分类号_____　　密级 _____
U D C _____　　编号 10741

硕 士 学 位 论 文

（专业学位）

论文题目 基于某电商平台用户行为的个性化推荐

研 究 生 姓 名：王娜

指导教师姓名、职称：庞智强 教授

学 科、专 业 名 称：统计学 应用统计硕士

研 究 方 向：大数据分析

提 交 日 期：2021 年 6 月 6 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：王娜 签字日期：2021年6月6日

导师签名：张玉宏 签字日期：2021年6月6日

导师(校外)签名：张小宁 签字日期：2021年6月6日

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：王娜 签字日期：2021年6月6日

导师签名：张玉宏 签字日期：2021年6月6日

导师(校外)签名：张小宁 签字日期：2021年6月6日

Personalized recommendation based on user behavior of an e-commerce platform

Candidate: Wang Na

Supervisor: Pang Zhiqiang

摘 要

当前互联网数据规模急剧扩大，我们已坐拥海量信息，可真正找到对自己有用信息的效率变得越来越低，且目前电商平台面临的一大难题就是如何快速、准确的为用户找到合适的商品，提升用户的购物体验感。而个性化推荐服务就是应对这一难题的有力工具，它不仅能为用户带来优质的服务，而且能够为商家带来前所未有的利润。

电商平台中的用户行为都是有意义的，且其蕴含着无限的价值。因此，本文对阿里平台用户行为从时间演化、行为转化、行为时间间隔和复购情况等方面进行分析，发现浏览后用户流失率高且购买转化率低，以及其他一些用户行为特征。为了能够满足不同用户的个性化需求，提升用户的购物体验感，减少用户流失提高购买转化率，为商家创造更大的价值，本文针对电商平台的个性化推荐进行了研究。

本文介绍了三种常用的推荐方法，对比分析了它们的优缺点及其适用场景，发现基于内容的推荐方法比较适用于文本类推荐领域，基于关联规则的推荐方法主要被用来发现购物车之间的关联性，基于协同过滤的方法推荐的个性化程度较高，可以挖掘用户的潜在需求，而且可解释性强。针对电商平台，为了依据用户行为数据完成用户的个性化推荐，本文最终选择了基于用户的协同过滤推荐方法。但同时也发现协同过滤推荐算法存在数据稀疏性问题、冷启动问题和扩展性等问题。对此，本文将协同过滤推荐算法与 k-means 聚类算法相结合来进行商品推荐，并将其与传统的协同过滤推荐算法做实验对比。实验结果表明：对于电商平台，基于 k-means 聚类的协同过滤推荐算法推荐的准确率、召回率和 F1 值均优于传统的协同过滤推荐算法，且其计算复杂度也较低，在缓解数据稀疏性问题的同时也有效解决了扩展性问题，不论是在推荐性能还是推荐效率上都表现出更大的优势，这为电商平台的个性化推荐服务提供了一定的参考。

关键词：用户行为 个性化推荐 k-means 聚类 协同过滤推荐

Abstract

At present, the scale of Internet data is rapidly expanding. We are already sitting on massive amounts of information. The efficiency of finding useful information for ourselves has become lower and lower. At present, a major problem facing e-commerce platforms is how to quickly and accurately obtain a large number of products. From the information, the products that the user is interested in are filtered out and presented to the user. The personalized recommendation service is a powerful tool to deal with this problem. It can not only provide users with high-quality services, but also bring unprecedented profits to businesses.

The user behavior in the e-commerce platform is meaningful, and it can even be said that every user's behavior operation reflects the essential needs of the user's heart. Therefore, this article analyzes the user behavior of Ali platform in terms of time evolution, behavior conversion, behavior time interval, and repurchase situation, and finds that the user churn rate after browsing is high and the purchase conversion rate is low, as well as some other user behavior characteristics. In order to meet the personalized needs of different users, improve the user's shopping experience, reduce user churn, increase purchase conversion rate, and create greater value for merchants, this article conducts research on personalized recommendations for e-commerce platforms.

This article introduces three commonly used recommendation methods, compares and analyzes their advantages and disadvantages and their applicable scenarios. It is found that content-based recommendation methods are more suitable for text recommendation fields, and recommendation methods based on association rules are mainly used to discover shopping carts. The relevance between collaborative filtering methods is highly personalized, and the potential needs of users can be explored, and the interpretability is strong. For the e-commerce platform, in order to complete the user's personalized recommendation based on user behavior data, this paper finally chooses the user-based collaborative filtering recommendation method. But at the same time, it is also found that the collaborative filtering recommendation algorithm has data sparseness problems, cold start problems and scalability problems. In this regard, this article combines the collaborative filtering recommendation algorithm with the k-means clustering algorithm for product recommendation, and compares it with the traditional collaborative filtering recommendation algorithm. Experimental results show that for e-commerce platforms, the accuracy, recall, and F1 value of the collaborative filtering recommendation algorithm based on k-means clustering are better than those of the traditional collaborative filtering recommendation algorithm, and its computational complexity is also lower. While alleviating the problem of data sparsity, it also effectively

solves the problem of scalability. It shows greater advantages in both recommendation performance and recommendation efficiency, which provides a certain reference for the personalized recommendation service of e-commerce platforms.

Keywords: user behavior; personalized recommendation; k-means clustering; collaborative filtering recommendation

目 录

1 引言

1.1 研究背景和意义

1.1.1 研究背景

在互联网迅猛发展的今天，智能移动设备也愈加普及，中国已步入一个信息化、数字化的社会，数据充斥着每个人的生活。近几年来，各大领域信息量持续攀升，尤其是电商行业，这离不开信息技术的高速发展和高效的智能化数据处理技术的运用。中国互联网络信息中心(CNNIC)发布的第 46 次《中国互联网络发展状况统计报告》中显示，截至 2020 年 6 月，我国网民规模达 9.40 亿，互联网普及率达 67.0%。从图 1.1 可以看出，我国网民规模平稳增长，互联网普及率也越来越高；从图 1.2 可以看出，截至 2020 年 6 月，网络购物用户规模达 7.49 亿，占网民整体的 79.7%。我国互联网产业展现出巨大的发展活力和韧性，成为我国应对新挑战、建设新经济的重要力量。

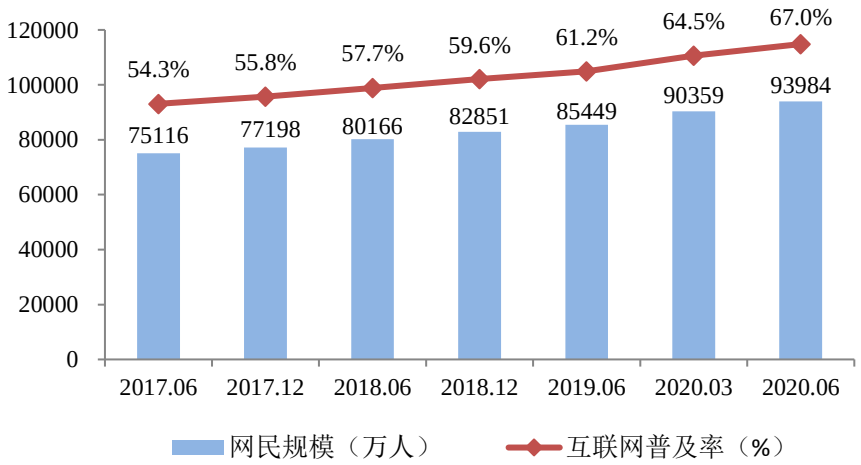

图 1.1 中国网民规模和互联网普及率

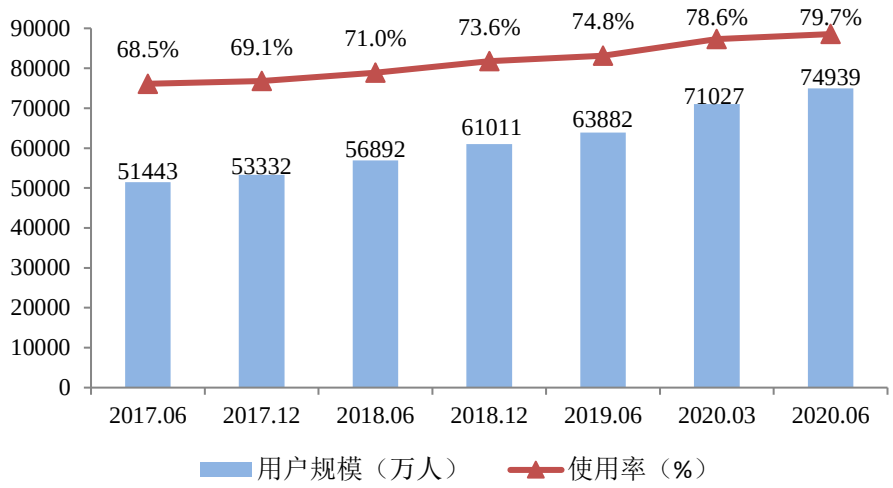

图 1.2 网络购物用户规模及使用率

互联网虽然为人们的生活带来了方便，但随之而来的“信息超载”（李勇和徐振宁，2002）问题成为了用户的一个困扰。以“天猫”“京东”为代表的网络购物平台，在 2020 年“双十一”活动总成交额分别达到 4982 亿元和 2715 亿元，反映出我国网上消费的巨大潜力，随着经济的转型和新的营销模式的转变，消费者也愈加依赖网上购物。数亿人集聚于电商平台选购自己想要的商品，可想而知其中产生的数据量非常庞大，面对庞大的数据量和不同形式的数据类型，当用户进行信息检索时肯定会产生很多冗余数据，会出现检索结果不够精确的问题，甚至根无法找到真正需要的信息。

为应对“信息超载”的问题，方便用户快速的检索自己想要的信息和产品，当前主要有两类解决办法：一类是使用搜索引擎，比如谷歌、百度、搜狐等，用户可以根据自己的需求按照关键词在特定网站进行搜索，搜索引擎通过信息匹配将检索结果反馈给用户，用户就可以得到相关的信息。但是这类方法需要用户明确清晰的表达自己的需求，若用户对自己的需求描述的不够清晰，那搜索结果的准确性也会受到影响，而且搜索引擎只能提供通用的服务，针对不同用户、不同环境，其呈现的内容并不会有所差异，这无法满足用户的个性化需求。在当今社会信息多样化、传播方式多样化的前提下，搜索引擎就不再能够很好的应对“信息超载”问题。另一类是使用推荐系统，比如淘宝、京东、亚马逊等都在使用推荐系统，在这些平台上用户不需要十分明确的表达出自己的需求，它会依据用户日志数据，通过构建模型来挖掘用户的兴趣偏好预测用户可能喜欢的商品，并将

预测结果反馈给对应的用户，个性化推荐不仅能够帮助用户快速发现其可能感兴趣的商品或信息，而且能够挖掘用户的潜在需求，帮助用户筛选信息做出决断，从而实现用户的个性化推荐。并且随着电商企业的增多，各大电商平台都在不断优化自己的推荐系统希望能够提供更好的个性化服务，推荐系统也成为各大平台一大竞争点。

随着个性化推荐系统的不断创新与发展，其在各大领域得到了广泛的应用，比如购物时的商品推荐、外出时的路线推荐、餐饮中的菜品推荐、学习中的图书推荐、娱乐中的电影推荐、休闲时的音乐推荐等，推荐系统正在一点一滴的渗透我们的生活。推荐系统在各大网络平台得到了广泛应用，如亚马逊、Netflix、淘宝、京东等都采用了推荐系统（Atzori，2010），其不仅能为用户带来便利，还可以有效提高了商家收益。比如，推荐系统的运用使谷歌文章的阅读量提高了 38%，分别为亚马逊图书推荐业务和 Netflix 影片租赁业务贡献 35%和 60%（王毅，2013）；另有研究表明，在商品种类多、价格低的领域，如电影、图书、日用百货运用个性化推荐服务能够使销额提高 2%-8%（张晓彬，2010）。可以看出，在面临大规模数据的前提下，推荐系统的成功使用为各大平台创造了巨大的利润，推动了电子商务的智能化发展，有利于企业迎接新的挑战。因此，研究推荐系统更符合当今时代的节奏和需求，且蕴藏着无限的商机。

1.1.2 研究意义

经过二十几年的发展，推荐系统不仅在产业界获得巨大的成功，在学术界也有越来越多的学者纷纷投入到电商平台的推荐系统研究中。各大电商平台也经过不断的尝试和研究开发了推荐系统，依据用户的历史行为数据，向用户推荐不同的商品，每个用户看到的购物界面并不完全相同，这也使得每个用户都有属于自己独特的网上购物商城。个性化推荐服务不仅满足了用户的个性化需求，提升了用户的购物体验和效率，还提高了购买转化率。在日益激烈的竞争环境下，如果能够有效地利用个性化推荐系统，其也必将为企业创造前所未有的利润。

优质的个性化推荐系统能为用户和商家创造互利共赢的局面，对电子商务的未来发展有十分深远的影响。对于用户而言，首先，我们目前正处在一个快节奏的社会环境中，人们总是希望能够用最少的时间完成更多的事，个性化推荐服务

可以节约用户的购物时间成本，使其花费最少的时间找到最合适的商品；其次，推荐系统可以挖掘用户的潜在需求，带领用户去关注新的领域，满足用户的多样性需求。对于商家而言，个性化推荐系统能够帮助商家做到广告的精准投放，降低多余广告对用户的无端骚扰，提升用户的消费体验，实现用户的高效转化，在用户的一举一动之间发现用户需求，为商家定位高价值的潜在用户。由此可见，个性化推荐系统在电商平台的应用对商家和用户都具有非常重要的实际意义。

在电商企业快速发展和用户需求不断增长的条件下，用户行为变化暗含了用户的消费心理变化，同时，电商推荐技术无论是在理论还是实践上都取得了很大的进步，在电商平台引入个性化推荐系统后，商家产品的销量有明显的增加。虽然目前电商平台推荐领域取得了一定的成果，但仍有一些问题亟需解决和优化，电商平台面临着新的挑战。在推荐系统被广泛应用的今天，其面临的一些限制性问题包括冷启动问题、稀疏矩阵问题和扩展性问题，同时，推荐系统的质量和效率也会随着用户和商品数量的增加受到影响。因此，对电商平台用户行为进行分析，发现用户行为特征与规律，然后并对个性化推荐系统中存在的问题进行解决，将其应用于电商平台中，以此实现电商平台高质量、高效率的个性化推荐正是本文研究的意义所在。

1.2 国内外研究现状

1.2.1 用户行为研究现状

用户行为分析是推荐系统设计的基础，发现用户行为规律能够在恰当的时机进行恰当的商品推荐，这有助于电商平台实现广告的精准投放。对用户行为进行分析是做好商品推荐的基础条件之一，并为进一步提高平台服务质量、提高用户满意度及个性化推荐服务提供依据（Wang 和 Zhao，2018）。电子商务网站内容在日益丰富的前提下，浏览、点击、收藏、购买等行为背后隐藏着用户的心理变化，其中暗含的规律也促使着研究者们进行进一步的分析。

从上个世纪末开始就有国外的学者对线上用户行为进行了分析，Kau 等（2003）采用因子分析和聚类分析方法将 3700 多名调查对象分为 6 类，探讨了他们的信息搜索模式、网上购物动机和关注点。Poel（2004）等研究了不同类型

的因子对网店购买预测行为的影响，发现详细的点击流变量是对客户进行购买预测的最重要变量。Benevenuto 等（2009）通过对详细的点击流和用户交互数据进行分析，发现浏览行为占所有用户行为的 92%，通过浏览好友页面会增加用户之间的交互量。Close 和 Kukar（2009）调查了线上消费者购物车的使用情况，发现用户使用购物车的动机有四个：（1）了解促销活动（2）获取更多的产品信息（3）方便多产品一起购买（4）将购物车看成个人愿望清单，同时还发现购物车的使用可以提高用户的购买意愿。Reza 等（2012）利用客户流失指数和决策树 CART 分析客户流失的原因，从而使企业选择更好的策略来减少用户流失。Krishna 等（2015)提出了 CIUBSM 模型为特定用户提供重要或有用的服务。Triyani 和 Diah（2018）对 100 名消费者分析，发现电子商务的易用性对用户行为会产生正向的影响，而有用性和信任并不影响用户的行为。Zhu（2021）通过对新媒体用户行为分析发现了用户的行为特点，为新媒体的业务运营提供支持。

