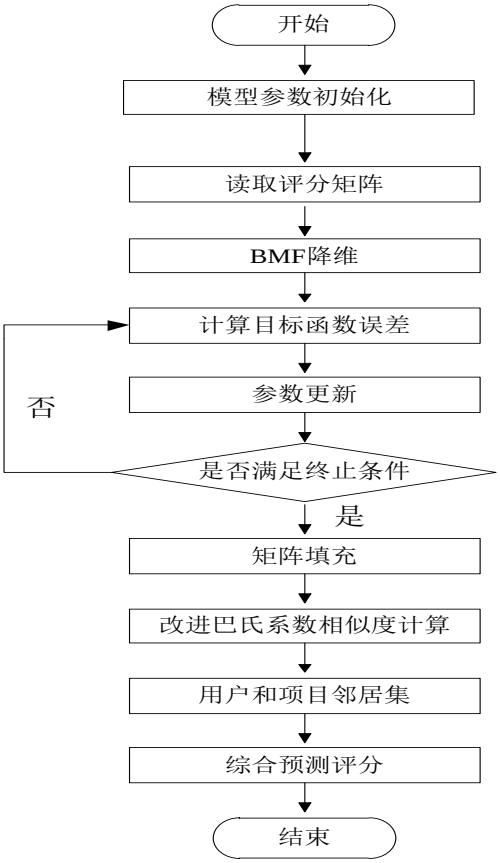


图 4.1 算法构建的基本流程

（2）改进巴氏系数相似度

巴氏系数在推荐领域的相似度计算，是将用户 U 和用户 V 的所有评分值的全局相似和局部相似的合并计算，如公式（4.3）所示。

$$Sim_{bc}(\mathrm{U},\mathrm{V})=\sum_{h=1}^{m}\sqrt{\widehat{P_{uh}}*\widehat{P_{vh}}}*\frac{[(R_{ui}-\bar{R}_u)*(R_{vi}-\bar{R}_v)]}{\sqrt{\sum_{i\in U}(R_{ui}-\bar{R}_u)^2}*\sqrt{\sum_{i\in V}(R_{vi}-\bar{R}_v)^2}} \quad (4.3)$$

在公式（4.3）中，m 为所有评分值，用户 u 和用户 v 的评分值为 h 的概率密度分布分别为$\widehat{P_{uh}}$和$\widehat{P_{vh}}$，R_{ui}是用户 U 对项目 i 的评分，$\bar{R}_u$是用户 U 对所有项目评分的平均值。

但该方法存在两大缺陷：第一，当$R_U \cap R_V = \emptyset$时，即用户 u 和用户 v 的评分值不一致时，此时巴氏系数为局部相似度，导致结果不准确。第二，没有考虑用户的偏好。因此，通过以上分析，本文加入调和平均权值因子和用户偏好对巴氏系数相似度度量方法进行改进，如公式（4.4）所示。

$$Sim_{hs}(\mathrm{U},\mathrm{V}) = \sum_{h=1}^{m} \sqrt{\widehat{P_{uh}} * \widehat{P_{vh}}} * \frac{2 * m_u * n_v}{m_u + n_v} * \frac{[(\mathrm{R}_{ui} - \bar{R}_u) * (\mathrm{R}_{vi} - \bar{R}_v)]}{\sqrt{\sum_{i\in U}(\mathrm{R}_{ui} - \bar{R}_u)^2} * \sqrt{\sum_{i\in V}(\mathrm{R}_{vi} - \bar{R}_v)^2}} + \exp\left(-\frac{\sum_{i\in I_{uv}}|r_{ui} - r_{vi}|}{|I_{uv}|} * |\bar{r}_u - \bar{r}_v|\right) * \frac{2 * |I_u| \cap |I_v|}{|I_u| + |I_v|} \quad (4.4)$$

在公式（4.4）中，$m_u = -\frac{1}{1\mathrm{n}\,\mathrm{C}} * \sum_{i=1}^{n} P_{ui} * lnP_{ui}$，$P_{ui}$是每个分数在用户 u 的所有评分中出现的频率，$1\mathrm{n}\,\mathrm{C}$是归一化系数，C 的取值取决于评分的范围，由于通常评分在 1-5 之间，故 C 取 5，m_u表示用户 u 的权值因子，n_v表示用户 v 的权值因子。I_{uv}表示用户 u 和用户 v 评分的总次数，$|I_u|$表示用户 u 评分的次数，$|I_v|$表示用户 v 评分的次数。

基于改进巴氏系数相似度度量方法，根据用户和项目的相似程度找到更符合其兴趣和特征的邻居集合。

（3）混合协同推荐 HCF 算法

混合协同推荐 HCF 算法，是考虑到基于用户或基于项目的协同过滤算法的推荐各会忽略掉一些信息，从而导致推荐结果不准确。因此，本文结合基于用户和基于项目的协同过滤两种推荐算法，加入权重系数综合计算用户 u 对项目 v 的预测评分，如公式（4.5）所示。

$$P_{uv} = \alpha(\overline{r_u} + \frac{\sum_{a\in U_x} sim(a,u)(r_{av} - \overline{r_a})}{\sum_{a\in U_x} sim(a,u)}) + (1-\alpha)\left(\overline{r_v} + \frac{\sum_{i\in V_x} sim(i,v)(r_{ui} - \overline{r_i})}{\sum_{i\in V_x} sim(i,v)}\right) \quad (4.5)$$

在公式（4.5）中，α为权重系数，$sim(a,u)$表示用户 a 和用户 u 的相似度，r_{av}表示 u 的邻居 a 对项目 v 的评分，$\bar{r}_a$和$\bar{r}_u$分别表示用户 a 和用户 u 对所有项目的平均评分。

最终基于混合推荐 BMF-HCF 算法，综合计算出用户对未评分项目的评分，从而提高推荐的准确率。

4.3 算法评估

（1）数据来源

为了保证算法的严谨性，本文采用 MovieLens 标准数据集[①]进行算法的性能评估。MovieLens 数据集是在推荐领域广泛使用的公开数据集，该数据集包含 610 名用户对 9742 部电影的评分情况，评分范围在 1-5 之间，评分记录有 100836 条。该数据集的数据稀疏度 Φ 为：

$$\Phi = 1 - \frac{100836}{610 \times 9742} = 98.3\%$$

由此可见，该数据集的评分矩阵是非常稀疏的，所以对其进行矩阵填充是有必要的。

（2）算法评估

1）基于偏置的矩阵分解 BMF 算法评估

为了得到较理想的实验效果，本文利用 BMF 算法在数据集上进行不同比例的填充，并在不同比例下与 PMF 算法进行比较，用平均绝对误差（MAE）作为衡量指标，从而确定合适的填充比例得到更准确的推荐效果，其填充的预测误差变化曲线对比如图 4.2 所示。

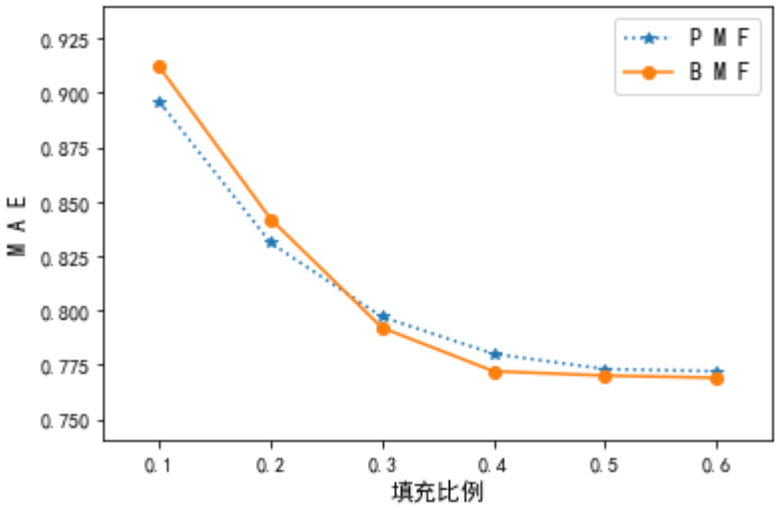


图 4.2 BCF 与 PMF 填充误差对比结果

从图 4.2 可以看出，不同填充比例下，BMF 算法的数据填充效果总体好于 PMF 算法，且当填充比例为 0.4 时，模型的性能趋向稳定。

2）改进巴氏系数算法评估

本文将改进的巴氏系数相似度算法与余弦相似度、皮尔逊相似度以及传统巴氏系数分别基于用户和基于项目的两种协同过滤算法进行比较，其结果如图

[①] 数据来源：http://movielens.org

4.3 和图 4.4 所示。

从图 4.3 和图 4.4 可以看出，随着 K 值的逐渐增大，其 MAE 值随之减少，说明预测准确率较高。并且在 K 相同时，本文提出的改进巴氏系数算法所得出的结果比其他三种相似度结果最小，说明取得了更好的预测效果。

