北方工业大学

硕 士 研 究 生 学 位 论 文

用户画像在推荐系统中的研究与应用

学　生　姓　名　孙吉祥

学　　号　2017311040102

学科(专业学位)　软件工程-工程硕士

研　究　方　向　软件服务工程

导　　师　方英兰

校　外　导　师　王新民

2020 年 6 月 22 日

Research and Application of User Profile in Recommendation System

By

Sun Jixiang

A Dissertation Submitted to

North China University of Technology

In partial fulfillment of the requirement

For the professional degree of

Master of Engineering

North China University of Technology

June, 2020

用户画像在推荐系统中的研究与应用

摘　要

随着互联网的飞速发展，网络购物成为了新常态。为了让用户快速找到所需商品，同时使得卖家快速定位对自己商品感兴趣的人群，推荐系统成为了优秀的解决方案之一。同时，用户画像作为用户信息全貌的抽象，能够更方便地对用户进行快速且全面的分析。本文研究了电商网站用户画像的构建技术，并对用户画像技术中的标签建立和标签权重的调优进行了研究，设计了基于BERT的文本情感分析模型对用户画像技术中非量化的用户信息进行分析，并结合兴趣遗忘曲线对用户画像进行调优。

在用户画像的构建研究中，对于非量化信息的研究相对较少。但有时非量化信息能够更好地反映用户的真实喜好，如用户评论信息。用户的评论信息主要以短文本的形式存在。因此，本文设计了基于BERT的文本情感分析方法对用户的评论信息进行分析，得到用户对于此商品的真实情感，并根据分析结果对用户画像进行了初步调优。

以往的研究大都忽略了用户的兴趣爱好会随时间发生改变，因此，本文结合记忆遗忘曲线设计了兴趣遗忘曲线，并根据该曲线对用户画像进行了二次调优，使得用户画像更加精确。

最后本文利用某电商网站的数据场景和上述方法，设计了用户画像的构建方案，同时开发了对应的推荐系统。该系统实现了用户画像的构建，同时结合基于BERT的文本情感分析模型和兴趣遗忘曲线对用户画像进行了调优。根据调优后的用户画像为用户提供了个性化推荐的功能并通过推荐结果验证了用户画像的合理性。

关键词：用户画像，情感分析，协同过滤，个性化推荐

Research and Application of User Profile in Recommendation System

Abstract

With the rapid development of the Internet, online shopping has become the new normal. In order to allow users to quickly find the required products, and at the same time make sellers quickly locate people who are interested in their products, the recommendation system has become one of the excellent solutions. At the same time, the user profile as an abstraction of the overall picture of the user's information, can more quickly and comprehensively analyze the user.This paper studies the construction technology of user profile on e-commerce websites, and focuses on the establishment of labels and the optimization of label weights in user profile technology. A text sentiment analysis model based on BERT is designed for non-quantitative users in user profile technology. Analyze the information, and adjust the user profile in combination with the interest-forgetting curve.

In the research on the construction of user profiles, there is relatively little research on non-quantitative information. But sometimes non-quantitative information can better reflect the user's true preferences, such as user comment information. The user's comment information mainly exists in the form of short text. Therefore, this paper designs a text sentiment analysis method based on BERT to analyze the user's comment information, get the user's true sentiment for this product, and preliminary tune the user's profile based on the analysis results.

Most previous studies have overlooked that the user's interests and hobbies will change with time. Therefore, this paper designs the interest-forgetting curve in conjunction with the memory-forgetting curve, and tunes the user's profile based on the curve to make the user's profile more accurate.

Finally, this paper uses the data scene of an e-commerce website and the above methods to design a construction scheme for user profiles, and develops a corresponding recommendation system. The system realizes the construction of user

profiles, and at the same time combines the BERT-based text sentiment analysis model and the interest-forgetting curve to optimize the user profiles. According to the tuned user profile, it provides users with personalized recommendation function and verifies the rationality of the user profile through the recommendation results.

Key words: user profile， sentiment analysis， collaborative filtering， personalized recommendation

目　录

第一章　绪论

1.1 研究背景及意义

随着互联网的飞速发展，网络购物成为了新常态。出于便捷性，越来越多的人开始适应了网络购物，各大电商网站也获得了飞速地发展。但在电商网站发展的同时，商品的规模与购物群体也越来越庞大。如何在海量的商品中找到自己需要的商品，变成了用户所面临的问题，而商家所面临的问题则是如何找到对自己的商品感兴趣的人群。对于像网络购物这样的海量数据的处理问题，比较成熟的方案有两种，一种是搜索引擎，一种为推荐系统。搜索引擎提供了用户主动从质量参差不齐的信息堆获取有用信息的途径，而推荐系统则通过分析用户习性，了解用户的喜好，为用户提供个性化推荐。

为了更好地抽象表征用户个体，用户画像技术应运而生。用户画像从提出至今，获得了高速地发展，并成为了研究热点[1]。从简单的静态标签发展为了成熟的抽象的用户模型。用户画像技术已经成为了各大电商网站运营的基础。用户画像技术的应用场景十分多样化，各大电商网站都会为用户构建画像模型。利用这个模型提高用户的购物满意度，同时不断改善模型，使之达到更好的效果，让企业获得更好的发展。

近年来，用户画像与推荐系统一直是学术界的热门研究内容之一[2]，二者都具有较高的研究价值与商用价值，共同推进着网络购物的发展，推进着互联网的发展，因此，本文将精确用户画像应用到推荐系统之中，用于改善用户网络购物的体验，提高用户网络购物的满意度。

1.2 国内外研究现状

1.2.1 用户画像构建研究

国外学者关于用户画像的研究比较早，也有了一些比较成熟的思想。

Zhiliang[3]等提出了一个从多个角度描述用户的偏好的用户画像模型，分析了用户对历史新闻的偏好程度，并提出一种根据用户的阅读行为和新闻流行度来计算历史新闻的偏好权重的方法。 此方法可以更准确地构造用户画像。此外，提供了一种动态的新闻推荐方法，其中考虑了短期和长期用户偏好。 实验结果表明，此方法可以显著提高推荐效果。

Chader[4]等以社交共享内容作为丰富用户画像的来源，强调了考虑用户的关系强度以导出相关社交画像的重要性。专注于以自我为中心的基于网络社区的用户画像过程（CoBSP），该过程仅考虑人与人之间的二元连接。认为与用户关系最密切的人可能会透露有关他的更有价值的信息。实验证明了此相关性。

Nguyen Phong[5]等提出了一种可视化分析方法，以帮助分析师全面了解多个级别（即个人和小组级别）的用户行为。采用以用户为中心的方法来设计视觉分析框架，对用户集合及其在数字应用程序中进行的众多活动的分析。围绕分层用户画像的概念建立框架，该分层用户画像是基于从会话派生的功能以及使用主题建模方法提取的用户任务来汇总和分层用户行为的基础上构建的。通过评估，借助交互式可视化分层用户画像，分析师能够有效地进行探索性调查研究，并能够了解用户行为的特征，从而在评估可疑用户和活动的同时做出明智的决定。