在国内，学者们通过研究深入挖掘了用户行为，同时，为搜素引擎的优化提供了参考。岑荣伟（2010）对 7.56 亿条线上用户数据进行分析，挖掘了用户搜索行为的相关特征和行为差异性，为搜索引擎算法和界面的优化提供了一些参考。袁兴福(2015)应用“状态-行为”模型分析了电商平台用户行为日志，通过建模、聚类等方法得到 8 类具有显著特征的用户，描述了更丰富的电商平台用户行为特征。另外，有学者提出了用户行为分析、建模和预测的各种算法，如雷名龙（2016）分别分析了用户行为和商品的特征，采用逻辑回归、SVM 和随机森林模型完成了用户行为的预测。于泽川（2019）利用 LSTM 算法构建了用户行为模型，更加准确的反映了用户的流量使用情况。朱珏樟（2020）利用模型进行排序的方法来提高预测用户行为、推荐用户商品的效率和准确率，这对于将推荐算法更好地应用于实际场景有着一定的借鉴意义。李志勇（2021）针对性的提出用户行为分析的六种模型，让那个用户行为分析更加可靠。

1.2.2 推荐算法研究现状

随着数据挖掘和人工智能等技术的不断发展，推荐系统所采用的推荐技术已经由传统的单一的推荐算法向集成的改进的推荐算法演进。依照推荐算法的不同，传统的推荐系统可以大致分为基于关联规则的推荐、基于内容的推荐以及基于协

同过滤的推荐。

基于内容的推荐方法主要是利用用户偏好特征和项目特征的相似性为用户产生推荐。其主要被用来处理文本类数据，且该推荐技术应用范围也比较广泛，例如 Yahya 等（2011）通过 Web 信息挖掘对用户进行网页推荐，Luis 等（2017）将健康消费者链接到 MedlinePlus 的知名健康教育网站，以获取 YouTube 上的健康视频，从而向用户推荐健康教育视频。Hariri 等（2018）提出了一种基于内容过滤和相关图书目录的多媒体教育视频推荐系统的体系结构，该系统提高了推荐的多媒体数据的动态性和准确性。Wang 等（2018）开发了计算机科学期刊和会议推荐系统，通过对文章摘要内容的分析来帮助作者推荐合适的期刊或会议。Manjula 等（2019）通过提取期刊内容的关键词，并计算期刊之间的相似度来为用户推荐期刊。

基于协同过滤的推荐方法主要通过用户相似性或项目相似性来产生推荐，该推荐方法应用时间较早且是当前最为流行的推荐方法。在数据量特别大时，协同过滤推荐算法存在着数据稀疏性问题、扩展性问题和冷启动问题，针对不同的问题学者们提出了自己的解决方案。

对于数据性稀疏问题和扩展性问题，首先，学者们巧妙的引入了降维技术，通过降维技术能有效缓解这些问题。比如，Peter（2007）使用 SVD 算法对用户-项目矩阵进行降维处理，优化了数据处理方法；方耀宁（2013）利用差分矩阵来表示局部结构信息，优化了基于 SVD 的推荐算法；Zhang（2005）利用反向传播神经网络算法来预测项目评分，弱化了降维推荐算法的缺点，提高了推荐的准确度。其次，虽然降维技术的引入能够减少推荐算法一定计算量，但其导致了大量信息的缺失从而影响推荐的质量。对此，机器学习和数据挖掘技术的快速发展又给了推荐系统新的研究方向，如基于 BP 神经网络的协同过滤算法（Chen，2009）、基于贝叶斯网络建立的推荐模型（Breese 等，1998）、随机游走推荐算法（Hilmi 和 Mukkai，2008）、基于二次多项式回归的推荐（Zhang 等，2018）、基于逻辑回归的推荐（Huang 等，2014）等等，学者们在提出理论的同时也验证了这些算法在实际应用的优势。最后，不论是降维算法还是数据挖掘技术都是在传统推荐算法的基础上做算法的集成，并没有从根本上考虑如何优化用户-项目评分矩阵。因此，有学者开始探究通过填充用户-项目评分矩阵的方法来对传统的协同过滤

算法进行优化，例如张玉芳（2013）等结合条件概率算法来发现最近邻用户，根据最近邻用户对项目的评分来预测填充评分矩阵；毕闰芳（2018）利用 SVR 模型对原始评分矩阵进行填充，再进行协同过滤给出初步的推荐结果，最后再借助用户画像产生最终的推荐结果，从而提升了推荐质量；王志远等（2020）提出了一种基于用户兴趣差异的评分矩阵填充方法。

对于冷启动问题，不少学者也为此进行了不少的研究，常用的方法就是引入用户相关属性信息来减小推荐的误差，如联合时间因素和地理因素的兴趣点构建推荐系统（夏萍萍，2020），将二值化用户属性加入隐语义模型，根据属性的相似性来寻找目标用户的相似用户（巫可等，2016）。另外，有学者同样做了算法的集成来优化单一算法，Russell 等（2007）提出离散小波变换方法来压缩评分矩阵，然后对压缩后的数据进行协同过滤的推荐，这提高了预测精度和计算速度。魏琳东（2017）提出了一种基于矩阵分解和神经网络映射的推荐算法来解决冷启动问题。任永功（2020）将基于协同过滤的推荐算法与关系挖掘方法相融合，解决了冷启动问题完成了个性化推荐。

基于关联规则的推荐系统是依据商品之间的关联效应而产生的推荐。Agrawal R 等在 1993 年最早提出了关联规则算法，其主要被用来发现购物车中商品之间的关联性。目前用来建立关联规则最为经典的两个算法分别是 Apriori 算法和 FP-Growth 算法。为了提高基于关联规则的推荐性能和效率，众多学者展开了深入的研究，贾桂霞（2016）提出了一种基于关联规则推荐算法设计的推荐系统，其在客户信息不完整的情况下实现了商品的二项关联推荐；陈淑英（2018）利用多维属性间关联规则数据挖掘技术来进行图书推荐，提升图书馆服务能力；李昌盛（2019）提出了服务于高效关联规则推荐的分布式计算框架，将关联规则挖掘与推荐算法无缝衔接；黎丹雨等（2019）提出了一种多层多维的关联规则挖掘算法，并验证了此方法能够优化模型的推荐性能。

1.2.3 推荐系统的应用研究

目前，在推荐算法领域人们已经有了将近三十年的研究历史。1992 年，美国施乐公司研发的 Typestry 系统是第一个推荐系统，同时也首次给出了协同过滤的概念（David 等，1992），这种新颖的推荐算法思想引起了众多学者的关注。

Grouplens 研究组在 1994 年使用协同过滤进行新闻推荐，该系统在推荐领域的成功应用为推荐系统商用化奠定了基础。明尼苏达州大学开发了 MovieLens 研究型电影推荐系统，并公布了电影评分数据集，其在现在推荐系统的研究中仍然被广泛使用（Dahlen 等，1998）。Carnegie Mellon 大学研发了用来推荐电子文档的 ACF 系统，这成为基于项目的协同过滤推荐的一个典型案例。亚马逊在线书城也采用此推荐算法，根据用户的历史反馈信息为用户推荐图书，提高了书城的销量。1997 年，推荐系统被应用于电子商务中，Resnick 和 Varian（1997）认为推荐系统在电子商务中充当导购的角色，主要是用来为用户提供建议和决策支持。2006 年，在当时颇具影响力的影片租赁网站 Netflix 举行了一场电影推荐系统比赛，并悬赏百万美金奖励能够设计出最好的推荐系统的团队，最终这场比赛的冠军由提出了一种基于模型的协同过滤推荐算法的团队获得。这场比赛的成功举办也引发了推荐系统的研究热潮。麻省理工学院研发了音乐推荐系统 Ringo，NEC 研究院为促进学术文献的传播和反馈，开发了学术论文在线数字图书馆，另外还有视频推荐、音乐推荐和社交好友推荐等。

相对而言，国内对推荐系统的研究起步较晚，但受国外先进水平的影响，其也逐渐发展起来。2009 年北京百分点信息科技有限公司成立，创建了第一个研究推荐系统的科研团队；2011 年百度世界大会召开，个性化推荐系统被列为未来互联网发展的一大重要方向，同时百度搜索首页应用到了推荐算法；2013 年，淘宝平台植入了推荐系统，其依托海量数据和强大的云计算功能，提出了千人千面计划，实现了产品的精准营销。近年来，抖音短视频、腾讯视频和网易云音乐等的快速发展都离不开推荐系统的助力，它们均能够投其所好为每个人展示不一样的内容，其强大的个性化推荐功能实现了软件的智能化，其为用户提供了更优质服务，吸引了大量的用户。

放眼望去，目前推荐系统已经被应用到各个领域，成为信息服务不可或缺的一部分。实践证明有效的电子商务推荐系统能够提升企业经济效益，同时也能更好的服务于用户，提升用户的购物体验感，满足用户的个性化需求。

1.2.4 研究综合评述及切入点

个性化推荐系统就是为缓解信息过载和查找有用信息效率低下等问题提出

的一种行之有效的方法，且其在各大领域都有着广泛的应用，尤其在网络购物中扮演着非常重要的角色。综合以上文献不难发现，目前的研究主要集中于协同过滤推荐算法，该算法过程易于理解，且考虑了用户项目之间的相似性，能够满足用户的个性化需求。但对于电商平台而言，当用户和产品数量过多时，由于用户只购买大量产品中的一部分，因此得到的用户行为数据就会非常稀疏，在数据严重稀疏的情况下，推荐系统很难给出推荐结果或者影响推荐结果的准确性。同时，随着用户和商品数量的增加，推荐系统所需的用户-项目评分矩阵会随之增大，推荐系统的计算复杂度也必然加大，这会消耗大量的时间从而无法满足用户的实时性需求，同时海量的数据对机器设备也会提出更高的要求。

针对于此，本文首先对电商平台用户行为进行分析，发现用户在行为转化、行为转化时间间隔和复购情况等方面的特征，然后以推荐系统中的协同过滤推荐算法为核心，将其与 k-means 聚类算法相结合来对电商平台用户进行推荐，希望能够通过此方法缓解推荐系统计算量大、计算复杂度高和数据稀疏性高的问题，进而实现用户的个性化推荐。

1.3 主要研究内容

本文主要研究内容如下：

第一章，绪论。本章介绍了研究的背景和意义，分别针对电商平台用户行为分析、个性化推荐算法和推荐系统应用的国内外研究现状做了简单的阐述，通过文献的阅读、整理和分析了解到目前的研究现状及不足之处，找到本文的切入点，为文章后续的展开做好铺垫。

第二章，电商平台用户行为的描述性分析。本章首先介绍用户行为数据的基本情况，并对数据进行了相应的预处理，然后主要从用户行为整体情况、行为转化情况、行为转化时间间隔和用户复购情况这四个方面进行分析，深入挖掘用户的行为模式，探索用户行为变化背后的影响因素，从而更好的了解用户，为进行个性化推荐打好基础。

第三章，推荐算法介绍、对比和选取。本章首先介绍了常用的三种推荐方法，并分别阐述了各推荐方法的优缺点及其适用的场景，然后将三种推荐方法放在一起进行了比较，根据本文的数据特点和研究的需要，选取了基于用户的协同过滤，

最后重点介绍了基于 k-means 聚类的协同过滤推荐算法的原理和步骤，为实证打好理论基础。

第四章，实验结果分析。本章首先介绍了评判推荐系统优劣的几个评价指标，并利用用户行为数据构建用户-项目评分矩阵，再对用户进行聚类，然后通过控制变量实验来寻找基于 k-means 聚类的协同过滤推荐算法的最优参数，并将其与传统的协同过滤推荐算法进行对比，验证了融合推荐算法在推荐质量和效率上的优势。

第五章，相关结论与展望。概括总结了本文的研究内容和相关结论，分析了本文研究中的不足之处，并为今后的研究工作进行展望。

2 电商平台用户行为分析

电子商务中的用户行为对商家来说蕴含着无限的价值，可以说用户的行为操作和行为变化都反映着用户的本质需求，包括浏览商品详情页、收藏、加购、购买以及对商品的评价等，这些数据是挖掘用户行为特征并完成个性化推荐不可或缺的一部分信息。因此，本章将介绍电商平台用户行为数据的基本情况，对数据做一定的预处理，主要从用户行为整体情况、行为转化情况、行为转化时间间隔和用户复购情况这四个方面进行分析，深入挖掘用户的行为模式，探索用户行为变化背后的影响因素，从而更好的了解用户，为进行个性化推荐打好基础。

2.1 数据介绍

2.1.1 数据描述

本文使用了阿里巴巴提供的电商平台用户真实的行为数据集，包括阿里天猫平台用户在 6 月 1 日到 6 月 30 日期间的消费行为数据。该数据来自阿里天池平台，是阿里天猫平台用户的线上线下消费行为数据。用户行为数据集中相关字段包括用户 ID（user_id）、商品 ID（item_id）、用户行为类型（action_type）、品牌 ID（brand_id）和时间戳（time_stamp）等，用户行为表和用户行为数据样例表如表 2.1 和表 2.2 所示：

表 2.1 用户行为表

变量名称	说明	具体描述
user_id	用户 ID	不同的 user_id 代表不同的用户
item_id	商品 ID	不同的 item_id 代表不同的商品
cat_id	商品类目 ID	代表商品所属的类目
seller_id	商家 ID	代表商品所属的商家
brand_id	品牌 ID	代表商品所属的品牌
action_type	用户行为类型	0 为点击行为 1 为收藏商品行为 2 为加购行为 3 为购买行为
time_stamp	时间戳	行为发生的时间

表 2.2 用户行为数据样例表

user_id	item_id	cat_id	seller_id	brand_id	time_stamp	action_type
328862	524981	664	2382	1272	602	0
356311	1017725	821	2768	7735	604	2
102269	224954	962	786	5245	607	0
92396	802458	662	554	6154	626	0
2859	1101304	302	1663	4874	625	3
153790	264177	517	4950	5472	615	0
359701	831737	273	3614	7573	618	0
231204	65743	151	1056		627	1
413606	376771	1075	4845	3345	607	0

数据来源：阿里云天池平台 https://tianchi.aliyun.com/dataset/dataDetail?dataId=42

2.1.2 数据预处理

在进行数据分析之前，我们首先要对数据进行清洗，将数据集中的“脏”数据进行剔除，以免其影响分析结果和模型的构建。通过分析发现部分用户的行为数据特别高，超出了合理范围之内，这可能与爬虫用户和刷单用户有关。爬虫用户为了获取网页信息会多次通过程序对某一界面进行操作，从而产生行为数据使得用户行为数据量过大，但是爬虫用户并不会真正的购买，这会对用户行为分析和产品推荐结果的准确性造成一定的干扰。刷单用户可能为了提高店铺产品的销量或为获得更多的商品好评，有意的对某一商品进行大量的操作，这种行为不能反映消费者的真实购买需求和用户的兴趣偏好，这些数据还会对后文的分析和建模造成干扰，所以对于这两类用户我们应将其从数据集中剔除。另外，通过查询发现部分用户行为为空值，对于用户行为缺失的数据也将其从数据集中剔除，从而保证数据的完整性。

2.2 用户行为分析

电子商务的用户行为主要指用户在网上购物过程中产生的一些消费行为。比如我们在网上购物时，首先打开 app 登录自己的账号，根据自己的需求按照关键词搜索想要的商品，然后点击浏览商品的信息，查看商品的规格和参数，并反复

对比价格、性能、质量、风格和颜色等，其次对于不感兴趣的商品我们会在浏览商品详情页后快速划走，对于感兴趣的商品我们可能会进行多次浏览，查看商品介绍视频和其他用户对商品的评价等，对于比较中意的商品我们会将其收藏、加入购物车或者直接下单购买，直到最终退出网站，这整个过程产生的数据都会被平台记录下来，都可以说是用户行为。