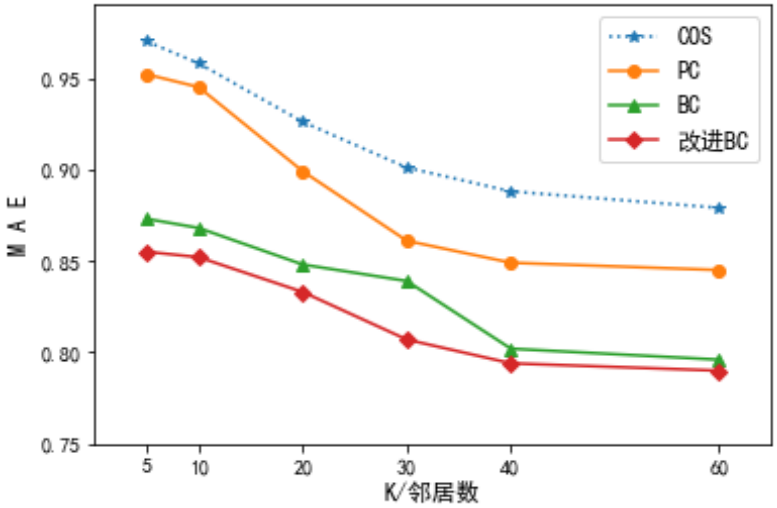


图 4.3 基于用户的协同过滤各相似度比较

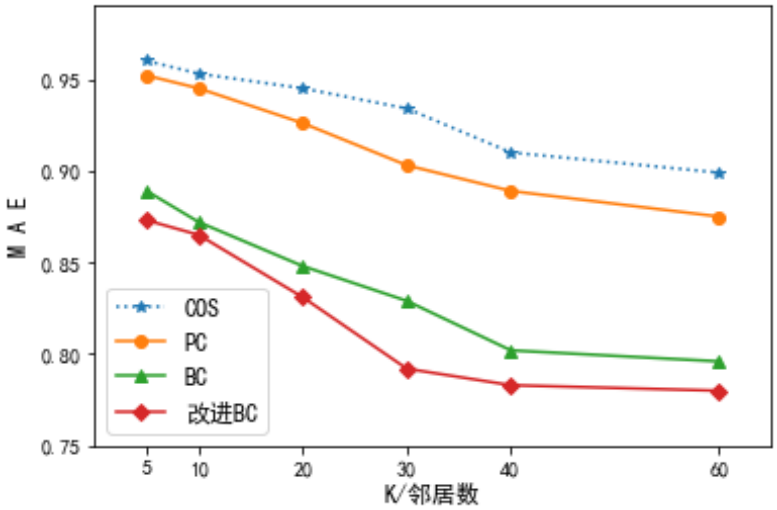


图 4.4 基于项目的协同过滤各相似度比较

3）混合协同推荐 HCF 算法评估

随机抽取数据集中 30 个用户，通过改变权重系数 α 的取值来分析其对本文提出的基于用户和基于项目的混合推荐算法（HCF）推荐精度的影响，以最终计算的 MAE 平均值来确定最优系数，如图 4.5 所示。

由图 4.5 可知，在基于用户和基于项目的混合推荐算法中，UCF 的权重系数为 0.4 和 ICF 的权重系数为 0.6 时，MAE 值最小，HCF 算法的性能最佳。

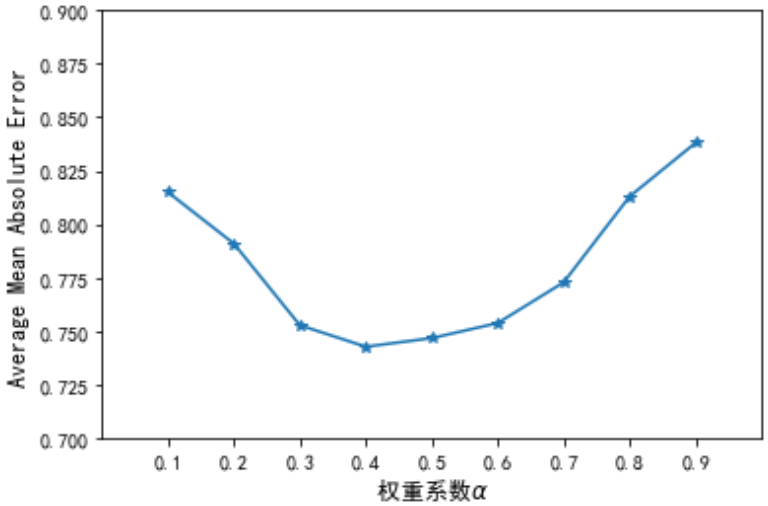


图 4.5　权重系数 α 的取值对 HCF 算法的影响

4）混合推荐 BMF-HCF 算法评估

本文使用 MovieLens 数据集对 BMF-HCF 混合推荐算法进行评估。算法的运行结果如图 4.6 和图 4.7 所示。图 4.6 和图 4.7 分别表明不同算法在 MovieLens 数据集上 MAE 和 RMSE 的对比结果。由图可知，总体上基于用户的协同过滤（UCF）、基于项目的协同过滤（ICF）、基于用户和项目的混合推荐（HCF）、本文提出的混合推荐（BMF-HCF）四种算法的训练结果的指标随着邻居数目 K 的增大都呈现减小趋势，且偏差逐渐缩小。BMF-HCF 算法的各指标误差相比于其他三种推荐算法有较大改进，并且在邻居数 K 为 40 时逐渐趋向稳定。综上，本文提出的 BMF-HCF 算法具有较好的推荐准确性，在 MAE 和 RMSE 衡量指标上与 UCF、ICF、HCF 相比具有较好的表现。

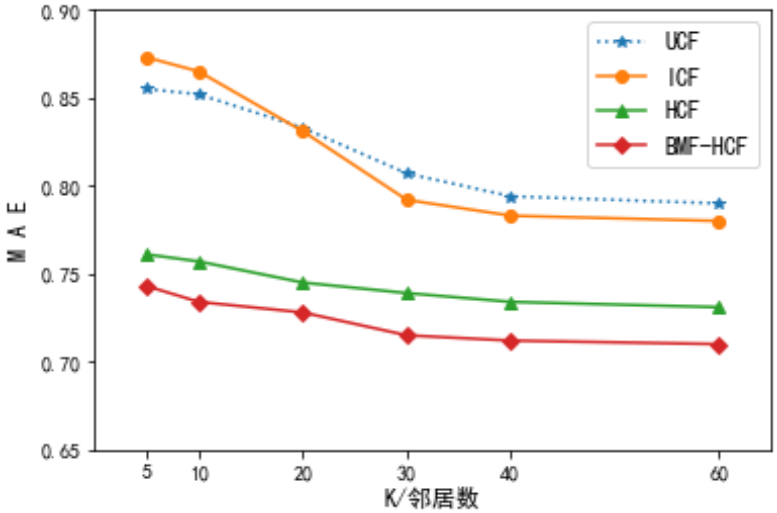


图 4.6　MovieLens 数据集上各算法的 MAE 值对比

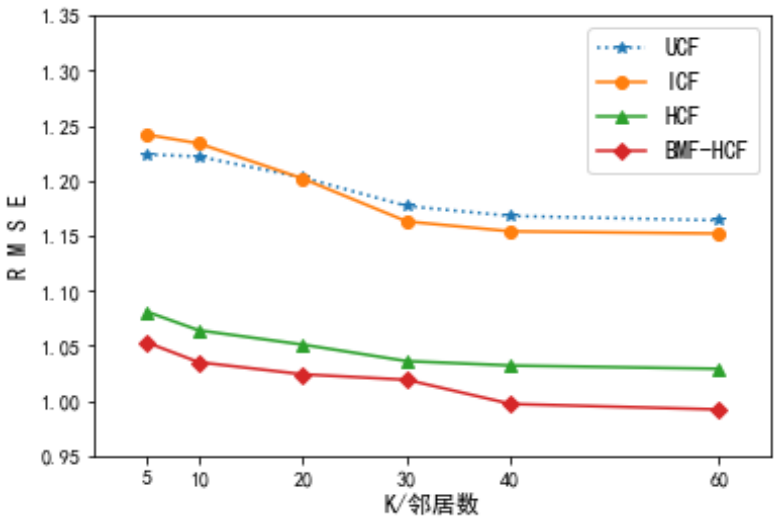


图 4.7 MovieLens 数据集上各算法的 RMSE 值对比

4.4 本章小结

本文首先介绍了融合矩阵分解和改进巴氏系数的混合推荐算法的构建思想与过程；然后对 BMF、改进的巴氏系数以及两种协同过滤混合推荐算法分别在标准数据集 MovieLens 上进行评估；最后在改进的巴氏系数相似度基础上，将 BMF-HCF 与 UCF、ICF、HCF 三种推荐算法作对比，结果显示，本文构建的 BMF-HCF 混合推荐算法在 MAE、RMSE 上具有明显优势。