国内的相关用户画像研究起步较晚，但也有了许多不错的研究成果。

过仕明[6]以群体为单位对用户进行了行为分析，用于构建数字图书馆用户的用户画像，并结合场景理论构建了数字图书馆场景重构模型。实验证明此方法能够反映不同群体用户的行为特征，挖掘出不同群体用户的潜在服务特征，能够较为准确地刻画用户画像，为用户提供比较准确的推荐，提升了用户使用数字图书馆的满意度。康存辉[7]分析了虚拟图书馆空间的再造必要性，提出强化四大空间再造，并智慧生成用户画像，满足读者需求。

谭浩[8]等以大数据环境为研究基础，为产品研发初期获取准确性较高的用户画像，优化了构建方法。采用三层次理论框架，首先研究了多模态的用户文本数据，提取了高影响力的用户属性因子。其次，使用 FCM 算法对用户属性因子进行聚类分析，获得用户属性原型类别。通过对样本用户进行调研，挖掘深层次需求动机。最终，构建出几类用户画像原型。实验通过研究从社交网络和电商网站等获取评论信息，最终分类取得了不错的效果。张长浩[9]等依托商旅大数据背景，通过分析用户的旅游决策，构建了一种能够快速识别敏感用户群体的用户画像。首先提取用户特征，然后采用双通道建模预测用户敏感程度，根据旅游信息提取了用户多个维度的特征，并采用了双层 XGBoost 的多视角融合模型，实验结果证明采用该模型后，分类的准确率有所提升。

单晓红[10]等从酒店平台获取数据，并对获取到的酒店数据和用户数据进行了分析，基于分析结果构建了基于酒店信息、用户信息和评论信息这三个维度的用户画像模型。在此模型的基础上使用 Protégé工具实现维度关联，并构建用户画像。提供酒店了解用户需求和开展精准营销的途径。

陈泽宇[11]等针对用户的历史查询词，使用 LDA 主题模型为每个查询词分配主题，使查询词和其主题共同放入神经网络模型中学习得到其主题词向量，最后

采用随机森林分类算法对用户基本属性进行分类构建用户画像。实验结果表明，该模型的分类精度要高于词向量模型。安璐[12]等，使用 LDA 主题模型提取恐怖事件情况下用户微博信息和评论信息，采用两步聚类刻画用户特征，加强反恐舆论引导。

标签广泛应用于社交系统中，但由于标签一般由人为定义，缺少理论支撑，因此极有可能引发标签歧义的问题，Xu[13]等提出了一种本体相似性，此方法在语义上比最新的相似度度量更准确，无需使用上下文信息进行模型训练即可解决标签歧义性问题。这种本体相似性的新颖性在于，它首先利用外部领域本体来消除标签信息的歧义，然后根据各个配置文件中标签匹配概念的语义相似性在语义上量化用户和项目配置文件之间的相关性。

现有的文献对于电商用户画像的构建研究是比较多的，针对不同领域的用户画像构建的研究是相对全面的，构建方法也是多种多样的。但还没有针对评论数据进行情感分析调整用户画像的文献。

1.2.2 个性化推荐算法研究

常见的推荐主要有四种：基于深度学习的推荐、基于树结构的推荐、基于图结构的推荐和基于内存的协同过滤的推荐。

国外的个性化推荐的研究比较成熟，涉及的领域比较广泛。

快速增长的科学论文提出了快速而准确地找到给定手稿的参考论文列表的问题。引用推荐是克服这一障碍的必不可少的技术。Cai[14]等提出了一种通过在三层图上相互增强的引文推荐方法，将每篇论文，作者或场所分别表示为纸层，作者层和场所层中的一个顶点。为了解决计算复杂度过高的问题，将三层交互式聚类方法应用于图中的相关顶点的聚类。然后在子图中提出个性化的引用推荐。提出的基于模型的引文推荐方法的性能可与其他最新的引文推荐方法比肩，同时还说明了个性化推荐方法比非个性化推荐方法更有效。

在线新闻推荐即时性要求较高，传统的用户资料进行静态表示的用户建模方法不能很好地处理此问题，为了解决这个问题，He[15]等提出了一个基于用户配置树（UP-Tree）的基于内容和协作过滤技术的新颖的新闻推荐框架。通过利用一种新颖的主题模型 UI-LDA，获得主题空间中新闻内容的表示向量，作为将用户兴趣与新闻主题相关联的基本桥梁。设计具有动态可变结构的决策树，根据用户的反馈来构建用户兴趣配置文件。提出了一种基于聚类的多维相似度计算方法，以有效地选择 UP-Tree 的最近邻居。提供了一个基于 Map-Reduce 框架的实现，将解决方案扩展到现实世界中的新闻推荐问题。该方法使得推荐的准确性有所提升。

鉴于社交图像的多样性和复杂性，通过从大规模社交图像中学习用户的兴趣来提高个性化推荐的性能具有重要意义。Zhang[16]等通过构建具有深层特征的用户兴趣树和标签树，提出了一种个性化的社会形象推荐方法。首先，通过重新排列标签来创建社交图像的标签树从而有效地利用标签。其次，通过训练 AlexNet 网络来学习深度功能紧凑地表示图像内容。然后，构建具有深层特征的用户兴趣树和标签树，最后，基于该功能构建个性化的社交图像推荐系统用户兴趣树。这个方法在准确性和个性化推荐的召回率方面均优于最新方法。

个性化推荐已显示出其在改善 Internet 信息过载问题方面的有效性。但是，由于个人隐私问题，用户不愿透露自己的个人信息已成为发展个性化推荐的主要障碍。Wu[17]等生成一组伪造的偏好配置文件，以掩盖用户敏感主题，从而在个性化推荐中保护用户的个人隐私。提出了一个基于客户端的用户隐私保护框架，该框架不仅不需要更改现有推荐算法，还不需要牺牲推荐准确性。其次，在该框架的基础上，引入了一个隐私保护模型，该模型阐述了理想的伪造的偏好配置文件应满足的两个要求：（1）特征分布的相似性，用于衡量伪造的偏好配置文件隐藏真实用户偏好的有效性；（2）敏感对象的暴露程度，用于衡量伪造的偏好配置文件掩盖敏感对象的有效性。最后，基于产品分类的主题库，提出了一种可以很好地满足隐私保护模型的实现算法。该方法于理论和实际上都取得了不错的效果。

国内的个性化推荐研究也有不错的进展。

传统的图书推荐系统普遍存在推荐与用户兴趣匹配度偏低和用户满意度较低等问题，张兰兰[18]引入关联数据技术，设计了一种新的智能图书个性化推荐系统，能够精准地匹配用户需求和图书信息。

传统的个性化推荐算法忽略了网络中的热门项目、用户间针对相同项目的评分等问题，邹洋[19]等提出了一种基于多权重相似度的随机漫步推荐算法。该算法分别采用万有引力公式和改进的协同过滤算法计算了相似度，然后将二者采用不同权重融合后结合随机漫步算法进行商品推荐。该算法比其他推荐算法具有更高的性能。