用户在浏览过程中会留下一些基本信息，比如用户是男性还是女性、大概处于什么年龄段和人生阶段、消费能力高低、用户所处的地理位置、关注的店铺有哪些、喜欢什么品牌、最近一段时间浏览什么类型商品的次数比较多等，这些特征都是可以通过分析得到的；有购买需求的人则会多次浏览同一商品或者同类型的商品，反复对比商品价格、规格、性能等，因此会出现浏览量上涨的情况，通过这些浏览行为就能分析出用户感兴趣的商品类型、想要购买的品牌、意向商品的价格区间等；用户收藏商品或者店铺就代表用户对其有浓厚的兴趣偏好，加入购物车代表用户对此商品有购买意向，这些行为都是用户行为数据中隐藏的重要信息；有些用户在浏览商品时一般不会立马下单，需要反复对比之后选取自认为最合适的产品再进行购买，当不同行为转化到购买的时间间隔越长，说明消费者愈加犹豫，尤其对于女生这种现象最为明显，犹豫时间越长，购买的可能性便会越小，用户的这些行为特征和行为习惯都是可以通过大量数据信息挖掘出来的。

2.2.1 用户行为整体分析

本节主要从整体上来对用户行为数据进行分析，观察用户浏览、收藏、加购、购买四种行为的变化趋势，以及在促销活动前后不同行为的异同。对用户行为的整体分析可以发现用户在不同时间节点的消费习惯和消费倾向，预测活动期前后互动人群量级，从而挖掘潜在的消费者，进而提高产品销量和店铺销额。由于在“618”活动[①]时，为了刺激用户消费提高商家的销额，电商平台会组织各种互动活动引导消费者直接与品牌或者商品进行互动，导致活动期各种行为的数据量均明显高于平销期。为了清晰的观察四种行为的变化趋势，本文计算了每日不同行

[①]“618”活动是指每年 6 月是京东的店庆月，在店庆月京东都会推出一系列的大型促销活动，其中 6 月 18 日是京东促销力度最大的一天，淘宝、天猫等电商平台也将这一天作为全民狂欢的购物节进行商品的促销。

为的占比情况，所得结果如图 2.1 所示：

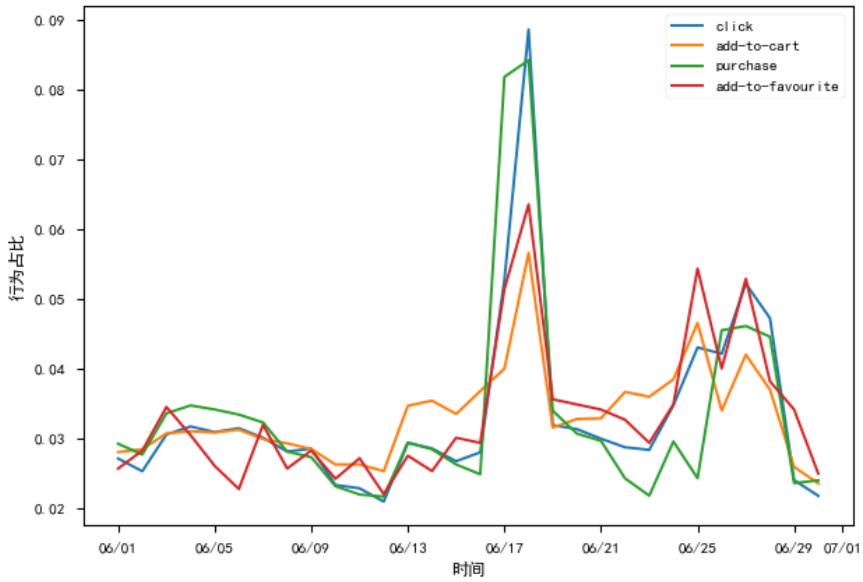

图 2.1 不同行为随时间变化的趋势

从图 2.1 可以发现，四种行为随时间的变化趋势具有相似性，即活动期的用户行为占比显著高于平销期。在“618”活动期间，四种行为的占比均明显高于平销期，其中购买行为最为突出，这体现了商家活动的“大促效应”和用户的“促销心智”，也即用户在平销期浏览、收藏、加购商品的行为较多，这主要为正式活动而做准备，活动开始以后用户直接进行购买，从而导致了活动期购买行为的爆发；另外，我们还能发现 6 月 25 日至 6 月 28 日期间，各种行为占比出现了第二次高峰，分析原因有三个：一是因为商家的二次促销活动。商家乘着活动的余热进行了二次促销活动，吸引用户在平台继续购物，从而导致各种行为大大增加；二是受已购用户的比价、退货、再购买、查询物流信息等行为的影响。对于在活动期已经发生购买行为的用户，用户可能会存在比价心理，用户在收到货后可能会查看自己在“618”活动期间购买的商品是否存在降价或者提价情况，这也是影响浏览量上涨的一大重要因素，对于活动期产生了购买但不满意的商品，用户会进行退货并出现二次购买，所以各种行为必然会增加。另外，由于活动期间商家销量大幅提升，店铺发货数量较多，物流压力大导致快递积压，部分商品无法及时送到用户手中，用户必然会查询物流信息，这也会使得用户行为增加。三是

受“超品”[②]活动的影响。在 7 月份各大品牌会举办“超品”活动，作为推广品牌的一次重要活动，商家会在此次活动中投入大量的资金，并在这之前投放大量的广告为活动蓄水，从而导致用户行为的增加。

商品浏览量、浏览的用户人数、跳失率等指标通常被用来分析用户最活跃的日期，从而挖掘用户的行为习惯。在时间段内，商品浏览量 5715979，浏览的用户数为 242012，进一步分析发现，只有浏览行为，没有进一步消费行为的用户数量为 28727，故跳失率为 11.87%，跳失率还是比较高的，说明店铺的商品详情页虽然能够吸引部分用户进行下一步的消费行为，但还是有 11.87%的用户在浏览商品后产生了流失。图 2.2 反映了每日访问量和访客数的变化轨迹，每日访问量是指用户行为中每日的浏览总量，每日访客数是指每日浏览商品的用户数。

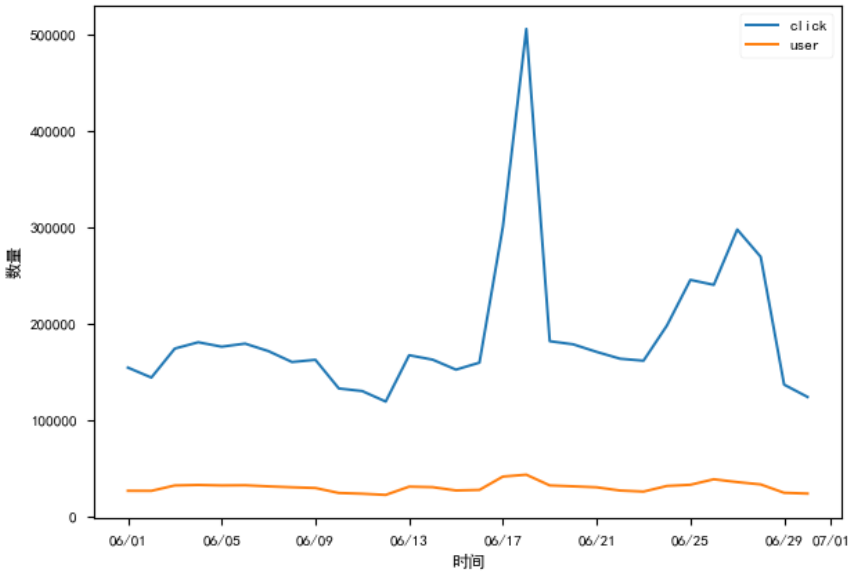

图 2.2 每日访问量和访客数

从图 2.2 可以看出，每日访问量和访客数具有相同的变化趋势，“618”活动期的人群量级和访问量级均显著高于平销期。在 6 月 1 日到 6 月 30 日期间，平销期时访客数基本保持在稳定的水平，变化幅度不大。但从“618”活动开始访客数量大幅增长，出现了两次小高峰；从每日访问量来看，其基本和每日访客数

② “超品”是指天猫超级品牌日，是淘宝商城（天猫）联合商家进行的一种促销活动。参与活动的商家在这一年设定了自己的“品牌日”，并于自己品牌日的当天将新品进行首发,或者将产品进行打折促销。

保持相同的变化趋势，6 月 18 日当天的访问量增长迅猛，约为平销期的 2.76 倍，6 月 25 日至 6 月 28 日也显著高于平销期，推测原因可能是商家的二次营销活动、用户的比价心理和“超品”活动的蓄水等。

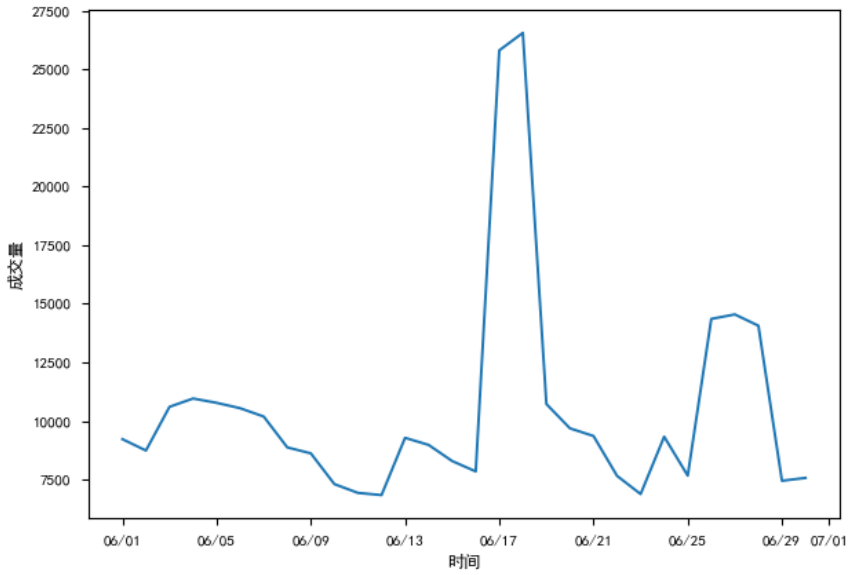

图 2.3 每日成交量

图 2.3 反映了每日成交量的情况，对比发现，其与每日访问量具有相同的变化趋势。从图中可以看出，在平销期每日成交量约为 1.1 万左右，在正式活动开始之前，成交量都有明显的下降，均显著低于平销期成交量，但在活动前某天也会出现峰值，这与商家在正式活动之前做的预售营销活动密不可分，用户会在预售期对自己想要购买的产品预付定金，这导致了活动前用户成交量的增加。从图中可以看出活动当天的成交量异常突出，约为平销期的 2.5 倍，这说明促销活动效果显著。大型促销活动是商家做好品牌宣传提高店铺销额的一次重要机会，商家应牢牢把握。因此，在活动之前商家应加大广告投放力度，重点针对目标人群进行广告投放，做好老客的购买转化和新客的拉新工作，为活动当天的爆发而蓄水，吸引更多的用户浏览商品、关注商品来提升成交量。网络平台也应该根据用户的历史行为，发现用户的行为特征从而挖掘其潜在需求，及时有效为用户做好产品的推荐满足其不同的需求，提升平台的服务质量和效率，为用户节省时间并让其能够快速找到自己喜欢的商品。商家对目标人群的确定和平台服务质量的提升是提高商家销售额的重要途径，是达到消费者和商家共赢的有效举措。

2.2.2 用户行为转化分析

对用户行为进行整体分析后，我们还希望能看到用户行为的转化情况如何。对用户行为转化情况进行分析，有助于企业发现高转化率用户，挖掘高价值人群，让企业的广告营销更加精准、有针对性，这样可以以最小的投入获得最大的收益，企业在降低广告成本的同时还能有效提高用户购买转化率。下面我们将从浏览转化率、收藏转化率、购物车转化率等指标来对用户行为转化情况进行分析，追踪用户的行为变化轨迹，寻找用户行为转化规律，从而更好的了解用户心理，增加成交量，提高购买转化率。本文首先对各种行为发生次数进行了统计，得到用户不同行为占比情况如图 2.4 所示：

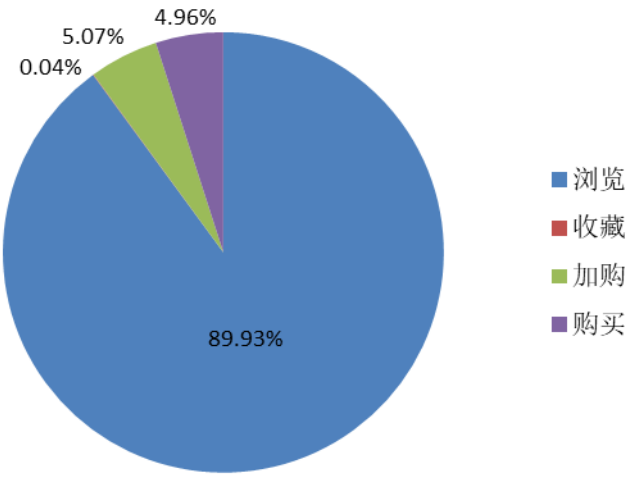

图 2.4 用户不同行为占比

从图 2.4 可以看出，在所有的用户行为中浏览行为占比最高，加购和购买行为占比较为相近，收藏行为占比最低。在所有用户行为数据中，浏览行为占比最大，达到 89.93%，而产生购买行为的只占总行为的 4.96%，也就是说，有约 95%的用户在初次接触商品以后是没有转化到购买，用户在浏览商品后出现了大量的流失。那么，从浏览到购买整个购物消费过程中，每一个环节的转化率是多少呢，哪个环节的流失量比较大？通过对每个购物环节的流失情况进行分析，分析可能的流失原因，可以在以后的销售活动中有效抓住目标用户，防止用户的大量流失。一般来说，用户的购物路径有三条，下面将分别展开讨论：

第一条路径：浏览→购买

这是指用户在浏览商品详情页后直接进行购买。我们定义浏览转化率=浏览商品详情页后购买的用户数/浏览商品详情页的用户数。由分析结果可知，在整个用户行为数据集中，浏览商品详情页的用户有 5715897 人，浏览后购买的用户数有 277745 人，用户从浏览转化到购买的情况如图 2.5 所示：

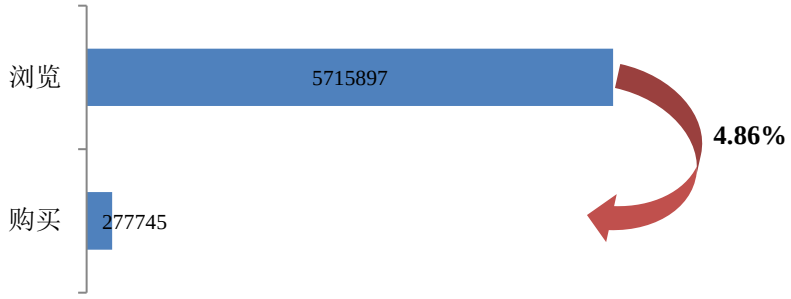

图 2.5 浏览→购买转化情况

从图 2.5 可以看出，浏览到购买的转化率约为 4.86%，也就是说，在浏览商品详情页之后直接购买的用户仅占浏览用户的 4.86%，其余的大部分用户发生了流失，流失用户占高。这可能是因为用户浏览的商品并不是自己想要的商品，他们只是在众多商品中浏览对比然后做出抉择，平台没有及时的挖掘用户的行为偏好为其做好产品推荐，或者是推荐结果不够准确没有满足用户的需求。因此，平台应该不断的完善自己的推荐系统，更深层次的挖掘用户的潜在需求，及时有效的为用户做好产品推荐，提升平台的服务质量和效率。另外，商家也应想办法在用户浏览商品详情页后留住用户，比如优化商品详情页页面、优化页面素材、组织逛店抽奖活动、明确清晰展现店铺的优惠活动、设置弹窗以突出活动重点和入会有礼等活动。

第二条路径：浏览→收藏→购买

这是指用户在浏览商品详情页之后先收藏商品，然后再进行购买。我们定义收藏转化率=添加收藏后购买的用户数/添加收藏的用户数，由分析结果可知，在用户浏览商品详情页后有 2296 人有收藏行为，收藏后购买的用户数有 362 人，用户从浏览转化到收藏再到购买的转化情况如图 2.6 所示：