5 基于情感分析的农产品个性化推荐模型的构建

个性化推荐是电商平台服务用户的重要环节之一。对于农产品电商平台来说，提供符合用户兴趣偏好的农产品尤为重要。用户在平台购买农产品过程中，通过评分来表达对产品、平台服务的评价，是平台进行用户兴趣分析，从而进行产品推荐的主要依据。但是，基于评分对用户兴趣的分析，并进行推荐有两方面缺陷。一方面，用户对产品的需求是多方面的，比如部分用户注重产品口感，部分用户注重产品质量安全，还有部分用户注重产品价格等，不同的需求产生的评价会有所偏差，即使评分相同未必代表他们对产品的兴趣相同。另一方面，评分一旦评价，后续无法对其进行更改与追加评论，这限制了用户的评价需求。而在线评论可以弥补评分的不足，原因有两点：第一，在线评论是对产品评价的细化，用户可以通过评论的方式表达在平台购买过程中，对产品从各个角度进行评价，其中包括对产品特征的观点评价。而不同的用户对于相同产品的特征需求是不同的，所以，在线评论蕴含着用户对产品的多方面需求。第二，平台上没有对评论的限制，用户在使用产品过程中，可以在平台上以追加评论的方式表达自己对产品的观点，对之前观点进行补充或纠正，从而完善了产品评价。

因此，在线评论对于用户的兴趣挖掘很有必要。本文以农产品在线评论为研究对象，采用情感分析方法将用户的情感偏好量化到农产品属性特征层面，确定不同用户的兴趣偏好，并与推荐方法相结合，从而更好的提升农产品推荐的准确率。

5.1 模型的构建

本文构建的基于在线评论情感分析的农产品个性化推荐模型，是利用平台上的在线评论对用户的兴趣偏好进行深入挖掘，并结合矩阵分解和改进巴氏系数的混合推荐算法，建立基于评论的情感分析农产品个性化推荐模型，从而进行农产品的推荐。对于用户的兴趣偏好挖掘使用的是情感分析方法，首先需要确定用户关注的农产品特征，然后基于词典的情感分析得出特征情感值，将每个特征情感值综合得到最终的产品情感评分。同时，将情感评分作为混合推荐算法的输入数据，通过混合推荐算法进行评分矩阵填充、相似度计算、目标用户评分预测等过程，为用户进行农产品的推荐。

5.1.1 模型的构建思想

前期相关学者的研究表明[17][20][50]，在线评论文本中所蕴含的用户主观性情感因素对产品的推荐具有重要的影响。因此，本文构建的推荐模型，充分考虑用户评论的情感因素对推荐产生的影响，通过文本分类、情感词典的文本分析方法得到用户的情感评分矩阵。此外，通过梳理以往的文献可以发现，在推荐领域，单一模型很难达到理想的推荐效果，只有结合不同的模型，对传统的推荐模型进行改进，才能提高推荐的准确率。所以，本文综合考虑了用户在线评论的情感因素与推荐算法的结合。在情感评分矩阵基础上，首先因数据稀疏问题，使用基于偏置的矩阵分解算法对评分矩阵进行部分填充；其次，因传统相似性度量方法过分依赖共同评分项，而且存在忽略用户整体偏好等问题，加入调和平均权值因子与用户偏好对巴氏系数相似性度量方法进行改进，寻找更加相似的邻居用户或项目；最后，因基于用户的协同过滤算法依据相似邻居预测用户偏好而对没有邻居的用户推荐准确度较低，且基于项目的协同过滤算法推荐的是历史相似项目而多样性不足、新颖性低，为此对基于用户与基于项目的两种协同过滤推荐算法进行融合。最终，构建基于情感分析的农产品个性化推荐模型。

5.1.2 模型的构建过程

基于情感分析的农产品个性化推荐模型是在情感评分矩阵基础上，使用基于偏置的矩阵分解算法（BMF）对评分矩阵进行部分填充，运用改进巴氏系数寻找最相似的邻居集和项目集，并融合基于用户和基于项目的两种协同过滤推荐算法，进行农产品的个性化推荐，其基本流程图如 5.1 所示。

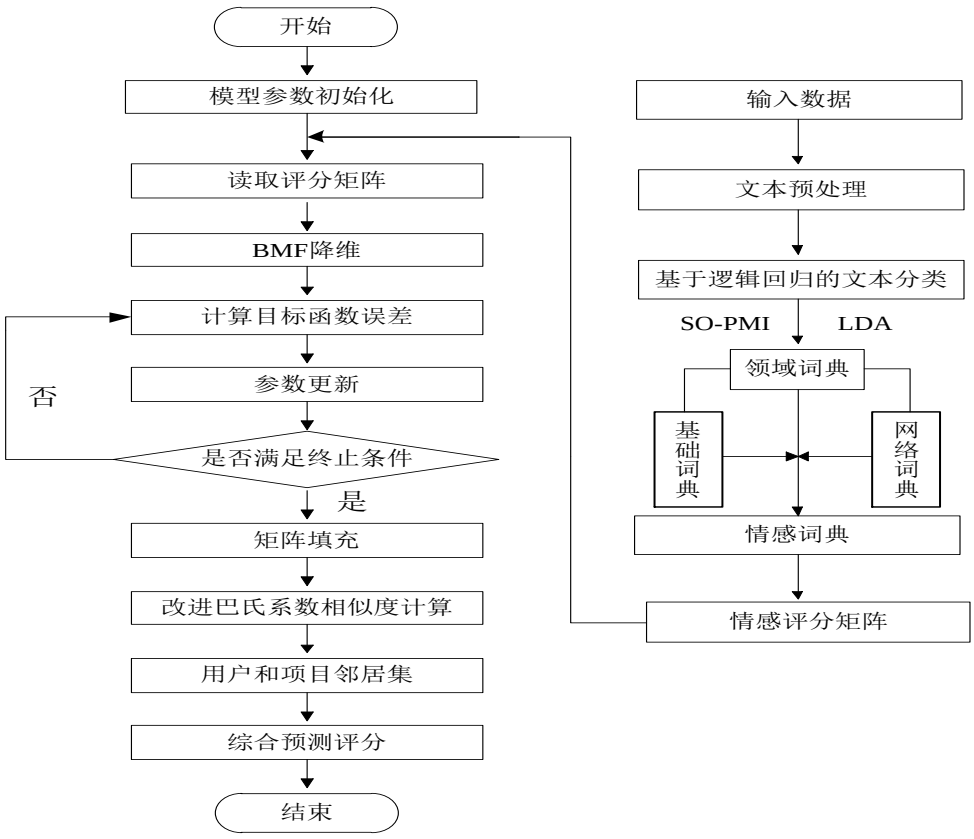


图 5.1 模型构建的基本流程

5.2 模型训练及检验

对在线评论中用户的偏好深入挖掘与分析，进行针对性的个性化推荐，不仅方便消费者迅速、便捷地找到自己需要的农产品，而且提升了平台的服务质量。因此，本文将情感分析与推荐方法结合，构建在线评论情感分析的个性化推荐模型，以甘肃省特色农产品为研究对象，基于其在电商平台的产品评论进行情感分析，挖掘用户的兴趣偏好，结合混合推荐算法为用户进行农产品的推荐。

（1）数据来源

本文所使用的文本数据均来自“京东-农资频道”网站[①]的评论数据，通过网络爬虫软件--八爪鱼采集器，对甘肃省的扶贫馆、特色馆等前 20 个排名较高的店铺产品进行爬取，共采集到“兰州百合”、“天水苹果”、“苦水玫瑰”等 36 种甘肃省具有地方特色的农产品，按照用户名、评分、评论文本、类别、购买日期等信息对好、中和差三种类型的评论文本进行分别爬取并进行汇总，共采