目前的会话型推荐主要考虑用户的短期和长期的兴趣，王雅青[20]等引入中期兴趣的概念，提出一个基于循环神经网络的个性化分层循环模型，通过在同一框架下联合利用用户的会话、区块和全部行为序列来学习用户的综合兴趣，更精准地获取了用户的行为模式，提升了推荐准确率。

以用户标签为基础的个性化推荐，标签或者资源的热门程度的不一致导致推荐的准确性存在问题，汪涛[21]等提出了一种融合时间权重的张量分解模型，对标签生成时间进行建模并计算得到权重，再将权重融入张量分解模型，最后利用分

解后的特征向量进行推荐。算法在准确率-召回率和 F1 指标上均高于其他流行标签推荐模型。刘晓飞[22]等结合用户行为分布提出基于用户行为特征挖掘的个性化推荐算法。根据用户行为构建特征组结合模糊信息感知方法构建社交网络用户偏好特征的混合推荐模型，根据语义分布和用户的行为偏好实现社交网络的个性化信息推荐。

传统的推荐系统分析了用户与物品之间的联系而忽略了现实物品与历史物品之间的关系，邓凯[23]等将一个更有表现力的 Top-N 推荐系统的物品相似性因子模型解决方法与多层感知机方法相结合，建模物品之间的高阶关系，捕获更复杂的用户决策。该方法对比传统的基准方法，推荐性能方面有明显的提升。

现有的个性化推荐的方法遍布领域广泛，相同领域的个性化推荐方法也很丰富，但还没有用户画像建模用户倾向及兴趣爱好的个性化推荐方法。

1.3 课题主要研究内容

1.3.1 研究目标

随着互联网的飞速发展，新技术不断涌现，用户如何在质量参差不齐的商品堆中找到适合自己的商品成为了大数据背景下的热门研究问题。因此，本文针对电商网站的特点，首先分析电商网站用户的基本信息以及行为信息，并在此基础上建模基本的用户画像，描述用户的基本兴趣爱好及倾向。然后采用本文方法对用户画像中的权重进行调优，生成更加合理的用户画像，最后根据调优后的用户画像采用本文的推荐算法对用户进行个性化推荐。

1.3.2 研究内容

本文所研究的内容分为以下四个部分进行：

首先，根据用户在电商网站的基本信息数据以及行为数据对用户进行用户画像的初步建模；

其次，对用户的评论数据进行情感分析，得到用户的真实情感，随后根据喜好情况对用户画像的标签权重进行调整，初步精准化用户画像；

然后，考虑到用户的兴趣会随时间会存在一定的周期性变化，兴趣遗忘曲线反映了兴趣随时间的演变情况，因此引入兴趣遗忘曲线再次对标签权重进行调整，进一步精准化用户画像；

最后，根据精准的用户画像，采用协同过滤算法，根据用户画像计算用户相似度，根据相似度选取邻近用户对用户进行个性化推荐。

1.4 论文组织架构

本文由六个章节构成，各个章节的内容如下：

第一章为绪论。描述了课题的研究背景、意义、研究现状和主要研究内容；

第二章为相关理论与技术研究。本章首先对用户画像的研究作综述性的描述，对比了用户画像的各种方法以及分析了不同方法的特点，然后介绍了个性化推荐的常用方法及特点，同时介绍了不同方法针对不同场景的应用，最后介绍了常用的评价指标；

第三章为用户画像构建设计。本章介绍了基于电商网站数据对用户画像进行建模的过程以及对评论文本进行情感分析和引入兴趣遗忘曲线对用户画像进行调整，得到精准的用户画像；

第四章为个性化推荐算法设计。本章介绍了基于调优用户画像的协同过滤，采用用户画像进行计算，一定程度上降低了计算所需时间。根据用户画像计算用户相似度，并依据最邻近用户对当前用户进行个性化推荐；

第五章为推荐系统的设计与实现。本章介绍了商品后台推荐系统的设计与实现以及推荐结果的展示与分析；

第六章为总结与展望。本章是对论文的研究工作和成果的总结，分析了当前研究可能存在的问题，并对下一步研究提出了研究方向与内容。

1.5 本章小结

本章根据课题的研究背景及意义，介绍了用户画像和个性化推荐的国内外研究现状，并引出了本文的研究目标和主要研究内容。

第二章　相关理论与技术研究

2.1 用户画像综述

2.1.1 用户画像概念

用户画像的概念由 Alan Cooper[24]于 1995 年首次提出，他在书中强调了“以人为中心”的思想：角色用户源于真实的用户数据，是真实用户的抽象，服务于用户。王宪朋[25]将用户画像定义进行了分解，主要包括三点：一是如何有效地获取用户数据；二是用户画像需要与特定的业务场景相结合，针对不同的业务场景，需要体现出不同的特色；三是通过数学模型去挖掘那些没有明显显现的信息。余孟杰[26]将用户画像定义为标签化的信息产物。综上，首先，用户画像源于真实的用户数据，是真实数据的抽象形态，是具有相似信息或行为的用户在使用系统或者接受服务时所呈现出的具有共同特征的集合。其次，用户画像更加注重的是由静态和动态信息提取后得到的 “代表用户”，是具有某些明显且相似特征的用户的抽象模型。最后，用户画像更加注重用户的主导地位，更能突出用户的个性化需求。

2.1.2 用户画像构成要素

曾建勋[27]提出从用户的个人信息及习惯中提炼用户画像标签。通过分析这些方面的信息，可以更清楚地了解所服务的用户群体，以便提供更好的用户体验并为个性化营销提供数据支持。李映坤[28]将构成扩展到行为属性和信用属性构建用户画像。同时将营销中的 FRM 指标结合聚类方法进行用户行为分析，更好的刻画了用户的行为特征。刘海鸥[29]等考虑情景要素，将用户画像分为自然、兴趣、社交和能力四个维度。用户画像由用户属性、用户特征和用户标签三个要素构成[30]。其中，用户属性又由两部分构成：用户基本信息和行为信息。以电子商务系统为例，其中，前者一般由用户在注册账号时填写；后者为用户在使用系统时留下的操作、浏览痕迹。而用户特征则为通过聚类等分类方法从用户属性中提取出来的一些共性的特征。在用户特征的基础上，通过精确地分析计算得到标签化的文本，可以精确地表达用户的用户的特征。最终的用户画像模型本质上为标签化文本与标签权重的组合，是用户全貌的标签化展现形式。构建用户画像的过程其实就是信息提炼的过程。

同时，用户画像还应满足三大特征：标签化、时效性和动态性[30]。用户画像由用户的属性挖掘得到，本质是标签与权重的组合。所以用户画像首先需要满足标签化的特征。其次，用户的行为信息时时刻刻都在更新，这使得一次分析计算得到的用户画像随着时间的推移，不能够再准确地体现用户的兴趣爱好，所以时效性也是用户画像需要满足的特征。最后，随着用户数据的不断变化，用户的标签也时时刻刻随之改变，所以构建的用户画像需要不断更新才能够准确地表达用户的特征。