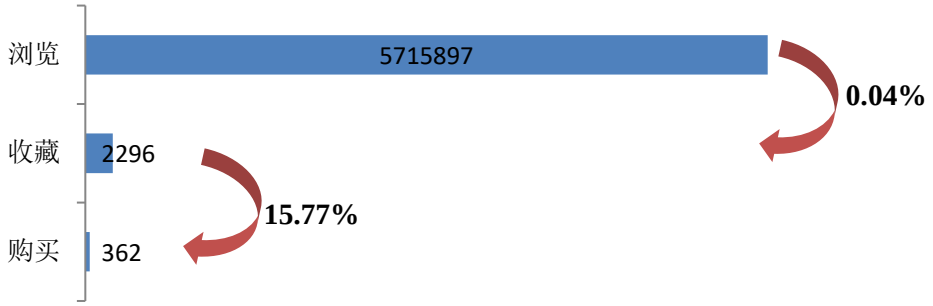

图 2.6 浏览→收藏→购买转化情况

从图 2.6 可以看出，浏览到收藏的转化率仅为 0.04%，转化率极低。推测原因可能有以下几点：一是用户在网上购物时没有养成收藏商品的习惯；二是用户在将商品添加到收藏后并没有可以直接下单的页面，如果想要进行购买，用户必须重新点击商品链接进入详情页，重新选择产品规格才能下单，这中间多了一个操作步骤，对用户来说极为不便，因此从浏览转化到收藏的用户较少。从图中可以看出，虽然浏览商品详情页后收藏的用户不多，但是收藏后购买的用户却达到了 15.77%，收藏转化率较高。推测原因可能是，收藏商品的用户是对商品具有浓厚兴趣也有强烈购买意愿的用户，收藏后转化到购买的可能性比较大，因此对于有收藏行为的用户，商家应注重提高其购买转化率。

第三条路径：浏览→加购→购买

这是指用户在浏览商品详情页之后先加购商品，然后再进行购买。购物车转化率=加入购物车后购买的用户数/加入购物车的用户数，由分析结果可知，浏览商品详情页后有加入购物车行为的用户有 227231 人，加入购物车后再发生购买的用户有 23268 人。由此可知，浏览到加入购物车的转化率约为 3.98%，购物车转化率约为 10.24%，也就是说，加入购物车的用户中，约有 10%的用户会进行购买。用户从浏览到加入购物车再转化到购买的情况如图 2.7 所示：

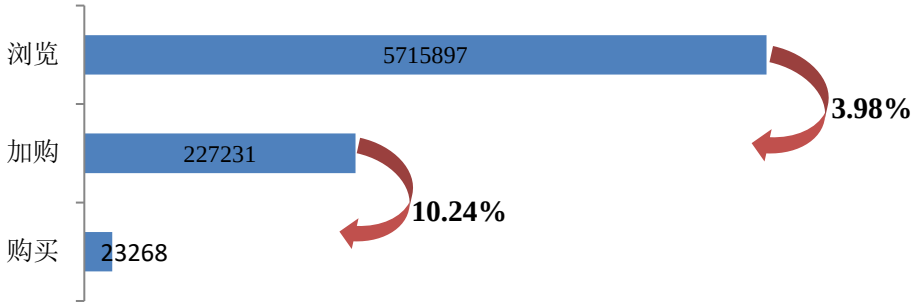

图 2.7 浏览→加购→购买转化情况

从图 2.7 可以看出，用户比较倾向于使用购物车，且购物车转化率也较高。浏览商品详情页之后约有 3.98%的用户有加入购物车的行为，对比发现浏览到加入购物车的转化率高于浏览到收藏的转化率，这说明，如果用户对商品有购买意愿，相对于收藏行为，用户更倾向于将商品加入购物车。但是，在用户加入购物车到购买的环节中，只有 10.24%的用户产生了真实的购买，而约 90%的用户是没有进一步的购买行为，在加入购物车以后发生了流失。推测原因可能有以下几点：一是用户加入购物车是为了与其他商品作价格的比较，最终选择自认为性价比较高的一个来购买，所以加入购物车行为比较多，而购买行为比较少；二是用户在等待活动优惠。将商品加入购物车说明用户有购买意愿，但可能因目前商品优惠活动力度不足，用户在等待活动优惠，所以迟迟没有下单；三是为了凑单满减而加购。将商品加入购物车可能是为了活动凑单，进行满减，最终可能不会进行购买。四是用户将购物车中的商品作为愿望清单。加入购物车的商品只是作为自己的一个愿望清单，不会转化到购买。对此，商家应该引导消费者多使用购物车，对于有加购行为的用户，当商品有降价时应及时提醒消费者，并设置活动倒计时提示等。

2.2.3 用户行为时间间隔分析

用户的每种购物行为一定程度上反映了用户的心理变化，更深层次的挖掘用户行为变化背后的信息，就可以更好的把握用户购买动向。因此本文分析了浏览、收藏、加入购物车三种行为分别转化到购买的时间间隔。当不同行为转化到购买的时间间隔越长，说明消费者愈加犹豫，其可能在从众多的商品中进行对比从而挑选出最中意的产品，犹豫时间越长，购买的可能性便会越小。

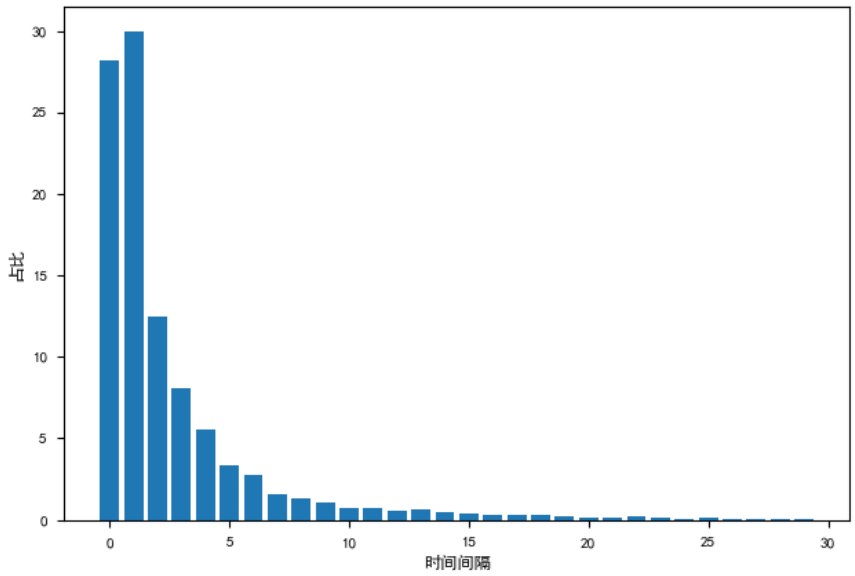

图 2.8 浏览→购买的时间间隔

图 2.8 反映了用户从浏览转化到购买不同时间间隔用户占比情况，x 轴表示用户从浏览转化到购买间隔的天数，y 轴表示符合这种间隔天数的行为数占总行为数的比例。从图中可以发现，从浏览转化到购买的时间间隔大多在 15 天以内，即用户通常是在浏览商品 15 天内就会购买此商品。因此商家在做广告投放时，可以重点针对近 15 天有过浏览商品行为的用户进行广告投放，这样既能确定产品购买目标群体，又能有效提高产品的购买转化率。

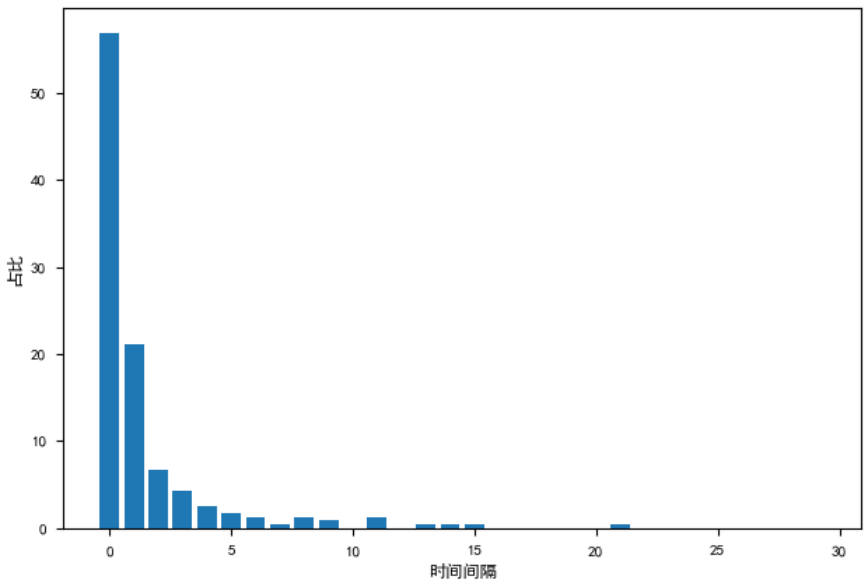

图 2.9 收藏→购买的时间间隔

图 2.9 反映了用户从收藏转化到购买不同时间间隔用户占比情况，x 轴是收藏转化到购买间隔的天数，y 轴是符合这种间隔天数的行为数占总行为数的比例。从图中可以发现，从收藏转化到购买的时间间隔大多数在 10 天以内，即用户通常是在收藏商品 10 天内就会购买此商品。在收藏后近 1 天内发生购买行为的占比达到 75%以上，即用户在收藏后会及时购买产品。同时，从上图可以看出，有部分用户会在收藏商品 20 天后产生购买，这可能是因为用户收藏商品后忘记付款，且没有购买其他类似产品，因此，对于收藏了产品却长时间未购买的用户，商家可以适当“提醒”用户让其付款。

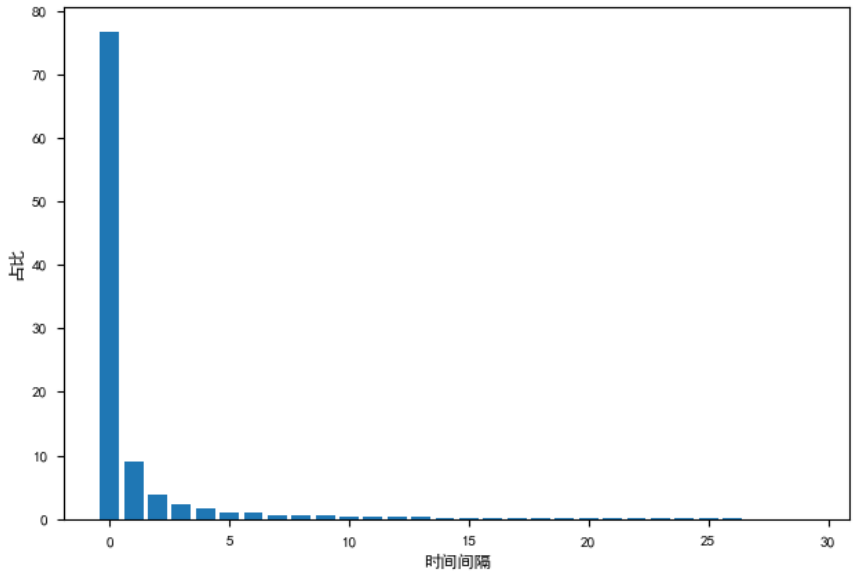

图 2.10 加购→购买的时间间隔

图 2.10 反映了用户从加入购物车转化到购买不同时间间隔用户占比情况，x 轴是加入购物车转化到购买间隔的天数，y 轴是符合这种间隔天数的行为数占总行为数的比例。可以发现，加入购物车转化到购买的间隔天数大多数在 5 天以内，即用户通常是在加购商品 5 天内就会购买此商品，且在加入购物车近 1 内发生购买行为的占比达到 85%以上，即用户在加购后会及时购买产品。同时也不难发现，上图存在明显的拖尾现象，原因可能是用户加入购物车在等待商品的降价，或者部分用户纯属因为喜欢某些商品而将其加入购物车，但是不会产品购买，更有甚者将购物车的商品作为自己的愿望清单。因此，对于有加入购物车行为的用户，店铺可以根据用户的兴趣偏好为其推荐最佳组合搭配商品，在让用户有商品的最佳使用体验的同时还能提升店铺产品的销量；对于加入购物车但长时间未付款的

用户，店铺可在商品降价时及时提醒用户购买或者为用户推荐店铺内价格低、性价比高的相似产品，提高店铺整体的销量。

2.2.4 用户复购行为分析

购买行为可以被视作用户对产品的一种认可，尤其是重复购买。对于优质的产品，重复购买率自然会相对较高，产品重复购买率越高，说明用户对产品的评价越高，其也更容易获得用户的青睐和持续关注。

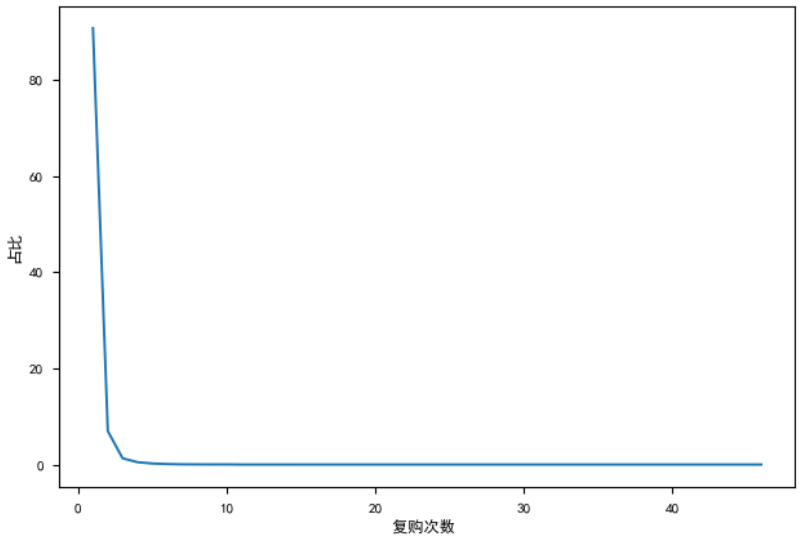

图 2.11 不同复购次数用户占比

图 2.11 反映了不同复购次数用户占比情况，其中 x 轴为用户复购次数，y 轴为对应复购次数用户所占比例。通过分析发现重复购买用户通常较少，大多数用户对于某个产品只购买一次。这可能是因为数据集本身时间周期较短，仅有一个月的数据，对于大多数商品来说，用户不会在短期内发生复购行为，从而导致购买一次的用户占比较高。

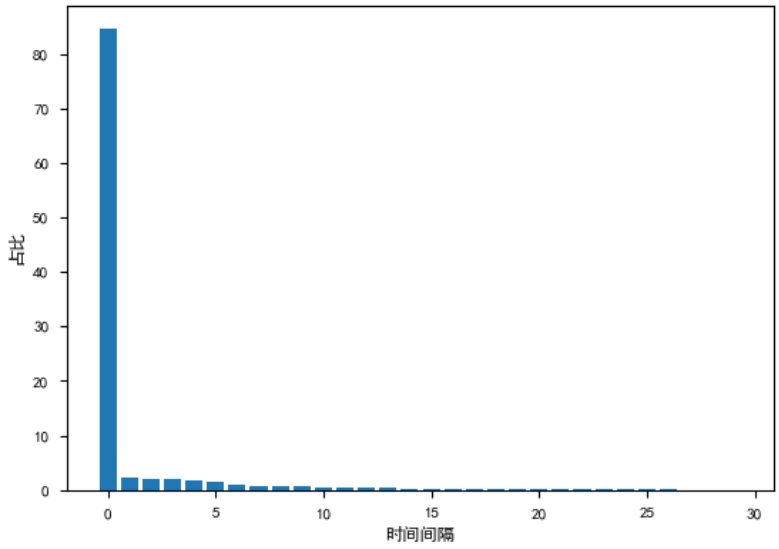

图 2.12 重复购买时间间隔用户占比

图 2.12 反映了用户重复购买时间间隔，X 轴是发生重复购买时距离上次购买的天数，Y 轴为对应复购时间间隔用户所占的比例。进一步统计发现，复购时间间隔在 15 天以内的用户占比较大。即对于某些商品，比如像零食、快消品等，用户在购买过此商品之后，如果对商品的整体使用感较好，短时间内重复购买的可能性较大，因此，在进行用户产品推荐时，也要充分考虑用户的复购周期，不同的商品类型其复购周期必然也会存在差异。

2.3 本章小结

本章通过对阿里平台真实的用户行为数据集进行分析，发现了用户行为在平销期和活动期的不同变化，以及用户行为转化、用户购物时间间隔和用户复购等方面的特征。我们的发现如下：