[①] 数据来源：https://nong.jd.com/

集到 15284 条评论。最终经预处理，得到 827 个用户对 36 种农产品的 12743 条评分记录，评分范围在 1-5 之间。因用户评论会受到产品品质、客服态度、物流速度等多种因素影响，所以不同店铺的同种产品评分会有所差异。因此，本文对不同店铺的农产品分别进行推荐，即对共 230 种农产品进行推荐。

该数据集的数据稀疏度 Φ 为：

$$\Phi = 1 - \frac{12743}{827 \times 230} = 93.3\%$$

在实验中，评分矩阵中的数据训练集和测试集的比例为 8:2，并将构建的基于情感分析的农产品个性化推荐模型与其他三种模型对比，验证模型的可行性。

（2）模型的参数设置

经模型的多次迭代训练，最终确定的主要参数如表 5.1 所示。

表 5.1 模型的主要参数

主要参数	参数值
学习率 γ	0.02
正则参数 λ	0.01
迭代次数 iteration	50
权重系数 α	0.4

（3）推荐结果与分析

本文构建了基于在线评论情感分析的农产品个性化推荐模型，将情感分析与构建的混合推荐算法（BMF-HCF）结合，对特色农产品进行推荐。为了验证模型的有效性，设置了三组实验，分别是：①衡量指标为 MAE 时，该模型与 UCF、ICF、HCF 三种模型的对比；②衡量指标为 RMSE 时，该模型与 UCF、ICF、HCF 三种模型的对比；③基于情感分析的 BMF-HCF 与 BMF-HCF 的 MAE、RMSE 对比。

1）情感评分矩阵下 BCF-HCF 模型与其他三种模型 MAE 指标对比

在本实验中，将 BCF-HCF 模型与其他三种模型 UCF、ICF、HCF 进行对比。设置邻居集数目 K 的大小分别为 5、10、20、30、40、60 时，以上四种模型 MAE 值的变化。结果如表 5.2 与图 5.2 所示。

表 5.2 四种模型的 MAE 值

K	UCF	ICF	HCF	BCF-HCF
5	0.901	0.913	0.852	0.843
10	0.897	0.892	0.848	0.834
20	0.890	0.874	0.832	0.811
30	0.883	0.872	0.815	0.802
40	0.877	0.862	0.803	0.798
60	0.875	0.860	0.800	0.795

表 5.2 中的第一列表示 K 取的不同邻居数目，取值范围为 5-60，第二至五列分别是 UCF、ICF、HCF 和 BCF-HCF 四种模型随着 K 的不同变化的 MAE 值的大小。从表 5.2 可以看出，BCF-HCF 模型比其他三种模型有较小的预测误差，如图 5.2 所示。

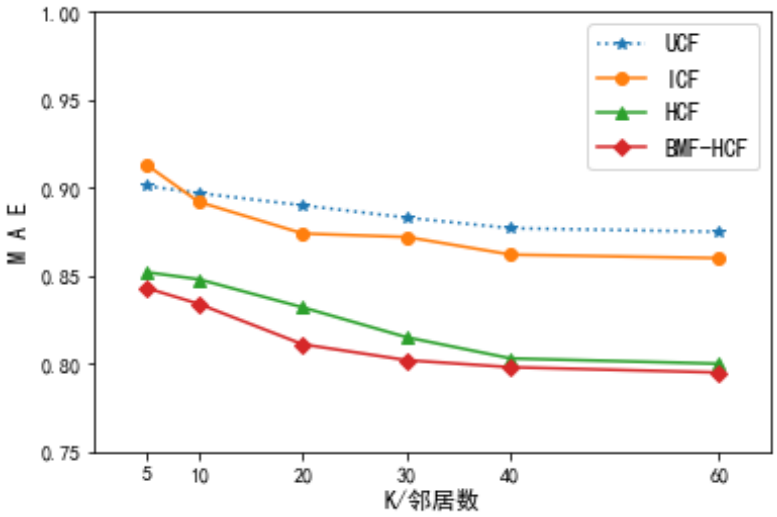


图 5.2 基于情感分析的推荐模型 MAE 对比

由图5.2可以发现，四种模型随着邻居数K的增大呈现出不同的下降趋势，并且邻居数在 5、10、20 范围内时下降幅度最大，在 K=30 时逐渐趋于平稳。可见，本文提出的 BCF-HCF 模型的预测误差明显小于其他三种模型，验证了模型的有效性。

2）情感评分矩阵下 BCF-HCF 模型与其他三种模型 RMSE 指标对比

同样，本实验将 BCF-HCF 模型与 UCF、ICF、HCF 三种模型进行对比。设置邻居集数目 K 的大小分别为 5、10、20、30、40、60 时，以上四种模型 RMSE 值的变化。结果如表 5.3 与图 5.3 所示。

表 5.3 四种模型的 RMSE 值

K	UCF	ICF	HCF	BCF-HCF
5	1.272	1.284	1.129	1.121
10	1.270	1.239	1.125	1.108
20	1.251	1.221	1.107	1.074
30	1.244	1.217	1.094	1.059
40	1.238	1.207	1.067	1.053
60	1.236	1.204	1.065	1.050

表 5.3 中的第一列表示 K 取的不同邻居数目，取值范围为 5-60，第二至五列分别是 UCF、ICF、HCF 和 BCF-HCF 四种模型随着 K 的不同变化的 RMSE 值的大小。从表 5.3 可以看出，BCF-HCF 模型在预测准确率上有一定程度的提升，如图 5.3 所示。

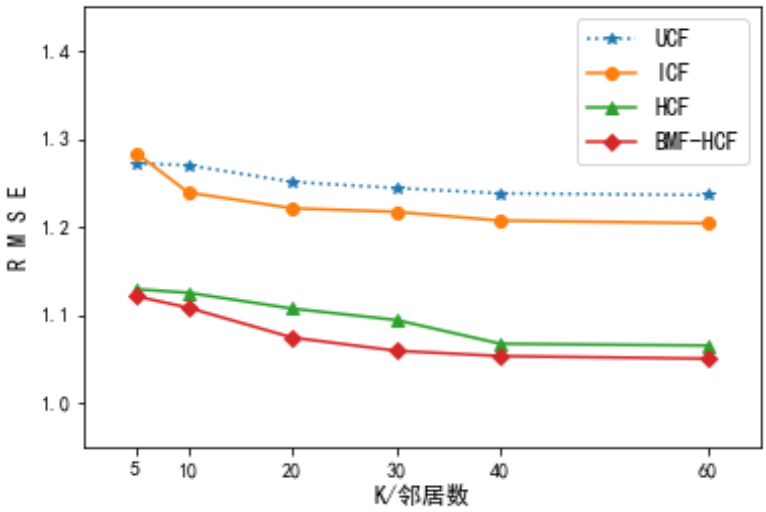


图 5.3 基于情感分析的推荐模型 RMSE 对比

由图所示，BCF-HCF 模型的 RMSE 曲线明显低于其他三种模型，并且邻居数在 K=30 时逐渐趋于平稳。可见，本文提出的 BCF-HCF 模型预测误差更小，说明推荐结果较其他模型更准确。

3）基于情感分析的 BMF-HCF 与 BMF-HCF 的 MAE、RMSE 指标对比

本实验是通过基于情感分析的 BMF-HCF 与 BMF-HCF 在 MAE、RMSE 指标下的对比，以此检验推荐的预测误差。同样，设置邻居集数目 K 的大小分别为 5、10、20、30、40、60 时，基于情感分析 BMF-HCF 与 BMF-HCF 的 MAE 与 RMSE 值的变化。结果如表 5.4 与图 5.4、图 5.5 所示。

表 5.4 基于情感分析 BMF-HCF 与 BMF-HCF 的 MAE 与 RMSE 值

K	BMF-HCF		基于情感分析的 BMF-HCF	
	MAE	RMSE	MAE	RMSE
5	0.872	1.152	0.843	1.121
10	0.859	1.149	0.834	1.108
20	0.846	1.136	0.811	1.074
30	0.824	1.124	0.802	1.059
40	0.813	1.113	0.798	1.053
60	0.807	1.106	0.795	1.050