2.1.3 用户画像构建流程

用户画像的构建由三个步骤组成：数据采集、数据处理、分析建模。构建用户画像，首先，一般使用网络爬虫技术或者采取直接从数据库后台获取用户基本数据与用户行为数据，将数据保存起来供下一步研究使用；然后，对保存的数据进行数据清洗、空值填缺等操作后处理为需要的数据格式；最后，对处理后的数据应用深度学习或者机器学习的方法进行分析、计算，获得标签及其权重，得到用户画像。

2.1.4 用户画像构建方法

目前常用的用户画像的构建方法有六种[31]，通常基于六种分析方法：设计与思维、本体或概念、主题或话题、兴趣或偏好、行为或日志、多维或融合。

基于设计与思维的方法通常需要调查用户习性，根据调查的结果得到用户的共同点与差异，然后在这些特征的基础上进行分析与设计，得到不同的用户画像。

基于本体或概念的方法利用已经定义的信息，对用户进行刻画，得到用户画像。这些信息主要以结构化的形式存在于本体或者是概念之中。其中，本体是一种强大的知识表示、推理方法。牛温佳[32]等利用本体对用户画像中的标签进行刻画。

基于主题或话题的方法基于用户行为信息的文本形式，这些信息由不同出现概率的一系列相关的词语构成，通常表示不同的主题或者话题。通过对用户文本信息进行分析，挖掘出隐藏的主题或者话题，然后依据所挖掘的信息刻画用户画像。曾鸿[33]等分析对同一话题感兴趣的人群的信息，并通过算法提取标签，对微博粉丝人群进行用户画像。郭光明[34]结合了 LDA 建模和特征提取的优势，提出隐行为模式用户信用画像。该画像基于社交数据，其中的社交数据分为用户的基本信息数据和用户的行为信息数据。该方法适用于大多数社交网络。

基于兴趣或偏好的方法根据用户参与度较高的信息对用户进行刻画得到用

户画像。用户的行为跟兴趣息息相关，行为很大程度上收到兴趣的支配。张小可[35]等使用贝叶斯网络实时获取并分析用户的数据，并进行了实时推荐。

基于行为或日志的方法对用户的活动历史数据进行分析，刻画用户，得到用户画像。用户行为数据及日志信息反映这用户的真正的需求，通过分析这些数据能够得到用户的兴趣，对用户画像进行很好的补充。汪强兵[36]等通过分析用户在移动设备上阅读文献产生的手势及手势所对应的文本及时间等对用户进行建模，为个性化营销提供了渠道。

基于多维融合的方法从多个维度分别分析用户特性，将分析后的结果结合起来刻画用户画像。

各构建方法的特点见表 2-1 ：

表 2-1　用户画像构建方法比较表

构建方法	特点	局限性
基于设计与思维	1.通过问卷等形式获得用户的真实需求，具体化用户特征； 2.维度通常可解释。	1.无法确定刻画的用户是否存在； 2.画像的实际效果基本上完全由设计人员的专业性决定。
基于本体或概念	1.推理过程容易解释。	1.各领域均需要单独分析； 2.需要及其专业的领域知识； 3.耗时耗力。
基于主题或话题	1.只考虑词语出现次数，与位置无关； 2.可以捕捉隐含的信息。	1.模型中无法引入用户本身特征； 2.分析情感时可能存在偏差。
基于兴趣或偏好	1.利用了不同用户的相似偏好； 2.无关词影响被弱化。	1.实时性差； 2.只能从历史数据中的到用户的兴趣偏好。
基于行为或日志	1.充分利用用户的个性化信息； 2.充分利用不同用户的相似行为。	1.对非理性因素不能很好处理； 2.不能很好的解决数据稀疏性
基于多维或融合	1.综合考虑多方面用户特	1.模型计算量大且复杂；

	征； 2.从多角度分析信息间的关系。	2.维度太大。

通过分析以上六种构建方法，考虑到本文依据用户的基本信息和购物信息刻画用户画像，主要分析用户的历史购物数据，最终得到用户画像。因此，对比六种方法的利弊，本文决定采用基于行为或日志的画像构建方法进行用户画像的构建更为合适。

2.2 推荐系统及相关技术

2.2.1 个性化推荐技术

推荐系统是信息过载时信息检索的解决方案之一。现实生活中，人们在做出选择时，大脑潜意识会进行信息的检索，这些信息通常是朋友间闲聊时透露的意见或观点、阅读过的文档、别人对事物的评价与评论、专家意见等。而推荐系统则模拟了这一过程，由用户潜意识自发地去检索这些信息转变为推荐系统主动对用户发起推荐。因此，推荐系统要求用户提供自己的个人信息、行为信息，通过对信息进行分析，并对用户进行建模，然后根据一定的算法在已经建立的用户模型的基础上计算得到返回给目标用户的推荐结果，实现个性化推荐。这也可看作项目与目标用户的一种匹配关系。

推荐系统由三大模块组成：输入、推荐算法和输出。如图 2-1 所示：

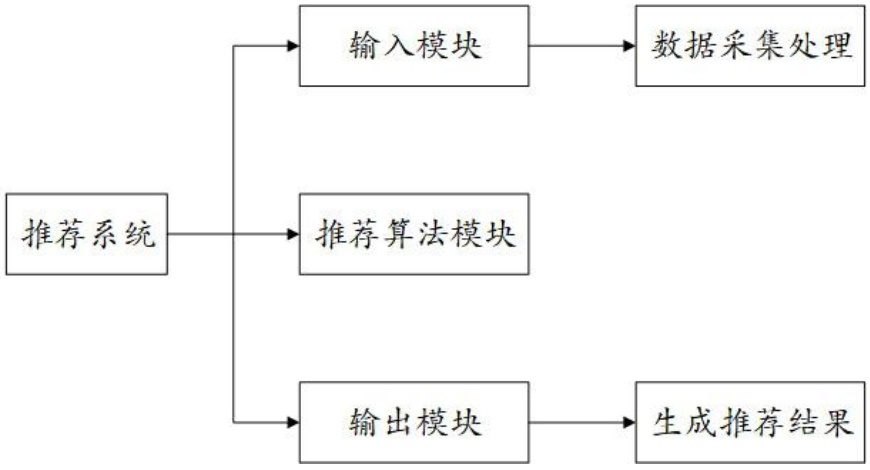

图 2-1　推荐系统功能模块图

推荐系统由三大模块组成：输入、推荐算法和输出。如图 2-1 所示：

（1）输入模块：使用合理的方法收集用户的数据，包括基本信息和行为数

据。其中，行为数据通常为历史行为数据。采集的数据经过如空值处理、去无效值等操作后进行保存。

（2）推荐算法模块：作为推荐系统最为重要的模块，推荐算法的选取决定了最终推荐系统的整体好坏。基于项目和用户得到目标用户的推荐结果。Adomavicius[37]等将推荐算法定义如下：U 表示用户集合，I 表示项目集合，R 表示全序集合，即全部排序推荐列表，使用效用函数 $r（u，r）：U\times I\rightarrow R$ 计算项目 i 值得向用户 u 推荐的程度。推荐算法所要研究的主要问题就是对任一用户 u∈U，找到使推荐度 r 最大的项目 i∈I，如式（2-1）所示：