（1）“618”活动期间，四种行为的占比均明显高于平销期，尤其购买行为占比突出，这体现了活动的“大促效应”及用户的“促销心智”。

（2）用户在浏览商品详情页后流失率高，原因可能是电商平台没有将用户真正需要的东西推荐给用户，导致浏览的商品并不是想要购买的。

（3）用户不习惯收藏商品，但如果有收藏行为那其购买的可能性会上升。

（4）相对于收藏行为，用户更倾向于将商品加入购物车，但购物车转化率不及收藏转化率。

（5）用户从浏览、收藏、加购转化到购买的时间间隔分别是 15 天、10 天、5 天，用户从加购到购买的转化时间间隔最短。

（6）短期内，用户很少会进行重复购买；对于有重复购买的用户，重复购买一个产品的时间间隔一般在 15 天以内。

上述的发现在提供个性化推荐服务时，都能够提供一定的参考和帮助。

3 个性化推荐方法

随着研究的不断深入和科学技术的不断发展，目前已有的个性化推荐方法多种多样，那么选择合适的推荐方法进行电商平台的个性化推荐成为至关重要的一步。因此，本章将对比基于内容的推荐、基于关联规则的推荐和基于协同过滤的推荐三种推荐方法，在了解各推荐方法优缺点的基础上，根据本文的研究需求，选择恰当的推荐方法，为实现电商平台的个性化推荐打好基础。

3.1 个性化推荐概述

随着科技的迅速发展，用户在网上的一举一动被记录下来形成了大量数据，这种数据具有交互性、实时性和社会性等特点，其蕴含着无限的价值。不可否认的是信息时代的不断发展虽然为我们带来了很多机遇，但与此同时也为我们带来了很大的挑战：信息不断以海量方式供给，对海量信息的挖掘成为企业和商家面对的困难和挑战，面对电商平台琳琅满目的商品，用户如何在网上购物中快速找到自己真正需要的也是一大问题。而个性化推荐系统是应对这一问题的有力工具，且其在各大领域已经有了广泛的应用。

个性化推荐本质上是建立在海量数据基础上的一种预测性方法，它可以辅助平台为不同用户提供个性化服务，主动帮助用户做出选择并将其反馈给用户满足其个性化需求。被推荐的可以是任意物品，比如衣服、鞋子、零食，书籍、电影、美食，甚至可以是楼房、景区等等。电子商务推荐系统基于用户之前的行为特征，依托不同的推荐算法，挖掘用户的潜在需求，针对不同的用户推荐不同的商品，为顾客提供不同的购物体验，使每位用户都有属于自己的在线商城。与传统的搜索引擎相比，个性化推荐机制创造了“信息找人”的新模式，它反馈给用户的结果不再单一死板，能够满足不同人生阶段、不同消费能力、不同身份职位、不同兴趣偏好人群的不同需求。其在信息推送者与信息获取者之间架起了快速沟通的桥梁，从而使信息产生了及时有效的价值，且为用户提供了新的购物体验。

3.2 推荐方法的比较

3.2.1 基于内容的推荐

基于内容的推荐根据用户兴趣偏好特征和项目特征的相似性为用户产生推荐。该推荐方法本质上是对项目特征的提取和匹配，它不依赖于项目评分数据，主要取决于用户的偏好属性和项目的相关属性。基于内容的推荐方法首先通过特征提取构建用户偏好和项目属性标签，然后通过信息提取确定用户的偏好特征，再将用户偏好特征和项目特征相匹配，计算他们之间的相似度，最后将相似度高的项目推荐给用户。基于内容的推荐算法原理如图 3.1 所示：

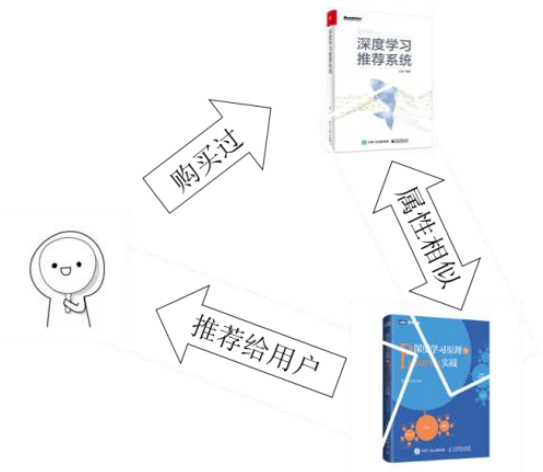

图 3.1 基于内容的推荐算法原理图

以图书推荐为背景，依据图书所属的不同类型，使用基于内容的推荐方法为用户推荐图书。从图 3.1 可以看出，假如某一用户购买了《深度学习推荐系统》这本书，其是一本数据挖掘领域的书，我们根据该用户的历史行为发现他比较喜欢阅读数据挖掘相关的书籍，在收集到他的偏好之后，系统最终会将类似《深度学习原理与 Python 实战》系列的书推荐给该用户。可以看出，基于内容的推荐算法是挖掘了用户过去的兴趣偏好，依据文本内容的相似度来为用户做产品推荐，该推荐方法有如下优缺点：

优点：1）推荐过程并不复杂，推荐结果比较直观，且可解释性强；2）在一定程度上可以解决项目冷启动问题，当有新项目加入时，只需要通过数据挖掘技

术提取新项目的特征属性，然后将此特征与用户偏好特征进行匹配，既可以完成推荐；3）对于一些新产品和比较冷门的产品，它也能够通过特征提取完成推荐。

缺点：1）新用户的冷启动问题，新用户在刚刚加入系统时没有历史行为，因此也就无法挖掘其兴趣偏好，推荐算法很难实现；2）此算法对数据信息分类要求比较高，内容的特征抽取比较困难，对于像音乐、视频等多媒体数据，其内容特征比较复杂，特征信息不易提取，推荐结果的准确度也就会随之降低；3）只推荐特征相似的产品且仅基于用户历史偏好推荐，不会产生多样性的推荐，具有一定的局限性。

3.2.2 基于关联规则的推荐

基于关联规则的推荐主要是根据商品之间的关联性，依据用户曾经喜欢的物品向其推荐与该物品关联性较高的物品。其关键点在于利用历史数据挖掘商品之间的关联效应。比如某位用户购买了一台电脑，在一般情况下该用户还会再购买鼠标等电脑配件，因此，在电脑和鼠标之间可以建立关联关系，在之后的推荐中可以根据这种关联关系来产生推荐。比较著名的关联效应是“啤酒和尿布”的案例（陈正明，2005），在超市中将啤酒和尿布陈列在一起，这一非同寻常的做法使得啤酒和尿布的销量同时增加，这正是沃尔玛超市通过对一年的交易数据进行详细的分析发现了这两者之间的关联效应。在电子商务网站购物时，也同样用到了这种思想。基于关联规则的推荐系统正是利用商品之间的关联规则来做商品的推荐，这不仅为消费者带来了方便，还可以促进商家产品销量的提升。

基于关联规则推荐的核心要点是需要通过对历史数据分析发现商品之间关联效应，然后根据用户的历史行为和这个关联效应进行推荐。形如：$X \Rightarrow Y$，表示 X 事件发生后 Y 事件也有一定发生的概率，这个概率就是对历史数据的统计与计算得到的。在利用关联规则做产品推荐之前，我们应该首先能够找出具有关联关系的产品，其中衡量关联规则的指标有两个，一个是支持度（Support），另一个是置信度(Confidence)。支持度是指在所有产生了消费行为的用户中，既购买了 X 产品又购买了 Y 产品用户占比。比如今天发生购买行为的用户总数是 100，其中，既购买 X 又购买 Y 的用户有 30 位，那该关联规则的支持度为 33.33%，如式 3.1 所示。

$$Support(X \Rightarrow Y) = P(XY) = Support(X \cup Y) \tag{3.1}$$

而置信度是指在所有购买了 X 的用户中有多少用户购买了 Y，也就是说依据 X 推荐 Y 的可信程度，如式 3.2 所示。

$$Confidence(X \Rightarrow Y) = P(X \mid Y) = \frac{P(XY)}{P(X)} = \frac{Support\_count(X \cup Y)}{Support\_count(X)} \tag{3.2}$$

在实际应用的过程中，我们可以根据自己的不同需求指定最小支持度阈值(Min_Support)和最小置信度阈值(Min_Confidence)，只有满足 Support($X \Rightarrow Y$) ≥ Min_Support，且 Confidence($X \Rightarrow Y$) ≥ Min_Confidence，才可以认为 X 和 Y 之间具有强关联，这样推荐成功的几率也会较高。

基于关联规则的推荐算法优点是非常简单易理解，且对数据的要求不高，不依赖用户对商品的评分数据，只要提前发现了关联效应那推荐结果就会比较准确。但它存在的缺点有：1）因为基于关联规则的推荐是根据用户的偏好特征，在已建立好的规则库中匹配对应的商品产生推荐，因此推荐的个性化程度不高；2）关联规则需要对海量的项目信息进行分析，当数据量特别大时，关联规则的发现过程尤为耗时，其规则的抽取和整理难度较大，关联规则难以建立；

3.2.3 基于协同过滤的推荐

在推荐领域，协同过滤推荐算法在推荐领域扮演着举足轻重的角色,它是目前使用最广泛的推荐方法之一。在日常生活购物中，面对众多的商品不知如何选择时，我们往往会向身边好友询问意见，如果自己身边的很多朋友都推荐某一种商品，那么我们就会很大概率的选择该商品；或者我们喜欢某一类产品，对于具有这种特征的产品我们也会更倾向于去选择它。协同过滤正是把这一思想运用到个性化推荐中来。协同过滤推荐算法有两个子类（Sarwar 等，2001）：一类是基于用户的协同过滤推荐（User-based Collaborative Filtering, UBCF），另一类是基于项目的协同过滤推荐（Item-based Collaborative Filtering, IBCF）。

（1） 基于用户的协同过滤

基于用户的协同过滤是建立在这样的假设基础上的：如果用户 U_1 和 U_2 对 t 个项目有过相似的评分,或者购买过相似的产品,那么可以认为用户 U_1 和 U_2 是相

似用户，他们也会对其他项目产生相似的评分或相同行为（周春华等，2019）。傅鹤岗和王竹伟（2010）提出，基于用户的协同过滤推荐实质上是使用其他用户的观点来过滤和评价商品的过程。

基于用户的协同过滤推荐算法首先是将一个用户对所有项目的评分作为一个向量，通过计算用户之间的相似度来找到和目标用户兴趣偏好相似的最近邻居，然后根据最近邻居的兴趣偏好和其对项目的评分，预测目标用户对未购买产品的评分，最后将预测评分最高的若干个项目推荐给用户。

如图 3.2 所示，假设用户 a 喜欢物品 A 和物品 C，用户 c 喜欢的物品有 A、C、D，如果需要给用户 a 做产品的推荐，那么首先根据用户的历史行为数据，通过相似度计算可以发现用户 a 和用户 c 相似，然后通过协同过滤，将用户 c 喜欢但目标用户 a 没有关注的物品 D 推荐给目标用户 a。

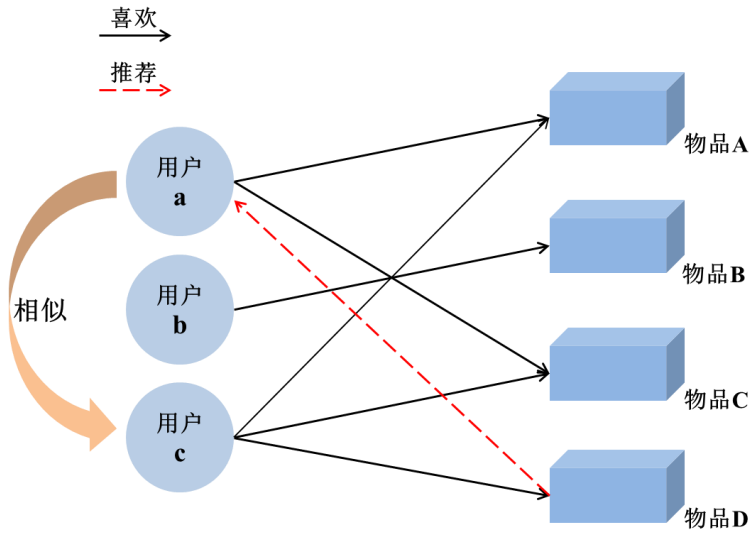

图 3.2 基于用户的协同过滤推荐原理

（2） 基于项目的协同过滤

如果基于用户的协同过滤是参考朋友的兴趣偏好产生的推荐，那么基于项目的协同过滤是基于项目的相似性做出的推荐。基于项目的协同过滤是建立在这样的假设之上的：如果用户均对产品 P_1 和 P_2 有相同或类似的评价，那么可以认为产品 P_1 和 P_2 具有相似性，用户在购买其中一个产品时如果给予恰当的推荐，其也有可能购买另外一个产品。

基于项目的协同过滤推荐首先是构建用户-项目评分矩阵，通过相似度计算

确定项目之间的相似性找到项目的最近邻集合，然后预测目标用户对项目的评分，最后将预测评分最高的若干个项目推荐用户。基于项目的协同过滤推荐是从商品的角度出发来计算的相似度。如图 3.3 所示，根据所有用户的行为数据可以发现，喜欢物品 A 的用户都喜欢物品 D，我们就可以认为物品 A 和物品 D 比较相似。那么在对用户 c 进行商品推荐时，我们根据他的历史行为发现用户 c 喜欢物品 A，而他没有购买过物品 D，根据物品 A 和 D 的相似度可以为用户 c 推荐其可能感兴趣的物品 D。

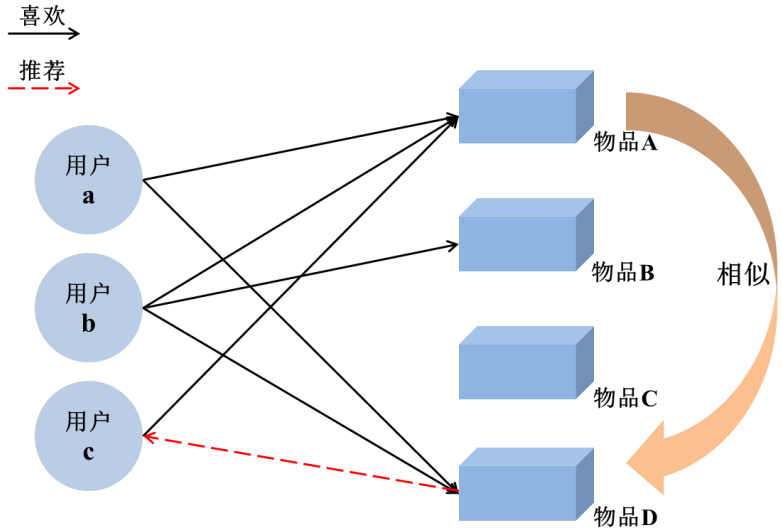

图 3.3 基于项目的协同过滤推荐原理

在实际应用中，基于用户和基于项目的协同过滤推荐应该如何做出选择呢？其实，这两个算法的区别相对来说是比较明显的：第一，二者基于的主体内容不同，一个是寻找相似用户，一个是寻找相似项目；第二，推荐的个性化程度不同，一个是根据相似用户来产生推荐的，因此个性化程度比较高，一个是根据用户偏好的项目来推荐相似的商品，推荐的个性化程度比较低，协同过滤推荐算法有如下优缺点：

优点：1）它是基于用户历史行为产生的推荐结果，因此无需对推荐的产品有先验了解，节省内存空间；2）可以挖掘用户潜在的需求偏好，甚至推荐与用户之前偏好完全不同的商品，商品推荐的个性化程度较高；3）产生推荐所需的数据结构比较简单，只需要用户-项目评分矩阵就可以做商品的推荐。

协同过滤推荐方法虽然被广泛应用,但仍然存在一些限制性问题：1）冷启动

问题，在缺少用户历史行为记录时，推荐结果可能不够准确或者根本无法进行推荐；2）稀疏性问题，电商平台上的商品数不胜数，但用户有过消费行为的商品却很少，因此就导致了数据稀疏性问题，极其稀疏的数据可能无法进行推荐。3）扩展性问题，当用户和商品数量的增多时，用户行为数据会显著增加，算法的计算复杂度也会随之增大，推荐系统将变得迟缓，从而降低了推荐的实时性。