表 5.4 中的第一列表示 K 取的不同邻居数目，取值范围为 5-60，第二至三列分别是 BMF-HCF 与基于情感分析的 BMF-HCF 随 K 的不同变化的 MAE 与 RMSE 值的大小。从表 5.4 可以看出，基于情感分析的 BMF-HCF 进行推荐的效果比 BMF-HCF 有很明显的提升。为了便于更直观的观测实验结果，其 MAE 与 RMSE 的对比如图 5.4 与图 5.5 所示。

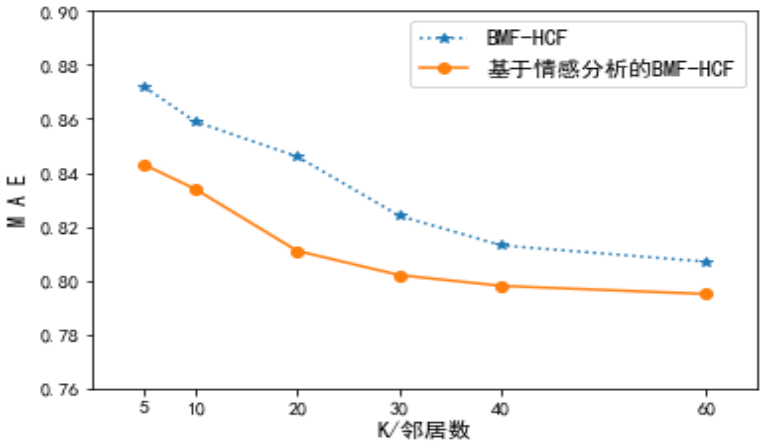


图 5.4 基于情感分析 BMF-HCF 与 BMF-HCF 的 MAE 对比

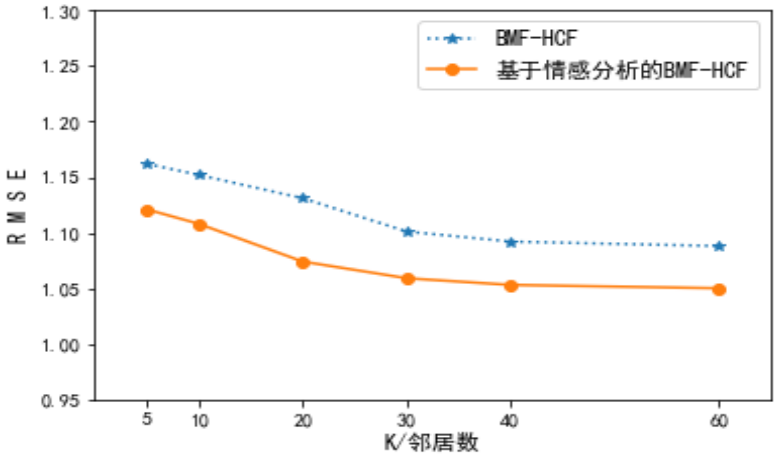


图 5.5 基于情感分析 BMF-HCF 与 BMF-HCF 的 RMSE 对比

5.3 本章小结

本章首先介绍了基于情感分析的农产品个性化推荐模型的构建思想与构建过程；然后在利用情感词典计算情感得分得到的情感评分矩阵基础上，与混合推荐算法 BMF-HCF 融合，构建基于情感分析的个性化推荐模型，并将该模型与 UCF、ICF、HCF 三种推荐模型作对比，进行推荐性能的检验。结果显示，本章构建的基于情感分析的 BMF-HCF 推荐模型在 MAE、RMSE 指标上较其他推荐模型具有明显优势。

6 基于情感分析的农产品个性化推荐系统设计

为了将第五章的农产品个性化推荐应用到实际推荐中，本文设计了基于在线评论情感分析的农产品个性化推荐系统，主要是将基于在线评论的情感分析、基于 BMF 矩阵分解算法、基于改进的巴氏系数相似性度量方法以及混合协同过滤算法进行融合，并运用在推荐系统中，从而为用户提供个性化的农产品推荐。

6.1 农产品推荐系统信息获取

本文设计农产品个性化推荐系统的主要任务是获得和分析用户对农产品的评论信息，根据评论信息运用基于在线评论情感分析的农产品个性化推荐模型，为用户提供其可能喜欢的农产品。系统实现个性化推荐必须获得用户对农产品的评论信息，主要包括：

（1）用户 ID：系统需要为每一个用户创建不同的 ID，以便根据历史评分数据准确识别用户。

（2）农产品 ID：每一种农产品在数据库中均拥有一个不同的 ID，以便区分这些农产品。

（3）用户评论文本：用户的评论文本主要包括农产品名称、农产品的主要特征以及用户的情感词等。

（4）用户评论时间：用户的评论时间与评论文本相对应，主要是统计该用户购买农产品的交易记录。

（5）网络特征：网络特征主要包括用户登录时使用的工具、登录时间、浏览时间、登录时的 IP 地址等，这些特征可以在 Web 日志中得到。

6.2 农产品推荐系统需求

推荐系统的任务是根据用户的评分数据，分析用户的兴趣偏好，为其推荐有可能感兴趣的农产品。该系统的需求主要包括两部分，即消费者的需求与商家的需求。

（1）消费者的需求

消费者对推荐系统的需求包括两点：第一，推荐结果应符合消费者的兴趣偏好；第二，应针对消费者的兴趣偏好，为其推荐有可能感兴趣但没有购买过的农产品。

（2）商家的需求

推荐系统的应用不仅要能维持原有的用户，而且可以拓展新的用户，达到增强用户粘性的目的。这就要求所要应用的推荐系统，一方面，针对原有的用户，深入挖掘与分析其兴趣偏好，主动为其推荐极有可能感兴趣的农产品；另一方面，根据新用户的历史评论，主动为其推荐可能感兴趣的农产品。

6.3 农产品推荐系统框架

本文设计农产品个性化推荐系统主要分为五个模块，即数据收集模块、数据预处理模块、评论特征抽取模块、情感分析模块、农产品推荐模块。其总体框架如图 6.1 所示。

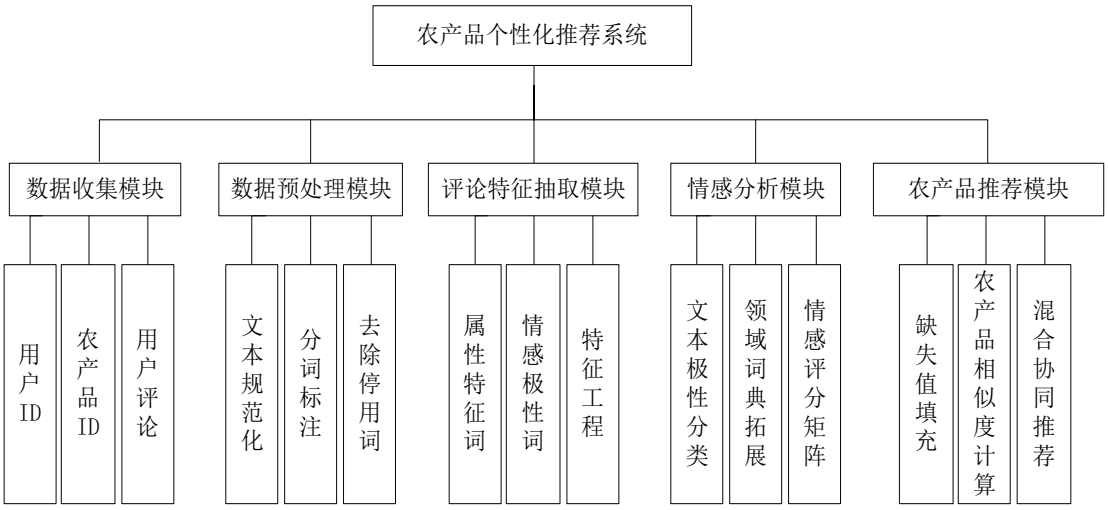


图 6.1 系统总体框架

（1）数据收集模块

数据收集模块主要通过农产品电商平台用户的基础数据、农产品数据收集该用户对所购买产品的评论，并将数据存储在 Oracle 数据库中。数据库存储的数据有用户 ID、农产品 ID、用户评论、用户的兴趣模型以及产生的推荐列表。其中，用户 ID、农产品 ID、用户评论数据是整个系统的原始数据，用户的兴趣模型与推荐列表是推荐系统经运行产生的数据。其中，农产品数据存储在 Agriculture_Products_Info 里，其属性如表 6.1 所示。