$$\forall u \in U, i'_u = \arg\max_{i \in I} r(u,r) \qquad (2\text{-}1)$$

（3）输出模块：将推荐结果列表展示给用户。

2.2.2 推荐算法分类

推荐算法大致分为三类：基于内容的推荐、协同过滤推荐和混合推荐。

（1）基于内容的推荐是最早采用的推荐算法，它所依据的是用户的历史信息。通过分析用户的历史信息，从历史项目中提取特征得到用户的偏好，并保存下来，然后进一步计算项目已保存数据之间的相似度，最后将相似度最高的项目推荐给对应的用户。基于内容的推荐过程如图 2-2：

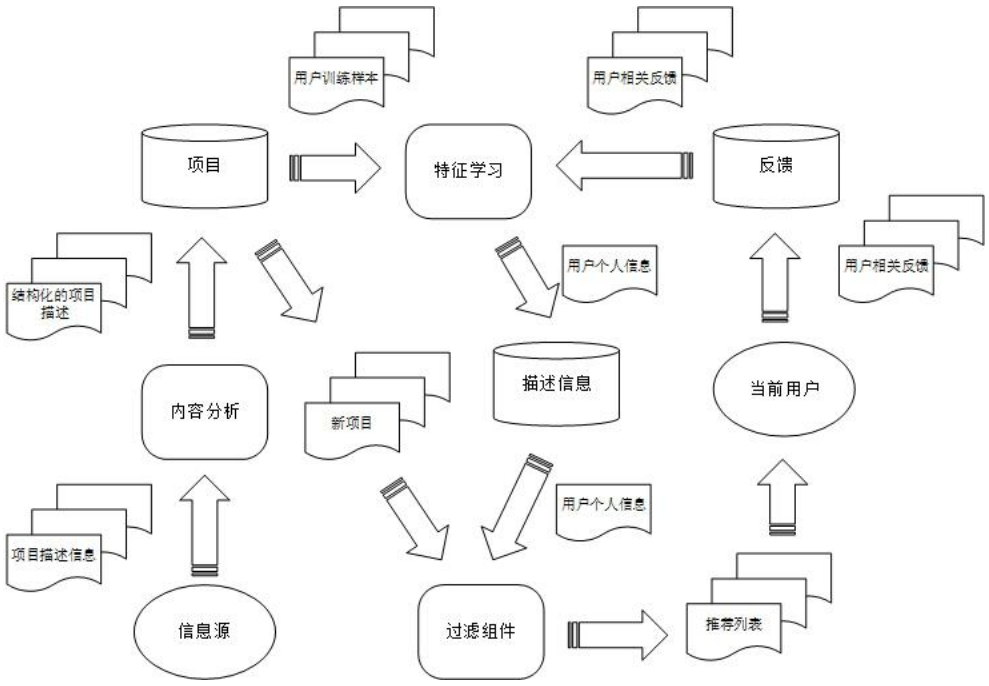

图 2-2　基于内容的推荐步骤

内容分析：分析得到项目特征，并将数据结构化。

特征学习：把用户的历史数据进行预处理后分析得到用户的偏好特征。

过滤组件：根据项目特征和用户偏好特征计算相似度得到相关性最大的一组推荐列表。

其中，部分算法可根据用户反馈不断调节优化算法。

基于内容的推荐对于新项目也可以进行推荐，但缺点之一是不能发掘潜在兴趣，之二是当信息源比较复杂时预处理会比较困难，如多媒体数据（音频、视频等）。

（2）协同过滤推荐的比较重要且核心的环节是相似度计算。目前的相似度计算方法通常分为两大类：基于用户和基于项目。二者核心思想的区别是去计算用户的相似度还是计算项目的相似度。但最终，都会根据用户的历史数据对当前项目进行打分，项目得分降序排列后取得分最高的集合得到最终的推荐列表。

基于用户的协同过滤分五步进行：

①根据历史数据中用户对项目的喜好程度，按照一定规则构造用户-项目评分矩阵如表 2-2：

表 2-2　用户-项目评分矩阵表

	I_1	…	I_j	…	I_n
U_1	$R_{1,1}$	…	$R_{1,j}$	…	$R_{1,n}$
…	…	…	…	…	…
U_i	$R_{i,1}$	…	$R_{i,j}$	…	$R_{i,n}$
…	…	…	…	…	…
U_m	$R_{m,1}$	…	$R_{m,j}$	…	$R_{m,n}$

表中行代表用户数，共 m 个；列代表项目数，共 n 个；$R_{i,j}$代表用户 i 对项目 j 的评分。

②计算用户相似度 sim（u,v）。通常采用的方法有余弦相似度、修正的余弦相似度和皮尔逊相关系数。

用户余弦相似度将每一项的评分当作维度，于是用户可看作 n 维向量，然后计算向量夹角，值越大则用户越相似，用式（2-2）计算：

$$sim(u,v)=\frac{\vec{u}\times\vec{v}}{\|\vec{u}\|\times\|\vec{v}\|}=\frac{\sum_{i\in I_{u,v}}R_{u,i}R_{v,i}}{\sqrt{\sum_{i\in I_u}R_{u,i}{}^2}\sqrt{\sum_{i\in I_v}R_{v,i}{}^2}} \qquad (2\text{-}2)$$

其中 $R_{u,i}$表示用户 u 对项目 i 的评分，$I_{u,v}$表示用户 u 和用户 v 的评分交集，I_u表示用户 u 的评分集。

修正的余弦相似度则充分考虑了用户的习惯，将分值减去平均值弱化因用户习惯不同引起的打分范围不同导致的影响，可用式（2-3）计算：

$$sim(u,v)=\frac{\sum_{i\in I_{u,v}}\left(R_{u,i}-\overline{R}_u\right)\left(R_{v,i}-\overline{R}_v\right)}{\sqrt{\sum_{i\in I_u}\left(R_{u,i}-\overline{R}_u\right)^2}\sqrt{\sum_{i\in I_v}\left(R_{v,i}-\overline{R}_v\right)^2}} \quad (2\text{-}3)$$

其中 $\overline{R}_u$ 表示用户 u 的评分集均值，$\overline{R}_v$ 表示用户 v 的评分集均值。

皮尔逊相关系数则只计算不同用户对相同项目有打分的情况，用式（2-4）计算：

$$sim(u,v)=\frac{\sum_{i\in I_{u,v}}\left(R_{u,i}-\overline{R}_u\right)\left(R_{v,i}-\overline{R}_v\right)}{\sqrt{\sum_{i\in I_{u,v}}\left(R_{u,i}-\overline{R}_u\right)^2}\sqrt{\sum_{i\in I_{u,v}}\left(R_{v,i}-\overline{R}_v\right)^2}} \quad (2\text{-}4)$$