3.2.4 推荐算法的比较与选择

不同的推荐算法都有各自的适用范围，但同时也有一定的局限性。各推荐算法优缺点的比较如表 3.1 所示：

表 3.1 推荐算法优缺点比较

推荐算法	优点	缺点
基于协同过滤的推荐算法	(1)不需要提取项目本身的特征，应用范围比较广泛 (2)可以有效的挖掘用户新的兴趣点 (3)个性化、自动化程度高 (4)可解释性强	(1)存在稀疏性问题 (2)存在冷启动问题 (3)存在扩展性问题 (4)系统开始时推荐质量差
基于内容的推荐算法	(1)方法简单有效，推荐结果直观 (2)用户之间具有独立性，可解释性强 (3)新项目不存在冷启动问题 (4)有比较成熟的分类学习方法能够为该方法提供支持	(1)实施较为复杂 (2)个性化推荐能力不强 (3)将内容抽取要求结构化较强 (4)存在新用户出现时的冷启动问题
基于关联规则的推荐算法	(1)技术成熟 (2)多离线处理、实时性好 (3)推荐伸缩性强	(1)个性化自动化程度不高 (2)关联规则难发现 (3)新项目的冷启动问题

相比较而言，基于内容的推荐算法在文本类推荐领域的应用比较广泛，推荐结果的多样性和个性化较差，对于电商平台中用户行为数据并不适合使用此方法；基于关联规则的推荐算法比较简单，推荐的实时性较好，但个性化自动化程度不

高，它主要被用来发现购物车中商品之间的关联性，当数据量特别大时，关联规则的发现过程尤为耗时，关联规则难以建立；基于协同过滤的个性化推荐方法相对来说计算比较简单，推荐的个性化程度较高，可以挖掘用户的潜在需求，而且可解释性强。协同过滤推荐系统的应用场景非常广泛，并且推荐效果和复杂度都是可以让人接受的，协同过滤算法通常可以作为一个稳定的基础应用于推荐系统中，在电商领域最为常见。因此，本文选择基于用户的协同过滤推荐算法来实现电商平台用户的个性化推荐。

3.3 改进的协同过滤推荐算法

众所周知，电商平台用户众多，包含的商品也极为丰富，但用户有过消费行为的物品却很少，且两个用户购买过相同物品的更少，所以根据用户行为数据构建的用户-项目评分矩阵中元素为 0 的项会非常多，这必然产生稀疏矩阵问题。传统的基于用户的协同过滤推荐算法在进行商品推荐之前需要遍历所有用户计算他们之间的相似度从而确定最近邻居，而随着用户和商品数量的增多，计算系统内所有用户之间的相似性变的不太现实，并且也会花费大量的时间，而推荐系统在能够实现推荐的同时也必须考虑到实时性问题。

为了缓解传统的协同过滤推荐算法存在的限制性问题和计算量大实时性差的问题，很多学者将数据挖掘算法和协同过滤推荐方法结合建模，尤其是分类技术的使用能够有效避免协同过滤推荐算法中寻找最近邻居不精确的问题。本文就是将 k-means 聚类算法与协同过滤推荐算法相结合，首先通过聚类压缩最近邻居搜索范围，然后再进行协同过滤推荐，这将使得传统的推荐算法在推荐效率方面有一定的提高。下面将会介绍算法的详细步骤并在最后用真实的用户行为数据进行实证分析。

3.3.1 k-means 算法

k-means 聚类算法是由 J.B.Mac Queen 在 1967 年提出的，其优点是算法简单、高效、准确率高等，尤其对于像商业数据等大数据集其能够表现出较好的可扩展性。k-means 算法在聚类之前首先要根据自己的需求确定 k 个聚类中心，并选择初始聚类中心，通过计算各样本点到聚类中心的距离，根据距离最小原则将

数据集中的样本划分到相应的簇中，更新聚类中心，不断重复这一过程直到目标函数收敛。算法具体过程如下：

（1）假设原始数据集 $R=\{x_1,x_2,...,x_n\}$,样本的个数为 n，其中每个样本向量为 m 维：选取 k 个初始聚类中心 C_j ,j=1,2,2,…k；

（2）计算每个数据对象 x_i 与初始聚类中心的距离

$$D(x_i,C_j),i=1,2,3,...,n,j=1,2,3,...,k \tag{3.3}$$

将该数据对象 x_i 划分到距离最近的类 C_{temp}，如果满足

$$D(x_i,C_{temp})=\min\{D(x_i,C_j)\},j=1,2,3,...,n,temp=1,2,3,...,k \tag{3.4}$$

其中距离测度本文采用欧式距离：

$$D(x_i,x_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+...+(x_{ik}-x_{jk})^2}=\sqrt{\sum_{k=1}^{m}(x_i-x_j)^2} \tag{3.5}$$

（3）计算误差平方和评价函数 J：

$$J=\sum_{j=1}^{k}\sum i\in C_{temp}\left\|x_i-M_j\right\|^2 \tag{3.6}$$

其中 M_j 是各类 C_j 中样本数据对象的均值。

（4）若 $\left|J_{new}-J_{old}\right|<\delta$，则算法结束；否则重新计算每个类的均值 M_j' 作为新的聚类中心， $M_j'=\frac{1}{N}\sum_{x_i\in C_j}x_i$ ，并重复步骤 2 直至算法收敛。

3.3.2 基于 k-means 的协同过滤推荐方法

基于 k-means 聚类的协同过滤推荐算法与传统的协同过滤推荐算法本质上并没有差异，只是在使用融合推荐算法做商品推荐时，使用的数据不再是原始的用户-项目评分矩阵，而是 k-means 聚类算法的输出结果。传统的协同过滤推荐在计算用户相似度寻找最近邻用户时需要遍历所有用户，因此计算的工作量相当大，而本文所使用的基于 k-means 聚类的协同过滤推荐算法首先会通过聚类将目标用户划分到与其距离最近的簇中，然后在所属簇中计算用户之间的相似性，这

两者的主要差别主要体现在最近邻用户的查找范围不同。具体的推荐算法步骤可描述如下：

步骤 1：构建用户-项目评分矩阵。依据用户对项目的评价数据构造用户行为偏好矩阵，且每个用户的行为偏好都可以表示为向量形式。由于本文采用的是阿里天猫平台真实的用户消费行为数据集，而不是用户对项目的评分数据，因此无法直接来构造用户行为偏好矩阵，对此，本文用 1、2、3、4 分别表示浏览、收藏、加购、购买行为。则该向量由用户、项目及用户对该项目的消费行为数值构成，那么所有用户的信息就构成一个矩阵，这个矩阵也称为用户-项目评分矩阵，表示为 R，如下所示：

$$R=\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{bmatrix}$$

其中，m 表示系统中的用户数量，n 表示项目数量，矩阵元素 R_{ui} 的值表示用户 u 对项目 i 的消费行为，也就是评分，R_{ui} 越大，说明用户 u 对项目 i 评价越高，反之越低。

步骤 2：利用 k-means 算法进行用户聚类。随机选取 C_j (j=1,2,2,…k)为 K 个初始聚类中心，根据公式 3.5 分别计算各数据对象到 K 个聚类中心的欧式距离，并将各数据对象划分到距离聚类中心最小的一类中。然后重新计算每个类的均值将其设置为新的聚类中心，并重复上述操作多次迭代之后把 R_{mn} 中的数据划分成 K 簇。

步骤 3：查找目标用户所属簇。根据公式 3.5 计算目标用户 u 与 K 个聚类中心之间的距离，根据距离最小的原则把目标用户 u 归到与其距离最近的簇中。

步骤 4：计算用户之间的相似性，寻找最近邻用户。计算目标用户 u 与簇中其余用户的相似性，将相似度按照从大到小的顺序进行排列，选取前 t 个最相似的用户为目标用户的最近邻居，得到目标用户的最近邻用户集 $N_{uj}(j=1,2,...,m)$。通常来说，相似性反映两个对象或两个特征之间的差异程度，差异程度越大相似度越低；相反，差异程度越小相似度越高。通过相似性计算为目标用户寻找最近邻居是协同过滤推荐算法最重要的一部分，其直接影响推荐结果的准确性，因此

我们必须慎重对待。用户之间相似性的计算方法主要有以下几种：

（1）Jaccard 系数

$$sim(u_i, v_j) = \text{Jaccard}(u_i, v_j)\frac{|N(\text{i}) \cap N(\text{i})|}{|N(\text{i}) \cup N(\text{i})|} \tag{3.7}$$

（2）Pearson 相关系数

$$sim(u,v) = \frac{\sum_{i \in I_{uv}} (R_{ui} - \overline{R_u})(R_{vi} - \overline{R_v})}{\sqrt{\sum_{i \in I_{uv}} (R_{ui} - \overline{R_u})^2} \sqrt{\sum_{i \in I_{uv}} (R_{vi} - \overline{R_v})^2}} \tag{3.8}$$

（3）余弦相似度

$$sim(u,v) = \cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\left\|\vec{u}\right\| \cdot \left\|\vec{v}\right\|} = \frac{\sum_{i=1}^{n} R_{ui} * R_{vi}}{\sqrt{\sum_{i=1}^{n} R_{ui}^{\ 2}} * \sqrt{\sum_{i=1}^{n} R_{vi}^{\ 2}}} \tag{3.9}$$

（4）修正的余弦相似度

$$sim(u,v) = \frac{\sum_{i \in I_{uv}} (R_{ui} - \overline{R_u})(R_{vi} - \overline{R_v})}{\sqrt{\sum_{i \in I_u} (R_{ui} - \overline{R_u})^2} \sqrt{\sum_{i \in I_v} (R_{vi} - \overline{R_v})^2}} \tag{3.10}$$

通过对比本文运用修正的余弦相似度来计算用户之间的相似性，因为利用电商平台用户行为数据构建的用户-项目评分矩阵会有严重的稀疏性，对于稀疏数据其能够表现出良好的性能，且很好的避免了不同用户评价标准不一致的问题。

步骤 5：预测评分，产生推荐结果。找到目标用户 u 的最近邻居后，找出最近邻用户消费过而目标用户 u 没有购买过的商品，利用公式 3.11 预测目标用户 u 对产品的评分，最后将评分靠前的 N 个商品推荐给用户 u。

设 I_u 为用户 u 有过消费行为的项目集合，$I_1, I_2, ..., I_k$ 分别为目标用户 u 最近邻用户集 N_u 中各用户消费的项目集合。则项目集合 $I_{\text{w}} = I_1 \cup I_2 \cup ... \cup I_k - I_u$ 为可能被推荐给目标用户 u 的项目集，那么对任意一个项目 $i \in I_{\text{w}}$，目标用户 u 对项目 i 的评分预测公式如下：

$$P_{u_i} = \overline{R_u} + \frac{\sum_{(v \in N_u(i))} sim(u,v) * (R_{vi} - \overline{R_v})}{\sum_{(v \in N_u(i))} |sim(u,v)|} \tag{3.11}$$

其中，$N_u(i)$ 表示目标用户 u 的最近邻用户中对项目 i 有过消费行为的用户集合，R_{vi} 表示用户 u 的最近邻用户 v 对项目 i 的评分值，$\overline{R_v}$ 表示用户 v 与用户

u 在共同消费的项目集合上的平均评分，$\overline{R_u}$ 则表示用户 u 在全部项目空间 I 上的平均评分，即

$$\overline{R_u}=\frac{1}{|I_u|}\sum_{j\in I_u}R_{uj}, \text{其中} I_u=\{j\in I \text{且} R_{uj}\neq\Phi\} \quad (3.12)$$

3.4 本章小结

本章首先介绍了目前常用的几种推荐方法，并对比分析了它们的优缺点及其适用场景。针对于电商平台，为了实现个性化推荐的目标，且根据本文所用的用户行为数据结构特征，最终选择了基于用户的协同过滤推荐算法来完成电商平台的商品推荐。但同时也发现传统的协同过滤推荐算法存在稀疏矩阵问题、冷启动问题和扩展性等问题，且对于用户和商品数量繁多的电商平台，推荐算法的计算复杂度也极高，推荐的实时性就会受到影响。因此，本文针对这些问题展开讨论，将传统的推荐方法与 k-means 聚类算法相结合，这种集成的推荐方法能够在一定程度上缓解稀疏矩阵问题和扩展问题，并且能够有效降低运算成本，减小运算复杂度，在提高产品推荐效率的同时还能满足实时性需求。

4 电商平台中商品的个性化推荐

4.1 实验流程

在上一章中，通过推荐方法的描述和对比，根据本文的研究特点，最终选择了基于用户的协同过滤推荐算法来做商品的个性化推荐，并将 k-means 聚类算法与协同过滤推荐算法相结合来优化传统的推荐算法在海量的稀疏数据上表现出的推荐性能差和推荐效率低的问题。为了针对于电商平台进行用户的个性化推荐，并验证基于 k-means 聚类的协同过滤推荐算法的性能，本章通过真实的用户行为数据进行了实验，主要流程如下：

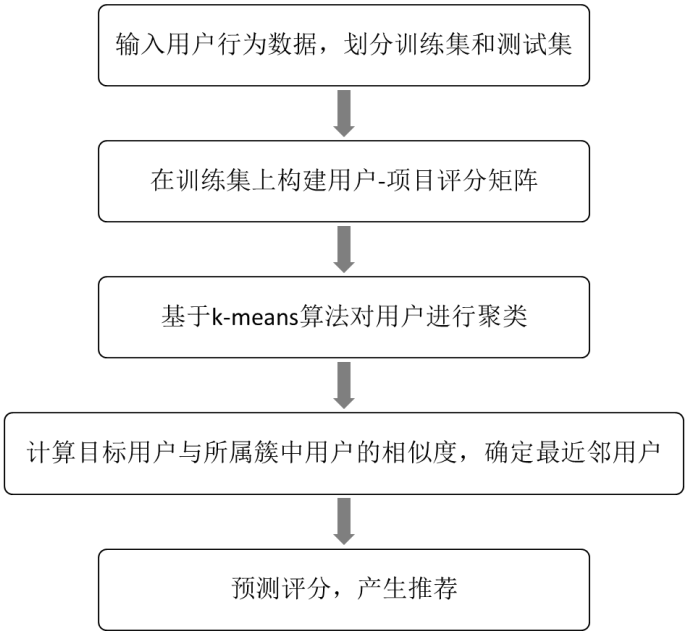

图 4.1 个性化推荐流程图

4.2 推荐系统评价指标

推荐系统中用到了不同的推荐算法，但无论何种推荐算法都需要有一定的评价指标来衡量推荐算法的优劣。目前常用的评测推荐算法的指标主要有：平均绝对误差（MAE）、均方根误差（RMSE）、准确率（Precision）、召回率（Recall）和 F1 评价指标，这些评估指标都是在离线的情况下来计算。

4.2.1 评分预测准确度

在推荐领域评估评分预测精度最著名的方法就是均方根误差（Root Mean Squared Error，RMSE）。RMSE 主要用于衡量推荐算法的评分预测值与实际值之间的差异。RMSE 值越小，说明预测的准确度越高，推荐算法的性能越好。RMSE 定义如式 4.1 所示。

$$RMSE=\sqrt{\frac{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})^2}{|T|}} \tag{4.1}$$

另一个衡量评分预测准确度的方法是平均绝对误差（Mean Absolute Error，MAE），MAE 值越小，说明评分预测的准确度越高，推荐算法的性能也就越好，MAE 定义如式 4.2 所示。

$$MAE=\frac{\sum_{u,i\in T}(r_{ui}-\hat{r}_{ui})}{|T|} \tag{4.2}$$

其中 r_{ui} 代表用户 u 对商品 i 的实际评分，$\hat{r}_{ui}$ 是推荐算法给出的预测评分，$|T|$ 为用户已经有过评分的物品个数。

4.2.2 准确率和召回率

一般衡量 Top-N 推荐的评价标准有准确率（Precision）和召回率（Recall）。假设对于用户 U 推荐商品 P，推荐结果可能出现以下情形：

表 4.1 商品推荐结果对照表

	被推荐的产品数	未被推荐的产品数
用户喜欢的产品数	N_{rs}	N_{rm}
用户不喜欢的产品数	N_{is}	N_{im}
N	N_s	N_m

准确率是指在推荐系统给出的所有推荐结果中，用户喜欢且被推荐的商品所占的比率。准确率越高，说明推荐算法越好，它主要是衡量的查准率。准确率定

义如式 4.3 所示：

$$precision = \frac{N_{\mathrm{rs}}}{N_s} \tag{4.3}$$

召回率是指在推荐系统给出的所有推荐结果中，用户喜欢且被推荐的商品占用户喜欢的所有产品的比率。召回率越高，说明推荐效果越好，它主要是衡量的查全率。召回率定义如式 4.4 所示：

$$recall = \frac{N_{rs}}{N_r} \tag{4.4}$$

其中 $N_r = N_{rs} + N_{rm}$， $N_s = N_{rs} + N_{\mathrm{is}}$

4.2.3 F1 评价指标

我们自然是希望准确率和召回率越高越好，但事实上在一般情况下这两者是矛盾的。在进行算法的评估时，我们不能片面的只使用某个指标来评价算法的优劣，只有同时综合来看才能进行全面的评价。因此，为了综合考虑有学者提出了 F1 评价指标，其也可以看成是准确率和召回率的调和平均数，F1 的数值越大，说明推荐效果越好。F1 评价指标如式 4.5 所示。

$$F1 = \frac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \tag{4.5}$$

4.3 实验步骤与结果分析

4.3.1 划分数据集

在第二章中通过对用户行为数据分析发现在平销期和活动期时用户行为存在明显的差异：在平销期时，用户行为数据基本处在比较稳定且较低的水平；在“618”活动期时，四种行为均有大幅度的提升。因此，为了更好的验证算法的可行性，完成电商平台的商品推荐，在实验时将 6 月 18 日之前的数据作为训练集来训练算法，将 6 月 18 日以后的数据作为测试集来检验算法的优劣。在分析数据时发现：数据集中部分用户行为数据非常少，而对于成千上万件的商品，这