表 6.1 Agriculture_Products_Info 表的属性

Dkey	Products_ID	Name	Price	Category
主键	农产品 ID	农产品名称	农产品价格	农产品类别

用户数据存储在 User_Info 里，其属性如表 6.2 所示。

表 6.2 User_Info 表的属性

User_Dkey	User_ID	Name	Created_time	Trade_time
主键	用户 ID	用户名称	创建时间	交易时间

用户评论数据存储在 User_Comments_Info 里，其属性如表 6.3 所示。

表 6.3 User_Comments_Info 表的属性

Comments_Dkey	User_ID	Products_ID	Comments	Comments_time
主键	用户 ID	农产品 ID	用户评论	评论时间

用户兴趣模型数据存储在 User_ Preference_Info 里，其属性如表 6.4 所示。

表 6.4 User_ Preference_Info 表的属性

Preference_Dkey	User_ID	Name	User_Preference	Create_time
主键	用户 ID	用户名称	用户兴趣模型	创建时间

用户推荐列表数据存储在 Products_Recommend 里，其属性如表 6.5 所示。

表 6.5 Products_Recommend 表的属性

Recommend_Dkey	User_ID	Products_ID	Predicted_Score
主键	用户 ID	农产品 ID	预测评分

（2）数据预处理模块

数据预处理模块主要通过对收集到的用户评论数据进行文本规范化、分词标注以及去除停用词后得到结构化的农产品评论集。

（3）评论特征抽取模块

评论特征抽取模块主要通过 LDA 主题模型挖掘出特色农产品的属性特征词与情感极性词，并利用 word2vec 词向量模型进行特征向量化处理。

（4）情感分析模块

情感分析模块，首先是对评论文本进行情感极性分类；然后进行农产品领域情感词典拓展；最后，基于情感词典计算农产品各特征的情感评分，得到情感评分矩阵。

（5）农产品推荐模块

农产品推荐模块，首先是利用 BMF 算法对情感评分矩阵进行缺失值填充；然后使用改进的巴氏系数计算特色农产品之间的相似性；最后，基于混合的协

同过滤算法预测未评分农产品的分数，为用户推荐预测分数高的农产品。

6.4 农产品推荐系统处理流程

农产品推荐系统包含五个主要模块，各模块相互协同，完成推荐任务。该系统的工作流程为：数据收集模块负责收集用户数据、农产品数据以及评论数据，为推荐系统提供数据支撑；数据预处理模块是对收集到的数据进行文本规范化、分词标注以及去除停用词等结构化处理，以备使用；评论特征抽取模块负责对用户评论中的农产品特征、情感词进行抽取，同时进行文本向量化处理；情感分析模块是对评论文本进行情感极性分类、农产品领域词典拓展，并基于词典进行情感分析，得到情感评分矩阵；农产品推荐模块是根据从情感分析模块中得出的情感评分矩阵，对其进行模型训练，使用 BMF 算法、改进的巴氏系数相似性度量方法以及混合的协同过滤算法对目标用户未评分的农产品进行预测，将预测分数高的农产品推荐给用户。本文设计的农产品个性化推荐系统的基本流程如图 6.2 所示。

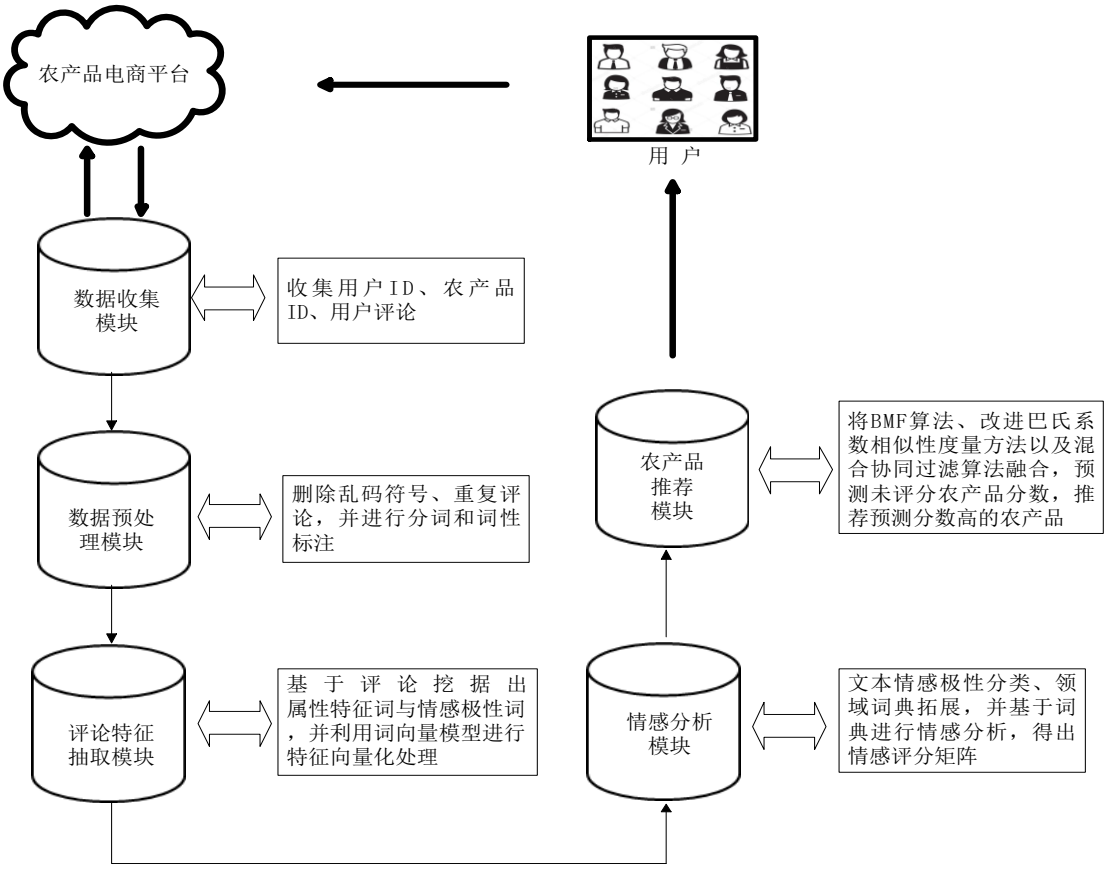


图 6.2 农产品个性化推荐系统基本流程

6.5 本章小结

本章主要设计了一个基于在线评论情感分析的农产品个性化推荐系统，阐述了系统对用户个性化信息的获取与系统需求，并对系统的总体框架以及各模块的功能与处理流程进行详细介绍。

7 总结与展望

7.1 总结

本文首先对评论文本进行预处理，在现有中文情感词典的基础上，对农产品领域的情感词典进行扩充，通过基于词典的情感分析计算得到用户的情感评分矩阵；其次，构建了 BMF-HCF 推荐算法，并与基于用户的协同过滤、基于项目的协同过滤以及混合协同过滤算法的对比，验证算法的合理性；再次，将情感分析与 BMF-HCF 推荐算法结合，构建基于在线评论情感分析的农产品个性化推荐模型，并通过与其他三种推荐模型的对比，验证模型的有效性。最后，设计了基于在线评论情感分析的农产品个性化推荐系统，为用户推荐符合其需求的农产品。

本文的主要研究成果如下：

（1）农产品领域情感词典的拓展。本文在现有情感词典的基础上，对农产品在线评论文本利用情感倾向点互信息算法和 LDA 主题模型进行领域词典扩充，并加入了近 3 年最新的网络热词完善情感词典。在此基础上，计算出用户评论的情感值，从而得出情感评分矩阵。

（2）构建了融合矩阵分解和改进巴氏系数的混合推荐算法。针对巴氏系数度量方法的不足，通过加入调和平均权值因子和用户偏好对巴氏系数相似性度量方法进行改进，寻找更相似的用户和项目。在此基础上，对基于用户和基于项目的两种协同过滤推荐算法进行融合，得到混合协同推荐算法，并利用 BMF 算法对评分矩阵进行缺失值填充，建立融合矩阵分解和改进巴氏系数的混合推荐算法。在数据集上验证了算法的有效性。

（3）构建了基于情感分析的农产品个性化推荐模型。通过词典的情感分析得到情感评分矩阵，并融合混合推荐算法，建立基于情感分析的农产品个性化推荐模型，通过与其他三种推荐模型的对比，验证模型的可行性。

（4）设计了农产品个性化推荐系统。将基于情感分析的个性化推荐模型应用在推荐系统中，从而为用户提供个性化的农产品推荐。

7.2 展望

本文提出的基于在线评论情感分析的农产品个性化推荐模型，虽然较其他

传统推荐算法预测准确度有一定提升，但是仍存在不足之处。因此，针对本文研究不足与局限之处，未来在以下几个方面展开进一步研究：

（1）电子商务和社会媒体的迅猛发展，评论文本随着网络热词的不断更新也在不停的发展与变化。情感词典作为量化用户评论文本情感值的方法，在情感分析中起到至关重要的作用。因此，保证情感词典的不断更新与完善，对情感分析准确度的提高是非常有必要的。

（2）农产品是日常生活的必需品，人们在电商平台购买农产品时会受很多因素的影响。本文只选取了用户的历史评论数据进行研究，但其实还有很多诸如自然环境变化、国家政策的调控等影响因素有待挖掘，进行综合分析，如何考虑将更多的影响因素加入到产品推荐中来，提高推荐效果是以后有待进一步研究的问题。

参考文献

[1] Liu B. Sentiment analysis and opinion mining [J]. Synthesis Lecture on Human Language Technologies, 2012, 5(1): 1-167.