③按照相似度由高到低排序选择最邻近的集合 N_u。

④根据近邻用户的评分对当前用户的无评分项目用式（2-5）进行评分预测。

$$P(u,i)=\overline{R}_u+\frac{\sum_{v\in N_u}sim(u,v)\times\left(R_{v,i}-\overline{R}_v\right)}{\sum_{v\in N_u}|sim(u,v)|} \quad (2\text{-}5)$$

其中 $P(u,i)$ 表示用户 u 的未评分项目 i 的预测评分。

⑤按照预测评分值选取前 N 项作为推荐结果。

基于项目的协同过滤：原理类似基于用户的协同过滤，相似度计算一般采用式（2-6），计算两项目之间的相似度。

$$sim(i,j)=\frac{\sum_{u\in U}\left(R_{u,i}-\overline{R}_i\right)\left(R_{u,j}-\overline{R}_j\right)}{\sqrt{\sum_{u\in U}\left(R_{u,i}-\overline{R}_i\right)^2}\sqrt{\sum_{u\in U}\left(R_{u,j}-\overline{R}_j\right)^2}} \quad (2\text{-}6)$$

其中，$R_{u,i}$ 表示用户 u 对项目 i 的评分，$\overline{R}_i$ 表示项目 i 的均分，U 表示对两项目均存在评分的用户集。

一般用公式（2-7）预测用户 u 对未评分项目 i 的评分。

$$P(u,i)=\frac{\sum_{j\in I}sim(i,j)\times R_{u,j}}{\sum_{j\in I}|sim(i,j)|} \quad (2\text{-}7)$$

其中 I 表示项目 i 的邻近项目集合。

基于项目的协同过滤时间复杂度比基于用户的协同过滤时间复杂度低，项目之间相似度相对稳定，一定程度上降低了生成推荐列表所需时间。

（3）混合推荐：纯粹的基于内容推荐和协同过滤都存在不可避免的缺点，比如基于内容的推荐对专业知识要求过高，而协同过滤也存在致命的冷启动问题。因此，为了弱化单一推荐算法对推荐产生的影响，混合推荐策略应运而生。比较常见的混合策略有加权、混合、级联等。其中，加权策略将不同的推荐方法得到的推荐结果进行加权后再得到最终的推荐列表；混合策略则是把所有的推荐方法的推荐结果全部展示给用户，用户自行选择；级联策略则是先由一个推荐算法产生一个粗略的推荐列表，再由另一个推荐算法在此列表的基础上进行筛选。

综合考虑上述三种方法，由于基于项目的协同过滤所要求的专业知识过高，本文决定采用改进的基于用户的协同过滤算法，采用用户画像的相似度计算代替普遍的评分矩阵的用户相似度计算，可以更加全面地使用用户的信息，使相似度更加可靠。同时，在计算耗时也得到了降低。因为用户注册会产生基本信息，根据用户画像进行推荐，也可以一定程度上减轻冷启动问题所带来的影响。

2.3 推荐系统评估方法

推荐系统常采用的评价指标有：平均绝对误差、召回率、准确率。

（1）平均绝对误差：表示预测评分于实际评分的误差程度，平均绝对误差越大则评分预测越不准确，通常用式（2-8）计算平均绝对误差。

$$MAE = \frac{\sum_{(u,i)\in T} \left| \hat{r}_{ui} - r_{ui} \right|}{\left| T \right|} \tag{2-8}$$

其中，T 是测试集，$\hat{r}_{ui}$ 是预测评分，r_{ui} 是实际评分。

（2）召回率：表示成功预测项目在全部项目中的占比，用式（2-9）计算。

$$R = \frac{N_i}{N_r} \tag{2-9}$$

其中，N_i 为成功预测个数，N_r 为用户喜欢的所有项目数。

（3）准确率：表示成功推荐项目在总推荐项目中的占比，用式（2-10）计算。

$$P = \frac{N_i}{N_a} \tag{2-10}$$

其中 Na 表示总得推荐项目数。

由于本文的推荐结果不涉及评分信息，因此，本文主要采用准确率和召回率这两个评价指标对推荐算法进行评估。

2.4 本章小结

本章介绍了用户画像和推荐系统的相关理论与技术研究，首先介绍了用户画像的概念、构成要素、构建流程以及构建方法，然后介绍了推荐系统的概念及意义，介绍了几种主流的推荐算法，最后介绍了推荐系统的几种常用的评估方法。

第三章　用户画像构建设计

3.1 基于 BERT 的情感文本分析

情感分析是自然语言处理的热门研究内容之一。文本情感分析方法主要基于三种方法研究：情感词典、机器学习及深度学习。其中，基于深度学习的情感分析步骤为：文本预处理、生成词向量、送入深度模型进行训练。针对传统的词向量生成方式不能很好地捕捉文本的双向语义特征，BERT 预训练模型通过真正的双向模型可以捕捉到文本的更完整的特征。本文在此基础上设计并建立 BERT+Bi-LSTM+ Attention 情感分析模型。

3.1.1 BERT 预训练模型

BERT（Bidirectional Encoder Representation from Transformers）是 google 的 Devlin J 等[38]于 2018 年 10 月提出的预训练模型，如图 3-1 所示。在 11 个 NLP（Natural Language Processing） 任务上的表现刷新了记录。该模型采用双向的 Transformer 作为编码器，创新性地提出了 Masked 语言模型（ Masked Language Model ） 与 下 一 个 句 子 预 测 （ Sentence-Level Representation）任务。

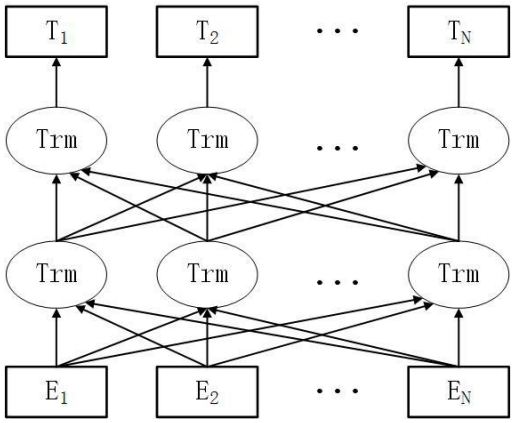

图 3-1　BERT 预训练语言模型

“Masked 语言模型”在训练集中随机给 15%的词打上[MASK]标记，具体规则分为三种：直接换为[MASK]、用随机的其它单词覆盖和保留原始 Token。其中，三种规则的概率分别是 80%、10%和 10%。“下一个句子预测”任务是为了学习句子之间的关系：随机替换一些句子，然后利用上一句预测是否是下一句。BERT