会让矩阵变的非常稀疏。因此在模型训练时，需要对数据进行一定的处理，选择行为数较多的用户来训练模型，计算用户的相似度。

4.3.2 构建用户-项目评分矩阵

协同过滤推荐算法是基于用户的历史行为数据为目标用户做产品推荐的一种算法，它主要依赖于用户-项目评分矩阵来进行评分预测和产品推荐。在电子商务系统中，我们无法得到用户对产品的评分，但我们可以利用隐式反馈信息，即用户在网上购物时产生的消费行为，通过一定的转化和处理可以构建用户-项目评分矩阵。本文采用的是真实的用户消费行为数据集，在此数据集中，用户对商品的行为分为：浏览、收藏、加购、购买四种行为，不同行为代表着用户对商品的不同偏好程度，因此需要对不同的行为赋予不同的权重来表示用户对商品的偏好程度，通过这种方式就可以将用户行为数据转化为用户对项目的评分。对此，本文分别对浏览、收藏、加购、购买行为赋权重为 1、2、3、4，代表用户对商品的评分。用户行为偏好评分表如表 4.2 所示：

表 4.2 用户行为偏好评分表

用户行为	评分
浏览	1
收藏	2
加购	3
购买	4
其他	0

每个用户对所有项目的行为代表着此用户的兴趣偏好，且都可以以向量的形式来表示，那么所有用户的信息就构成了用户-项目评分矩阵。

4.3.3 用户聚类

由于数据集包含大量的用户和商品，不同的用户购买的商品类型差异较大，导致用户对大部分商品没有行为数据，这使得在生成用户-商品评分矩阵的时候会产生稀疏性问题，所以本文利用 k-means 聚类方法先把相似的用户划分成一类，

再进行每一类用户的分析。图 4.3 为用户-项目评分矩阵示例：

表 4.3 用户-项目评分矩阵示例

	P_1	P_2	P_3	P_4
U_1	1	2		
U_2	1	1		
U_3			1	2
U_4			2	2

其中U_i代表用户，P_i代表商品，可见用户U_1和U_2对商品P_1和P_2有过消费行为，U_3和U_4对商品P_3和P_4有过消费行为，如果将这四个用户作为整体运行协同过滤推荐算法，必然存在矩阵稀疏性的问题，因此我们首先通过聚类将兴趣相似的用户划分为一类，然后再对每一类用户进行分析。在此我们可以将用户U_1和U_2划分为一类，U_3和U_4划分为一类分别进行分析。

4.3.4 参数调整

为了得到基于 k-means 聚类的协同过滤推荐算法的最优推荐结果，应该进行控制变量实验，查看不同参数的变化对算法结果的影响。通过前文对算法的介绍我们可以发现，基于 k-means 聚类的协同过滤推荐算法的重要参变量有 3 个：k-means 聚类的聚类个数 K、用户近邻个数 t、推荐列表长度 N。在实证中，将分 3 组进行控制变量实验，比较预测评分与测试集中评分的差异，用 MAE 和 RMSE 值来衡量算法的优劣。实验结果如下所示：

（1）用户近邻个数固定取 t=10，推荐列表长度固定取 N=10，观察随着聚类簇数 K 的变化 MAE 和 RMSE 的变化情况。基于 k-means 聚类的协同过滤推荐算法中聚类簇数 K 的取值对推荐结果的准确度有至关重要的影响，当 K 值较大时，在进行用户聚类时需要占有大量内存且计算消耗时间较长，簇中用户太少也会导致最近邻居的确定不够准确；当 K 值较小时，在计算用户相似度时需要遍历簇中所有用户，K 值较小簇中用户也就越多，计算复杂度也就越大，这将影响

推荐系统的效率。因此，我们需要根据实际情况和多次试验确定最优 K 值。下边从聚类个数 K 对推荐性能的影响进行了相应分析，分别选择 k=5,10,15,20,25,30 来进行聚类，当 K 选取不同的值时，预测评分结果的 MAE 和 RMSE 值如图 4.2 所示：

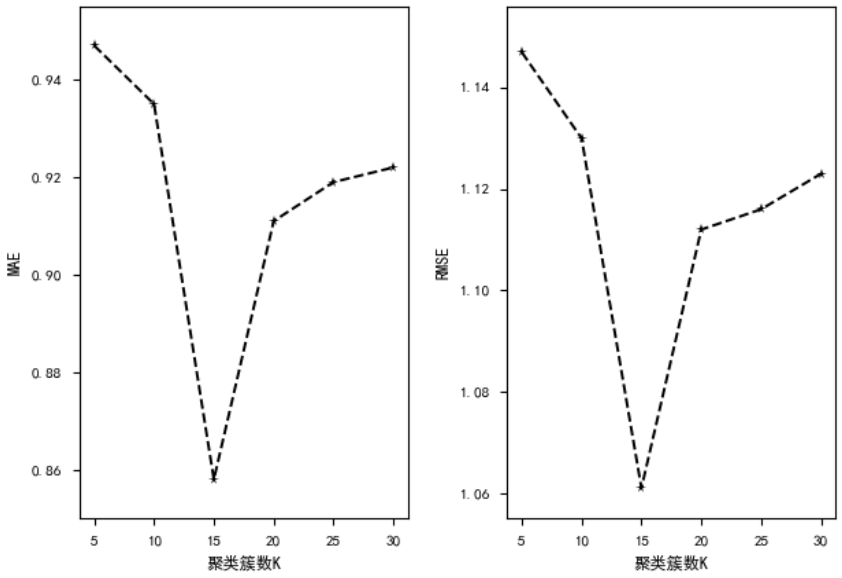

图 4.2 不同的 K 值对应的 MAE 和 RMSE 值

图 4.2 描述了不同聚类簇数 K 的取值下，评估推荐结果的 MAE 和 RMSE 值的变化情况。从图中可以看出，当聚类簇数 K<15 时，随着 K 值的逐渐增大，MAE 和 RMSE 值总体呈下降趋势；而当聚类簇数 K>15 时，MAE 和 RMSE 值总体呈上升趋势。由图可以看出，当 K 取 15 时，推荐误差 MAE 和 RMSE 达到了最小，推荐的效果是最佳的，因此在进行协同过滤推荐之前，设置保留的聚类簇数为 15。在聚类结束后，就可以通过计算目标用户与簇中最近邻用户的相关性，预测目标用户的评分。

（2）聚类簇数固定取 K=15，推荐列表长度固定取 N=10，观察随着近邻用户数 t 的变化 MAE 和 RMSE 的变化情况。基于 k-means 聚类的协同过滤推荐算法中近邻用户数 t 的选择很重要，当 t 值过大时，会导致噪声数据的加入，在确定最近邻用户时会掺杂一切相似度并不是很高的用户，从而影响推荐结果的准确性；另外，如果 t 值过小则会损失很多信息，推荐结果的多样性可能会受到一定的影响。因此，t 值的选择既不能过大也不能过小，需要根据实际情况和多次试

验选择合适的最近邻用户数。有学者提出在近邻用户数取值为 20-50 时协同过滤推荐算法的性能最好，因此本文选取目标用户的最近邻居数 t=10,20,30,40,50 来进行分析，当 t 取不同值时，预测评分结果的 MAE 和 RMSE 值如图 4.3 所示：

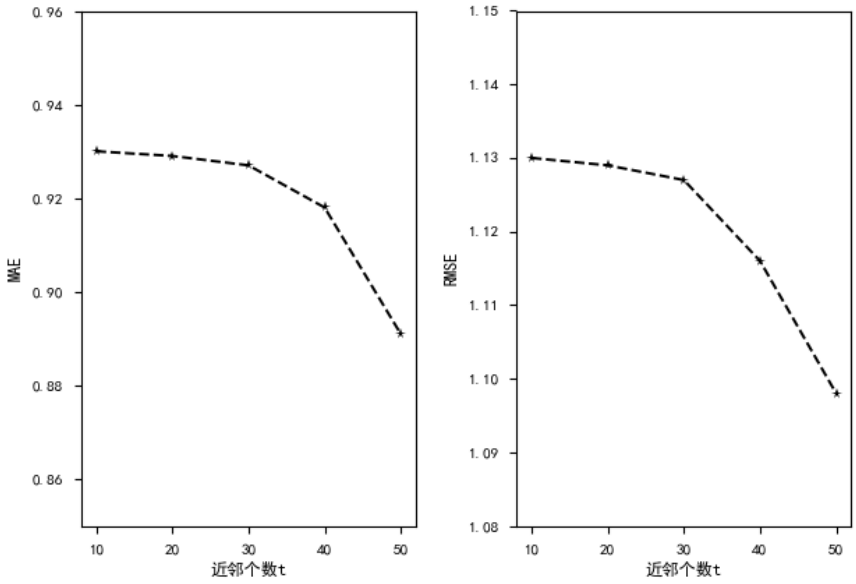

图 4.3 不同的近邻个数 t 对应的 MAE 和 RMSE

图 4.3 描述了随着近邻个数 t 的变化，评估推荐结果的 MAE 和 RMSE 值的变化情况。从图中可以看出，随着近邻个数 t 值的逐渐增大，MAE 和 RMSE 值总体呈下降趋势，当近邻个数取 50 时，此时的评分预测误差 MAE 和 RMSE 达到了最小，评分预测的准确度最高，因此设置保留的近邻个数为 50。

（3）聚类簇数固定取 K=15；近邻个数固定取 t=50；观察不同推荐列表长度 N 对推荐结果的影响。在协同过滤推荐中，常见的应用是 Top-N 推荐，但推荐列表长度 N 的选择对推荐的精度也具有不可忽视的影响。当推荐列表较长，N 值较大时，召回率可能会增大，但推荐精度会随着推荐类表列表长度的增加而减小；当推荐列表较短，N 值较小时，推荐精度可能会增大，但召回率可能会受影响。因此，下边分析了不同推荐列表长度对推荐性能的影响，本文分别选择 N=5,10,15,20,25,30 来进行推荐，当 N 取不同值时，预测评分结果的 MAE 和 RMSE 值如下图：

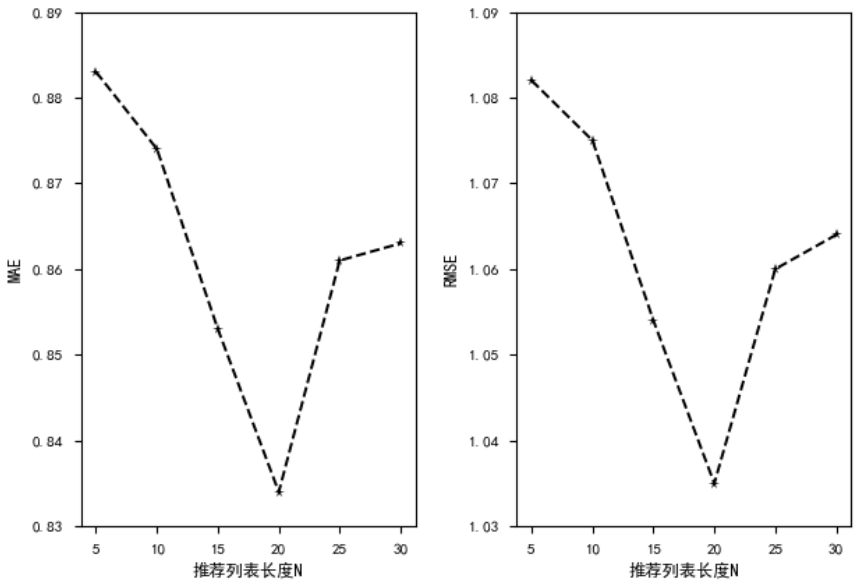

图 4.4 不同的推荐列表长度 N 对应的 MAE 和 RMSE

图 4.4 描述了不同的推荐列表长度 N 的取值下，评估推荐结果的 MAE 和 RMSE 值的变化情况。从图中可以看出，当推荐列表长度 N<20 时，随着 N 值的逐渐增大，MAE 和 RMSE 值总体呈下降趋势；而推荐列表长度 N>20 时，MAE 和 RMSE 值总体呈上升趋势。由图可以看出，当 N 取 15 时，评分预测误差 MAE 和 RMSE 达到了最小，评分预测的准确度最高，此时推荐系统的性能最好，因此，本文设置保留的推荐列表长度为 50。

4.3.5 结果分析

仅通过以上指标去评价一个推荐算法的优劣是很难有说服力的，为了能够进一步验证基于 k-means 聚类的协同过滤推荐算法的真实效果，本文选取了传统的协同过滤推荐算法与其进行比较。通过控制变量实验对重要参数的调整，分析其对推荐结果造成的影响后，本文选取聚类簇数 K=15，用户近邻个数 t=50，推荐列表长度 N=20 时与传统的协同过滤推荐算法进行对比，采用准确率、召回率、F1 评价指标来衡量推荐算法的性能，实验对比结果如下表 4.4 所示：

表 4.4 推荐结果对比

模型	召回率	准确率	F1
UBCF	0.0406	0.0178	0.0247
基于 k-means 的 UBCF	0.0458	0.0275	0.0344

由实证结果可知，与传统的协同过滤推荐算法相比，基于 k-means 聚类的协同过滤推荐算法在准确率、召回率和 F1 评价指标上均有了显著提升。该算法不仅保留了聚类算法拥有的可扩展性，而且在一定程度上提高了推荐的准确度，能为用户做更好的推荐。同时，基于 k-means 聚类的协同过滤推荐算法计算成本较低、计算复杂度较小，算法运算时间也相对较短，其在推荐效率有了很大的提升。因此，对于用户和项目数庞大的电商平台，基于 k-means 聚类的协同过滤推荐表现出更大的优势。

4.4 本章小结

本章介绍了几个评价指标来评估推荐算法的优劣，并根据第三章介绍的算法步骤做了实证分析。这其中主要包括数据集的划分、用户-项目评分矩阵的构建、用户聚类、通过控制变量实验确定最优参数、算法推荐结果的对比分析等步骤。首先，将用户行为数据划分为训练集和测试集，用训练集来构建用户-项目评分矩阵，其次，使用 k-means 算法对用户聚类，保证簇中用户相似度高，簇间用户差异较大，然后，在做协同过滤推荐之前通过控制变量实验确定了推荐算法的最优参数，使得其在最优参数下与传统的协同过滤推荐算法进行对比，最后，通过准确率和召回率等指标验证了本文使用的推荐方法在电商推荐领域的优势和使用价值。实证结果显示，针对于用户和商品数量繁多的电商平台，基于 k-means 聚类的协同过滤推荐算法更符合电商推荐领域的现状和需求，其在推荐性能和推荐效率上均优于传统的协同过滤推荐算法。

5 总结与展望

5.1 总结

中国迎来了 5G 时代，电子商务也愈加深入我们的生活，电商企业的快速发展为人们网上购物提供各种各样的便利。然而，海量数据信息被消费者有效挖掘是件很困难的事情，而推荐技术的应用就是一个缓解信息过载问题的有效方案。个性化推荐服务不仅满足了用户的个性化需求，提升了用户的购物体验和效率，其也必将为企业创造前所未有的利润。因此，本文在对电商平台用户消费行为多方面分析的基础上，将 k-means 聚类算法与协同过滤推荐算法相融合，对电子商务平台上的用户行为进行分析并推荐，并验证了基于 k-means 聚类的协同过滤推荐算法结果的好坏。本文主要做了如下几方面工作：