[2] Sanjiv Das and Mike Chen Yahoo! for Amazon: Extracting market sentiment from stock message boards [C]. In Proceedings of the Asia Pacific Fianace Association Annual Conference(APFA), 2001.

[3] Tong. R. M. An operational system for detecting and tracking opinions in online discussion [C]. The 24th Annual International ACM SIGIR Conference. New Orleans: ACM, 2001.

[4] Thet T T, Na J c, Khoo C S G. Aspect-based sentiment analysis of movie reviews on discussion boards [J]. Journal of Information Science, 2010:33-37.

[5] Abbasi A, France S, Zhang Z, et al. Selecting attributes for sentiment classification using feature relation networks [J]. Knowledge and Data Engineering, IEEE Transactions, 2011, 23(3): 447-462.

[6] Wang H, Yin P, Yao J, et al. Text feature selection for sentiment classification of Chinese online reviews [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2013, 25(4): 425-439.

[7] Wang H, Yin P, Zheng L, et al. Sentiment classification of online reviews: using sentence-based language model [J]. Journal of Experimental & Theoretical Artificial Intelligence, 2014, 26(1): 13-31.

[8] Turney P. Thumbs up or thumbs down: Semantic orientation applied to unsupervised classification of reviews [C]. Meeting on Association for Computational Linguistics: 2002: 417-424.

[9] A. Muhammad. Contextual sentiment analysis for social media genres [J]. Knowledge-Based Systems, 2016, 108: 92-101.

[10] Li Ji, Lowe Dan, Wayment Luke, et al. Text mining datasets of β-hydroxybutyrate (BHB) supplement products' consumer online reviews [J]. Data in brief, 2020, 23-30.

[11] Chih-Fong Tsai, Kuanchin Chen, Ya-Han Hu, et al. Improving text

summarization of online hotel reviews with review helpfulness and sentiment [J]. Tourism Management, 2020, 72-80.

[12] Onur Can Sert, Salih Doruk Şahin, Tansel Özyer, et al. Analysis and prediction in sparse and high dimensional text data: The case of Dow Jones stock market [J]. Physica A: Statistical Mechanics and its Applications, 2020, 539-545.

[13] Sarah Meldrum, Sherlock A. Licorish, Caitlin A. Owen, et al. Understanding stack overflow code quality: A recommendation of caution [J]. Science of Computer Programming, 2020, 199.

[14] Zafar Ali, Guilin Qi, Khan Muhammad, et al. Paper recommendation based on heterogeneous network embedding [J]. Knowledge-Based Systems, 2020, 210.

[15] Linden G, Smith B, York J. Amazon.com recommendations: Item-to-item collaborative filtering [J]. IEEE Internet computing, 2003, 7(1): 76-80.

[16] Kenny A. Rodriguez-Wallberg, Ida Wikander. A global recommendation for restrictive provision of fertility treatments during the COVID-19 pandemic [J]. Acta Obstetricia et Gynecologica Scandinavica, 2020, 99(5).

[17] Zheng Zeqi, Gao Yuandong, Yin Likang, et al. Modeling and analysis of a stock-based collaborative filtering algorithm for the Chinese stock market [J]. Expert Systems with Applications, 2019(prepublish).

[18] Natarajan Senthilselvan, Vairavasundaram Subramaniyaswamy, Natarajan Sivaramakrishnan, et al. Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data [J]. Expert Systems With Applications, 2019, 149.

[19] Celma O, Serra X. FOAFing the music: Bridging the semantic gap in music recommendation [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(4): 250-256.

[20] Sun Zhongbin, Zhang Jingqi, Sun Heli, et al. Collaborative filtering based recommendation of sampling methods for software defect prediction [J]. Applied Soft Computing Journal, 2019.

[21] Shardaand U, Maes P. Social information filtering: algorithms for automating "word of mouth" [C]. Proceedings of the SIGCHI conference on Human factors

in computing systems. ACM Press: Addison-Wesley Publishing Co., 1995: 210-217.

[22] Guang Xing Lye, Wai Khuen Cheng, Teik Boon Tan, et al. Creating Personalized Recommendations in a Smart Community by Performing User Trajectory Analysis through Social Internet of Things Deployment [J]. Sensors, 2020, 20(7).

[23] Yijia Zhang, Zhenkun Shi, Wanli Zuo, et al. Joint Personalized Markov Chains with social network embedding for cold-start recommendation [J]. Neurocomputing, 2020, 376-386.

[24] Xueping Su, Meng Gao, Jie Ren, et al. Personalized Clothing Recommendation Based on User Emotional Analysis. [J]. Discrete Dynamics in Nature and Society, 2020, 199-202.

[25] 周咏梅, 杨佳能, 阳爱民. 面向文本情感分析的中文情感词典构建方法 [J], 山东大学学报(工学版), 2013,43(6): 27-33.

[26] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报, 2006, 26 (11): 14-20.

[27] 娄德成, 姚天昉. 汉语句子语义极性分析和观点抽取方法的研究 [J]. 计算机应用, 2006, 26(11): 2622-2625.

[28] 龚安, 费凡. 基于多特征融合的评论文本情感分析 [J]. 计算机技术与发展, 2018(08): 91-95.

[29] 赵刚, 徐赞. 基于机器学习的商品评论情感分析模型研究 [J]. 信息安全研究, 2017, 3(02): 166-170.

[30] 赵志滨, 刘欢, 姚兰, 等. 中文产品评论的维度挖掘及情感分析技术研究 [J]. 计算机科学与探索, 2018, 12(03): 341-349.

[31] 曾子明, 杨倩雯. 基于 LDA 和 AdaBoost 多特征组合的微博情感分析 [J]. 数据分析与知识发现, 2018, 2(08): 51-59.

[32] 吴鹏, 应杨, 沈思. 基于双向长短期记忆模型的网民负面情感分类研究 [J]. 情报学报, 2018, 37(08): 845-853.

[33] 冯兴杰, 张志伟, 史金钏. 基于卷积神经网络和注意力模型的文本情感分

析 [J]. 计算机应用研究, 2018, 35(05): 1434-1436.

[34] 刘思琴, 冯胥睿瑞. 基于 BERT 的文本情感分析 [J]. 信息安全研究, 2020, 6(03): 220-227.

[35] 胡德敏, 褚成伟, 胡晨, 等. 预训练模型下融合注意力机制的多语言文本情感分析方法 [J]. 小型微型计算机系统, 2020, 41(02): 278-284.

[36] 安璐, 吴林. 融合主题与情感特征的突发事件微博舆情演化分析 [J]. 图书情报工作, 2017, 61(15): 120-129.

[37] 崔彦琛, 张鹏, 兰月新, 等. 面向时间序列的微博突发事件衍生舆情情感分析研究——以“6.22”杭州保姆纵火案衍生舆情事件为例 [J]. 情报科学, 2019, 37(03): 119-126.

[38] 何天翔, 张晖, 李波, 等. 一种基于情感分析的网络舆情演化分析方法 [J]. 软件导刊, 2015, 14(05): 131-134.

[39] 李宏媛, 陶然. 服装电商评论情感分析研究 [J]. 智能计算机与应用, 2017, 7(01): 27-30+34.

[40] 孟园, 王洪伟, 王伟. 网络口碑对产品销量的影响:基于细粒度的情感分析方法 [J]. 管理评论, 2017, 29(01): 144-154.

[41] 杨春晓, 张鹤馨, 黄家雯, 等. 卷烟在线评论的文本情感分析 [J/OL]. 中国烟草学报: 1-11[2020-04-26]. http: //kns. cnki.net/ kcms/ detail/ 11. 2985. ts.20200121.1622.018. html.