的核心是采用了双向的 Transformer 作为编码器，Transformer 的编码单元如图 3-2 所示。

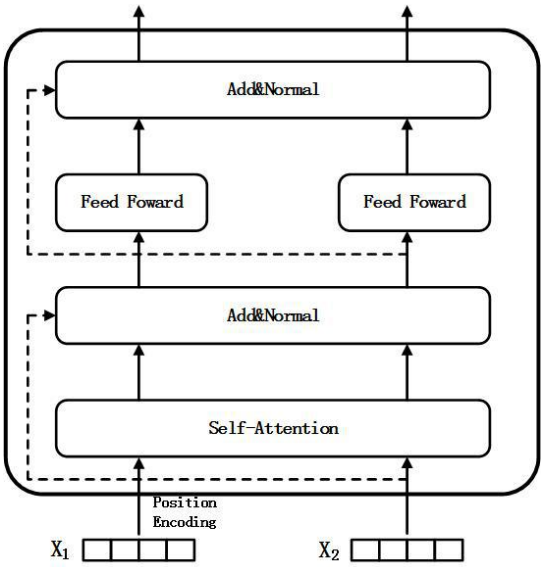

图 3-2　Transformer 编码单元结构

为了解决自注意力（self-attention）层提取出来的特征没有位置信息的问题，在输入时首先经过一层 position encoding，从编码器输入的句子首先会经过一个自注意力（self-attention）层，该层帮助编码器在对每个单词编码时关注输入句子的其他单词。自注意力层的输出会传递到前馈（feed-forward）神经网络中。每个位置的单词对应的前馈神经网络都完全一样。为了解决深度学习中的退化问题，Transformer 编码单元中加入了残差网络（图示虚线部分）和层归一化[39]。其中，Add 操作借鉴了 ResNet 模型的结构，其主要作用是使得 transformer 的多层叠加而效果不退化，Layer Normalization 操作对向量进行标准化以简化学习难度。

BERT 预训练语言模型与其它语言模型相比，可以充分利用词左右两边的信息，获得更好的词分布式表示。

3.1.2 LSTM 与 Bi-LSTM 模型

长短期记忆模型（Long Short-Term Memory，LSTM）是一种特殊的循环神经网络（RNN）模型，是 Hochreiter 等[40]于 1997 年提出的，主要解决了传统 RNN 中梯度消失与梯度爆炸的问题。LSTM 可以捕获长时依赖问题中的语句的长期依赖关系，从而可以更好地从文本的整体上进行情感分析。LSTM 中有 3 个控制门：遗忘门、输入门和输出门。记忆结构如图 3-3 所示：

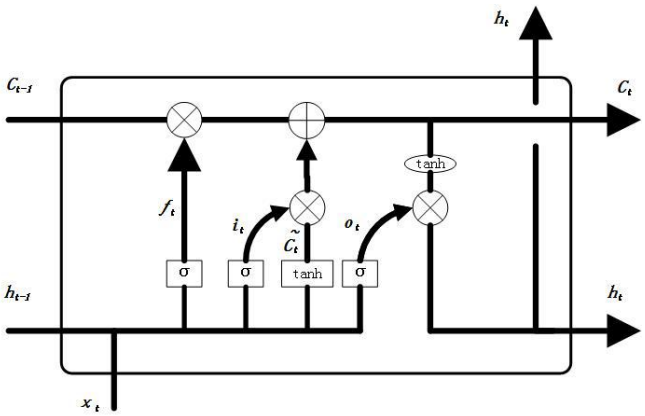

图 3-3　LSTM 记忆单元结构图

整个过程主要公式如下：

$$f_t = \sigma\left(w_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \quad (3\text{-}1)$$

$$i_t = \sigma\left(w_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \quad (3\text{-}2)$$

$$C_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \quad (3\text{-}3)$$

$$C_t = f_t \times C_{t-1} + i_t \times C_t \quad (3\text{-}4)$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \quad (3\text{-}5)$$

$$h_t = o_t \times \tanh\left(C_t\right) \quad (3\text{-}6)$$

其中 ht-1 表示的是上一个细胞的输出；ht 表示当前细胞的输出；xt 表示的是当前细胞的输入； σ 表示 sigmod 激活函数；tanh 表示 tanh 激活函数；ft 为遗忘门输出，决定从 cell 状态中丢弃什么信息；it 与 $\overline{C}_t$ 的乘积为输入门输出，决定让多少新的信息加入到 cell 状态中来；ot 为输出门输出，输出将会基于 cell 状态进行过滤。

在情感分类上，LSTM 比 RNN 耗时短且准确率更高。而在 LSTM 的基础上，双向长短记忆网络 (Bidirectional Long Short-Term Memory， Bi-LSTM)。Bi-LSTM 由前向 LSTM 与后向 LSTM 组合而成，分别从句子的两个方向对句子进行分析，然后将分析结果合并，使得 Bi-LSTM 可以更好地捕捉双向的语义依赖。

3.1.3 注意力机制

注意力机制（Attention）启发自人对事物的关注。注意力机制在模型进行输出操作的时候选择关注那些我们输入的需要关注的信息。而 self-attention 可以在没有注重的信息输入的情况下，从当前句子中选择需要关注的信息。同时，self-attention 能捕捉到句子中距离较远的词之间的特征关系。

3.1.4 BERT+Bi-LSTM+Attention 模型

首先对文本数据进行预处理，然后将预处理后的文本送入 BERT+Bi-LSTM+Attention 模型进行训练。

（1）数据预处理：经分析文本后发现长度超过 25 汉字（英文一个单词算一个汉字）的文本占总文本的 5.31%，为尽量保留文本信息的同时降低训练需要花费的时间，且使得数据满足模型的要求，将单条文本长度设置为 25 汉字，超过 25 个汉字的部分舍弃掉，长度不足的用 0 补齐。

（2）情感分析模型：传统的基于深度学习文本情感分析方法生成词向量的步骤为：分词、去停用词、生成词向量。最常用的生成词向量采用的是 Word2Vec，但它无法很好的捕捉文本的双向语义特征。由于 BERT 充分利用词左右两边的信息，获得更好的词分布式表示，本文将文本送入 BERT 生成字向量，代替 Word2Vec 生成的词向量。

模型结构如图 3-4 所示。将处理后的文本经过 BERT 预训练模型得到字向量；BERT 层之后接入 Bi-LSTM 网络执行特征提取操作； Dropout 层的作用是防止模型仅在训练集效果好而在测试集表现不佳；经过 Attention 层，得到关注权值不同的句子的向量表示，然后将所有词向量加权得到句子的向量表示；随后全连接层整合提取到的特征，进行全连接操作；最后通过定义的判别函数输出对应类别。