（1）本文基于电商平台用户行为数据，对用户行为整体情况、用户行为转化、用户购物时间间隔和用户复购行为进行分析。有如下发现：活动期和平销期的用户行为存在差异，用户有明显的“促销心智”；用户在浏览商品后会出现大量流失，所以如果有更加优质高效的个性化推荐系统，也许能够有效提高浏览用户的购买转化率；与收藏行为相比，用户更倾向于使用购物车，但收藏商品后的用户其购买转化率更高；用户从浏览、收藏、加购转化到购买的时间间隔分别是 15 天、10 天、5 天，其中加购转化到购买的时间间隔最短；复购用户较少，短期内用户不会发生复购行为。

（2）在对电商平台用户行为分析的基础上，本文借鉴众多学者的研究经验，介绍了基于内容、基于关联规则以及基于协同过滤三种推荐方法的基本推荐原理，并对比分析了它们的优缺点及其适用场景。为了能够借助用户行为数据实现电商平台用户的个性化推荐，本文最终选择了基于用户的协同过滤推荐算法，但是发现其存在冷启动问题、稀疏矩阵问题和扩展性等问题，同时，当数据量特别大时算法的计算复杂度会增加，推荐效率会受到影响。因此，本文将 k-means 聚类算法与基于用户的协同过滤算法相结合来完成电商平台商品的推荐。

（3）在进行个性化推荐时，考虑到不同的参数选择对推荐结果会产生至关重要的影响，基于 k-means 聚类的协同过滤推荐算法中影响其推荐结果的参数分别有聚类簇数 K、目标用户最近邻个数 t 和推荐列表长度 N。因此，本文通过控

制变量的实验来确定推荐算法的最优参数。另外，将本文所使用的推荐算法与传统的协同过滤推荐算法进行对比，发现本文所使用的推荐方法缓解了数据稀疏性问题和扩展性问题，在大大减少计算量的同时也提高了推荐质量。对于电商平台，基于 k-means 聚类的协同过滤推荐算法更符合电商推荐领域的现状和需求，其在推荐性能和推荐效率上均优于传统的协同过滤推荐算法。

5.2 展望

随着电商企业的增多，其竞争也愈加激烈，商家纷纷向数字化科技转型，争夺客户资源，寻找优质客户，挖掘潜在客户成为企业竞争的潮流，但是传统的客户分类方法很难从大量数据中获取潜藏的价值和规律。因此，以用户行为分析为基础，以推荐系统为代表的数据挖掘算法成为客户细分的新工具。本文研究的基于 k-means 聚类的协同过滤推荐算法在用户挖掘和商品推荐方面具有良好的应用前景，能够有效的对用户进行细分，确定每个客户的消费特征与客户价值，满足用户的个性化需求，同时也能在很大程度上促进企业的发展。除却本文已经完成的部分，还存在以下几点需要更加深入的研究：

（1）本文主要基于用户行为信息来做商品的推荐，但没有考虑用户属性和商品属性等其他信息，比如用户性别、年龄、人生阶段，商品的类别、规格、价格等。相信如果能够将这些信息应用到推荐系统中，推荐算法的性能会有进一步的提高。

（2）用户的兴趣是会随着时间的改变而改变的，而且在电子商务系统中需要给用户进行实时的推荐，及时满足用户的多种需求，而本文采用的是离线行为数据来进行商品的推荐，对于实时更新的数据如何更好的去挖掘还有所欠缺，这也是后续研究需要改进的地方。

（3）在使用基于 k-means 聚类的协同过滤进行推荐时，各参数的取值都是在数据稀疏的情况下取得的，这是一个相对较优值而不是绝对最优值。因此，可以考虑在进行聚类之前先对稀疏矩阵进行填充，从最开始就可以考虑解决数据的稀疏性问题，然后再进行商品推荐。

（4）在利用 k-means 算法来对用户聚类时，聚类簇数和初始聚类中心的确定具有一定的随机性，在以后的研究中可以考虑到这一点。

参考文献

[1] Ah Keng Kau,Yingchan E. Tang,Sanjoy Ghose. Typology of online shoppers[J]. Journal of Consumer Marketing,2003,20(2).

[2] Angeline G. Close,Monika Kukar-Kinney. Beyond buying: Motivations behind consumers' online shopping cart use[J]. Journal of Business Research,2009,63(9).

[3] Agrawal R,Imielinski T,Swami A.Mining association rules between sets of items in large databases[C]//Proceeding of the ACM SIGMOD Conference on Management of Data,1993.

[4] Atzori L., Iera A., Morabito G. The internet of things: a survey[J]. Computer Networks,2010, 54(15): 2787-2805.

[5] Bing Chuan Long Huang,Yang Xiang,Zhen Hua Huang. Use Logistic Regression to Predict User' Behaviors[J]. Applied Mechanics and Materials,2014,3512.

[6] Breese J S,Heckerman D,Kadie C. Empirical analysis of Predictive Algorithms for collaborative filtering. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence,1998,43-52.

[7] Carlos A. Gomez-Uribe,Neil Hunt. The Netflix Recommender System[J]. ACM Transactions on Management Information Systems (TMIS),2015,6(4).

[8] Chen D E. The collaborative filtering recommendation algorithm based on BP neural networks[C]// 2009 International Symposium on Intelligent Ubiquitous Computing & Education. Piscataway: IEEE Press, 2009: 234-236.

[9] Dahlen B J, Konstan J A, Herlocker J L, Good N, etal. Jump-Starting Movielens: User Benefits of Starting a Collaborative Filtering System with "Dead Data". University of Minnesota, 1998.

[10] David Goldberg,David Nichols,Brian M. Oki,Douglas Terry. Using collabor ative filtering to weave an information tapestry[J]. Communications of the AC M,1992,35(12).

[11] Dirk Van den Poel,Wouter Buckinx. Predicting online-purchasing behaviour[J]. European Journal of Operational Research,2004,166(2).

[12] Donghui Wang,Yanchun Liang,Dong Xu,Xiaoyue Feng,Renchu Guan. A content-based recommender system for computer science publications[J]. Knowledge-Based Systems,2018,157.

[13] Giles C L,Bollacker K D,Lawrence S.Cite Seer:An Automatic Citation Indexing System[C].Proceedings of the ACM Conference on Digital Libraries. New York,

1998:89-98.
[14] Hariri, Waleed,Ghauth, Khairil Imran,Eswaran, C. A Multimedia Content Recommender System Using Table of Contents and Content-Based Filtering[J]. Advanced Science Letters,2018,24(2).
[15] Hilmi Yildirim,Mukkai S. Krishnamoorthy. A random walk method for alleviating the sparsity problem in collaborative filtering[P]. Recommender systems,2008.
[16] Kau A K, Ghose S, Tang Y E. Typology of online shoppers. Journal of Consumer Marketing, 2003, 20(2):139-156.
[17] Krishna K. Mohbey,G. S. Thakur. Interesting User Behaviour Prediction in Mobile Ecommerce Environment using Constraints[J]. IETE Technical Review,2015,32(1).
[18] MANJULA W,VIVIEN P,NAOMAL D,et al.Selecting a text similarity measure for a content based recommender system [J].The Electronic Library,2019,37(3):506-527.
[19] Maltz D,Ehrlich K.Pointing the Way:Active Collaborative Filtering[C]. Proceedings of the 1995 ACM Conference on Human Factors in Computing Systems. New York, 1995.
[20] Paul Resnick,Hal R. Varian. Recommender systems[J]. Communications of the ACM,1997,40(3).
[21] Peter D Hoff. Model Averaging and Dimension Selection for the Singular Value Decomposition[J]. Journal of the American Statistical Association,2007,102(478).
[22] Resnick P, Iakovou N, Sushak M, et al. GroupLens: An open architecture for collaborative filtering of netnews. Proc 1994 Computer Supported Cooperative Work Conf, Chapel Hill,1994: 175— 186.
[23] Reza Allahyari Soeini,and Keyvan Vahidy Rodpysh."Applying Data Mining to Insurance Customer Churn Management".Proceedings of 2012 International Conference on Network and Computer Science.Ed.. IACSIT Press, 2012, 85-95.
[24] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms.Proceedings of the 10th international conference on World Wide Web. ACM, 2001. 285-295.
[25] Sanchez Bocanegra Carlos Luis,Sevillano Ramos Jose Luis,Rizo Carlos,Civit Anton,Fernandez-Luque Luis. HealthRecSys: A semantic content-based recommender system to complement health videos.[J]. BMC medical informatics and decision making,2017,17(1).
[26] Stephen Russell,Victoria Yoon. Applications of wavelet data reduction in a recommender system[J]. Expert Systems With Applications,2007,34(4).

[27] Shardanand U, Maes P. Social information filtering: Algorithms for automation “Word of Mouth”[C]. Proceedings of ACM Conference on Human Factors in Computing System. New York: ACM Press, 1995: 210-217.

[28] Triyani Budyastuti,Diah Iskandar. The Influence of Usefulness, Easy of Use and Trust Using E-Commerce To User Behaviour (Case Study To Lazada.Com Consumers)[J]. Journal of Marketing and Consumer Research,2018,46(0).

[29] Wang T, Zhao L. Study on Mining Method for User Location Based on the RaSequence GPS Dataset [J]. Geomatics World, 2018.

[30] YAHYA A M,MD N S,NORWATI M,et al.Improved web page recommender system based on web usage mining [C]//The 3rd International Conference on Computing and Informatics (ICOCI).2011:8-9.

[31] Zhang Feng, Chang Huiyou. A collaborative filtering algorithm embedded BP network to ameliorate sparsity issue[C]// International Conference on Machine Learning & Cybernetics. Piscataway: IEEE Press, 2005: 1839-1844.

[32] Zhang Libo, Luo Tiejian, Zhang Fei, et al. A recommendation model based on deep neural network[J]. IEEE Access, 2018, 6: 9454-9463.

[33] Zhu Zhixuan. New Media User Behaviour Research Based on Big Data Analysis[J]. Journal of Physics: Conference Series,2021,1802(4).

[34] CNNIC 发布第 46 次《中国互联网络发展状况统计报告》.

[35] 毕闰芳.基于 SVR 的协同过滤与用户画像融合的电影个性化推荐研究[D].郑州大学,2018.

[36] 岑荣伟.基于用户行为分析的搜索引擎评价研究[D].清华大学,2010.

[37] 陈淑英,徐剑英,刘玉魏,山洁.关联规则应用下的高校图书馆图书推荐服务[J].图书馆论坛,2018,38(02):97-102.

[38] 陈正明.从“尿布和啤酒”到数据挖掘[J].软件工程师, 2005 (1): 59-59.

[39] 方耀宁,郭云飞,丁雪涛,兰巨龙.一种基于局部结构的改进奇异值分解推荐算法[J].电子与信息学报,2013,35(06):1284-1289.

[40] 傅鹤岗,王竹伟.对基于项目的协同过滤推荐系统的改进[J].重庆理工大学学报, 2010,000(009):P.69-74.

[41] 贾桂霞,赵锡英,刘熠琦.电子商务中关联推荐算法的应用研究[J].工业仪表与自动化装置,2016(1):43-45,共 3 页.

[42] 陆楠,王喆,周春光.基于 FP-tree 频集模式的 FP-Growth 算法对关联规则挖掘

的影响[J].吉林大学学报(理学版),2003(02):180-185.

[43] 黎丹雨,陈怡华.一种多层多维的关联规则挖掘算法在推荐系统中的应用[J].计算机与现代化,2019(06):44-48+54.

[44] 李勇,徐振宁'.Internet 个性化信息服务研究综述[J].计算机工程与应用.2002,38(019):183-188.

[45] 李志勇.基于大数据技术的移动用户行为分析研究[J].电脑知识与技术,2021,17(05):34-35+41.

[46] 雷名龙.基于阿里巴巴大数据的购物行为研究[J].物联网技术,2016,6(05):57-60.

[47] 任永功,石佳鑫,张志鹏.融合关系挖掘与协同过滤的物品冷启动推荐算法[J].模式识别与人工智能,2020,33(01):75-85.

[48] 王毅.网络推荐系统的三大挑战——从用户体验出发[J].清华管理评论,2013(06):10-13.

[49] 王志远,王兴芬.基于用户兴趣差异改进矩阵填充的个性化推荐算法[J].计算机应用与软件,2020,37(12):224-230+237.

[50] 魏琳东,黄永峰.融合用户属性信息的冷启动推荐算法[J].电子技术应用,2017,43(10):137-140+144.

[51] 巫可,战荫伟,李鹰.融合用户属性的隐语义模型推荐算法[J].计算机工程,2016,42(12):171-175.

[52] 夏萍萍. 联合时间和地理因素的兴趣点推荐研究与应用[D].大连理工大学,2020.

[53] 袁兴福,张鹏翼,王军.电商用户“状态–行为”建模及其在商品信息搜索行为分析的应用[J].现代图书情报技术,2015(06):93-100.

[54] 于泽川. 基于大数据的用户精准定位与行为分析[D].北京邮电大学,2019.

[55] 余胜辉. 层次聚类算法基于 Spark 的实现及在推荐系统中的应用[D].南京邮电大学,2020.

[56] 张晓彬.基于可信度的协同过滤推荐算法研究[D].重庆大学,2010.

[57] 朱珏樟.客户购买行为建模分析预测[J].现代计算机,2020(21):27-32.

[58] 张玉芳,代金龙,熊忠阳.分步填充缓解数据稀疏性的协同过滤算法[J].计算机应用研究,2013,30(09):2602-2605.

[59] 赵洪英,蔡乐才,李先杰.关联规则挖掘的 Apriori 算法综述[J].四川理工学院学报(自然科学版),2011,24(01):66-70.

[60] 李昌盛,伍之昂,张璐,曹杰.关联规则推荐的高效分布式计算框架[J].计算机学报,2019,42(06):1218-1231.

[61] 周春华,沈建京,李艳,等.经典推荐算法研究综述[J].计算机科学与用,2019,9(9):1803-1817.

[62] 朱郁筱,吕琳媛.推荐系统评价指标综述[J].电子科技大学学报,2012,41(02):163-175.

后 记

行文至此，意味着我三年研究生学习生活即将落幕，总以为来日方长，不惜岁月，但时间流逝总让人猝不及防。回想在兰财学习和生活的点点滴滴，心中感触万千、五味杂陈，在这青春的校园里有过困顿、有过失落、有过迷茫，但也有过坚定、有过认可、有过拼搏，此时更多的是不舍。三年来的所见所闻、所学所想得离不开每一位可爱的身边人，在此，我要向所有给予我指点、帮助、包容和关爱的老师、家人、同学、朋友认真地道一声感谢。

首先，我要特别感谢我的导师庞智强教授，三年来我的成长与进步离不来导师的精心指导与教诲，他是位严师，也如慈父般指引我做人、教我做事。这篇论文从选题、主体框架的敲定、修改到最后定稿离不来老师的悉心指导，写作过程中老师提供了很多建设性意见，让我在遇到困难时瞬间豁然开朗。另外，庞老师严谨求实的治学态度、踏实努力的工作作风将使我受益终生。庞老师不仅是我硕士求学路上的老师,也将永远是我的人生恩师。

其次，我要感谢各位优秀的舍友，感谢你们无微不至的照顾，是大家的包容、陪伴与付出让我们三年的寝室生活融洽、温暖而欢乐，你们有趣的灵魂与积极向上的心态也将影响我未来的生活，祝福你们考博顺利、工作顺心、事事如意。感谢同窗好友们在生活和学习中给予的帮助，你们的出现让我的研究生生活变的更加丰富多彩。

最后，我要感谢我的父母，无论我得意或者失意，你们都无条件相信我，做我最坚强的后盾，是你们多年来的关心、理解与鼓励支撑我走到今天，让我可以无忧无虑的在学校完成学业。养育之恩，无以为报，只能全力以赴，带着这份爱继续奋力前行，成为你们永远的骄傲。祝愿我的父母身体健康，平安顺遂。

前路漫漫，我们即将飞往人生中的下一个起点，未来也将面临更多的困难与挑战，我定不忘初心，勇敢前行，全力以赴迎接生活中的每个惊喜，不断学习，成为更好的自己！