[42] 王梓萌, 周亦鹏, 苏兵杰. 基于用户评论下的生鲜农产品优选排序 [J]. 江苏农业科学, 2020, 48(03): 305-310.

[43] 梁霞, 姜艳萍, 高梦. 基于在线评论的产品选择方法[J].东北大学学报(自然科学版), 2017, 38(01): 143-147.

[44] 涂海丽, 唐晓波. 基于在线评论的游客情感分析模型构建 [J]. 现代情报, 2016, 36(04): 70-77.

[45] 骆亮. 基于内容推荐算法和余弦相似度算法的领导决策辅助信息系统 [J]. 广西科学院学报, 2018, 34(02): 143-150.

[46] 崔春生, 王梦冉, 王国成. 一种基于可拓学的电子商务内容推荐算法研究 [J]. 运筹与管理, 2018, 27(06): 75-81.

[47] 高晟. 基于关联规则与贝叶斯网络的高校图书馆个性化图书推荐服务 [J]. 情报探索, 2019(08): 87-94.

[48] 陈双双, 王晓军. 基于关联规则的标签推荐 [J]. 计算机技术与发展, 2018, 28(12): 43-47.

[49] 袁泉, 成振华, 江洋. 基于知识图谱和协同过滤的电影推荐算法研究 [J]. 计算机工程与科学, 2020, 42(04): 714-721.

[50] 李昆仑, 戎静月, 苏华仃. 一种改进的协同过滤推荐算法 [J]. 河北大学学报(自然科学版), 2020, 40(01): 77-86.

[51] 崔国琪, 李林. 加入惩罚因子的电商平台协同过滤推荐算法 [J]. 软件导刊, 2020, 19(01): 103-107.

[52] 郑修猛, 陈福才, 黄瑞阳, 等. 面向协同推荐的评论文本情感打分机制研究 [J]. 信息工程大学学报, 2017, 18(04): 464-469.

[53] 卢竹兵, 李玉州. 基于网络评论情感信任分析的推荐策略 [J]. 计算机科学, 2019, 46(06): 75-79.

[54] 钱春琳, 张兴芳, 孙丽华. 基于在线评论情感分析的改进协同过滤推荐模型 [J]. 山东大学学报(工学版), 2019, 49(01): 47-54.

[55] 彭敏, 席俊杰, 代心媛, 等. 基于情感分析和 LDA 主题模型的协同过滤推荐算法 [J]. 中文信息学报, 2017, 31(02): 194-203.

[56] 熊旭东, 杜圣东, 夏琬钧, 等. 基于二分图卷积表示的推荐算法 [J/OL]. 计算机科学: 1-11[2021-02-19]. http: // kns. cnki. net/ kcms/ detail/ 50.1075. TP. 20210112.1126.044.html.

[57] 杨丹, 张鹰. 基于评论文本的动态生成对抗网络推荐算法 [J]. 太原师范学院学报(自然科学版), 2020, 19(04): 40-44.

[58] 何婧, 胡杰. 融合矩阵分解和 XGBoost 的个性化推荐算法 [J]. 重庆大学学报, 2021, 44(01): 78-87.

[59] 杨兴雨, 李华平, 张宇波. 基于聚类和随机森林的协同过滤推荐算法 [J]. 计算机工程与应用, 2018, 54(16): 152-157.

[60] 沈晶磊, 虞慧群, 范贵生, 等. 基于随机森林算法的推荐系统的设计与实现 [J]. 计算机科学, 2017, 44(11): 164-167+186.

[61] 陆君之. 基于随机森林回归算法的电影评分预测模型 [J]. 江苏通信, 2018, 34(01): 75-78.

[62] 滕传志, 赵月旭. 基于随机森林-马尔可夫用户冷启动推荐系统 [J]. 计算机工程与设计, 2020, 41(11): 3094-3098.

[63] 叶霞, 曹军博, 许飞翔, 等. 中文领域情感词典自适应学习方法 [J]. 计算机工程与设计, 2020, 41(08): 2231-2237.

[64] 吴璠, 王中卿, 周夏冰, 等. 基于用户和产品表示的情感分析和评论质量检测联合模型 [J]. 软件学报, 2020, 31(08): 2492-2507.

[65] 李卓冉. 逻辑回归方法原理与应用 [J]. 中国战略新兴产业, 2017(28): 114-115.

[66] 彭云, 万常选, 江腾蛟, 等. 基于语义约束 LDA 的商品特征和情感词提取 [J]. 软件学报, 2017, 28(03): 676-693.

[67] 樊艳清, 梁宏宇, 纪佳琪. 协同过滤算法中相似度计算问题研究 [J]. 计算机技术与发展, 2020, 30(08): 91-96.

[68] Patra B K, Launonen R, Ollikainen V, et al. A new similarity measure using Bhattacharyya coefficient for collaborative filtering in sparse data [J]. Knowledge-Based Systems, 2015, 82: 163-177.

[69] 项亮, 陈义, 王益. 推荐系统实践 [M]. 人民邮电出版社, 2020, 19-23.

[70] 姜伶伶, 何中市, 张航. 基于 Good-Turing 平滑 SO-PMI 算法构建微博情感词典方法的研究 [J]. 现代计算机(专业版), 2018(10): 15-20.

[71] 张艳丰, 李贺, 彭丽徽. 基于模糊情感计算的商品在线评论用户品牌转换意向研究 [J]. 现代图书情报技术, 2016(05): 64-71.

[72] 王鹏, 高铖, 陈晓美. 基于 LDA 模型的文本聚类研究 [J]. 情报科学, 2015, 33(01): 63-68.

[73] 李国, 张春杰, 张志远. 一种基于加权 LDA 模型的文本聚类方法 [J]. 中国民航大学学报, 2016, 34(02): 46-51.

致 谢

转眼间，三年的研究生学习生活临近尾声。在兰州财经大学三年的学习与生活当中，自己渐渐成长，让我成为更好的自己。在此，我要对我的导师王玉珍教授、信息工程学院的领导和老师以及同学们表达我最诚挚的感谢。在未来的学习、工作中，我会牢记“博修商道”的校训，并以此严格要求自己。

首先，我要感谢我的导师王玉珍教授。在两年多的时间里，王老师是我研究生求学道路上的引路人，不仅在学业上给我耐心指导，而且在生活中给我无微不至的关怀。可以说，王老师渊博的专业知识、严谨的学术态度、和蔼可亲的处事风格，让我受益匪浅。当自己在论文撰写遇到困难时，王老师总会从论文选题、文章结构、实验实施等方面及时给我提出中肯的建议，如果没有王老师的悉心指导与帮助，论文一定不会如期完成。在此向王老师表示深深的感谢。

其次，感谢信息工程学院的各位领导和老师，为我们提供良好的学习环境以及在理论知识上地传授与分享。同时，感谢我的同学们，在学习和生活上给我的鼓励与帮助。

再次，还要感谢我的父母，在学习和生活上给我的支持与鼓励，使我能够全身心投入到学习中，顺利完成学业。

最后，诚挚的感谢参加论文评阅的各位老师！

攻读硕士学位期间发表的论文及科研情况

[1] 李佳儒, 王玉珍. 新零售背景下生鲜食品超市配送路径优化研究 [J]. 邵阳学院学报(自然科学版), 2019, 16(03): 27-35.

[2] Yuzhen Wang, Jiaru Li. Text Analysis of Cross-border E-commerce Policy Based on Co-word Clustering Method: A Case Study of Gansu Province [C]. The 5th Intelligent Computing and Signal Processing Conference. Suzhou: ICSP, 2020.

[3] 李佳儒, 王玉珍, 丁申宇. 基于逻辑回归的在线评论情感分类方法研究 [J]. 东莞理工学院学报, 2020, 27(05): 50-54.

[4] 李佳儒, 王玉珍, 丁申宇. 在线评论情感分析的影院推荐 [J]. 宁德师范学院学报(自然科学版), 2020, 32(03): 253-258.

[5] 参与甘肃省科技厅项目：甘肃省电子商务信用管理研究——构建基于大数据的甘肃省网络供应商信用评估体系；项目编号：17CX1ZA024.

[6] 参编中国铁道出版社出版的《电子商务概论》，参编共计 10.6 万字.

[7] 参编清华大学出版社出版的《电子商务概论》，参编共计 10.1 万字.

[8] 参加 2018 年第六届“发现杯”全国大学生互联网软件设计大奖赛，全国总决赛三等奖；证书编号：2019049049550.