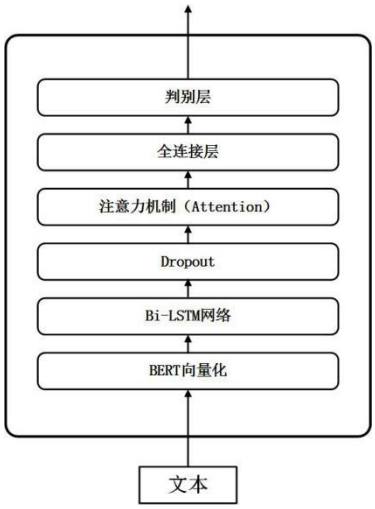

图 3-4　BERT+Bi-LSTM+Attention 模型结构图

3.2 兴趣遗忘曲线研究

3.2.1 记忆遗忘曲线

德国心理学家艾宾浩斯研究人脑对于新事物的遗忘与记忆规律发现：遗忘在学习之后立即发生，起初以较快的速度开始遗忘，随后遗忘速率逐渐缓慢[41]。同时，他指出记忆随时间变化的规律，见表 3-1。

表 3-1　记忆遗忘规律

时间间隔	记忆量
刚记完	100%
20 分钟后	58%
1 小时后	44%
8-9 小时后	36%
1 天后	33%
2 天后	28%
6 天后	25%
31 天后	21%

3.2.2 兴趣遗忘曲线

根据艾宾浩斯曲线模拟兴趣遗忘规律，见式（3-7），可以很好的拟合艾宾浩斯曲线：

$$M(t)=\frac{0.75}{1+0.42t}+\frac{0.25}{1+0.0003t} \tag{3-7}$$

其中，t 表示用户最后一次对某兴趣的访问时间间隔，由于不同于记忆，记忆通常采用的时间单位为小时，考虑到兴趣偏好是长期形成的习惯，且通常保持周期较长，因此，通过类比记忆遗忘曲线，式（3-7）中的 t 以月为单位。

3.3 用户画像标签模型设计

3.3.1 用户画像标签设计

用户在使用电商 APP 或者购物网站时的历史数据经过处理得到用户画像。用户画像是用户信息全貌的抽象体现，用户画像标签由两部分组成：静态标签和行

为标签。如图 3-5。

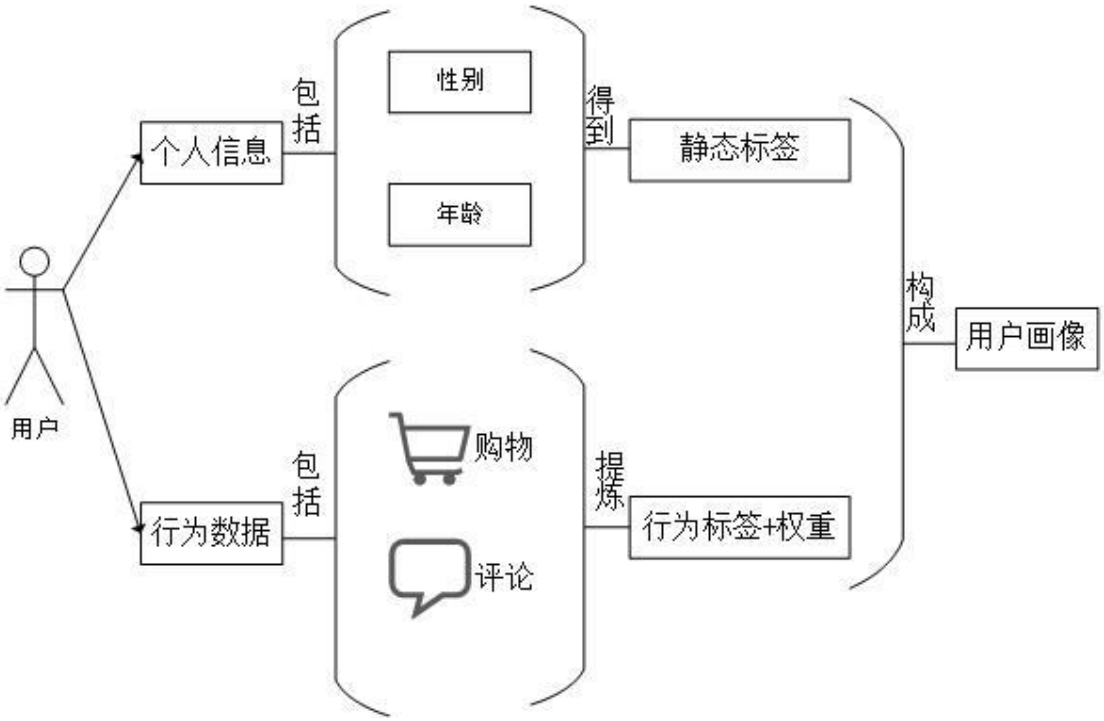

图 3-5　用户画像标签组成

其中，静态标签为用户在使用 APP 或者网站时所填写的个人信息，如年龄、性别等；动态标签则来源于用户的行为数据，如评论、购物信息等。

（1）数据采集与处理：从某电商后台数据库获取用户基本数据，如图 3-6。

用户编号	同一微信公众平台编号	用户昵称	用户姓名	手机号
000004f46f7d4917ad8d3	(Null)	(Null)	(Null)	189558
000044a3c7e34f7580e21(	op0DtjnlNVsA-yXstHMAD	<梅>琳彬	(Null)	(Null)
000056e0e1d14468ad941	(Null)	(Null)	(Null)	185302
000069eb176842d1b49f5	oit5W0oE3mZ5EywcGliQ	多雨的季节	(Null)	(Null)
00007311c9024cbe8066d	oit5W0pvhsg6r-lZOUgtp)	张雷	张雷	137165
000095d5d6d34d83a556	oit5W0iFGum9yv4Z87ByQ	乌托邦	吴庆业	133636
0000a3fac0114325b949d	op0DtjlvMyjboiEB923qm	N1ce 小美	王仟仟	150012
0000bdf0ef8a454d95547	op0DtjnceXcoqxhEGRDE	猫九	杨军	187938
0000e7e8503b456aa7325	op0Dtji8Cs7YP3_32qFLyo	Mr.QTแล้วแต่	齐勇	153531
00012ce67c4c4ed4ada66	(Null)	小 雄	(Null)	(Null)
000133b1886d4a78ad2d	oit5W0oD_DrnLzzTpfMN	Deep Blue	(Null)	(Null)
00013916970443f8afcbca	(Null)	(Null)	(Null)	133533
00014426b59b490e8f6a7	(Null)	自观	李帅	133313
0001500ec2d44be59c362	op0DtjslHxAK53feDJ3Vm	韩建欧	我	139345
0001661aba224098ba192	op0Dtjuj-M0kXhZ5UZyNS	单纯	(Null)	(Null)

图 3-6　后台获取的用户基本数据

其中，用户为了保护用户隐私，首先对数据进行处理时，对于静态数据，去掉用户的敏感信息，仅保留实验所需数据，并进行脱敏操作。将脱敏后的数据保存在数据库中。由于用户的隐私保护意识和用户的 APP 使用习惯，部分用户在个人信息填写部分对于性别这种非必填信息没有录入，由于空值会极大程度干扰画像的生成，所以需要对空值进行一定的处理。常用的空值填充的方法有随机、中位数和众数等。对于性别，采用随机填充。根据用户填写的生日计算用户的年龄，生日空缺则不计算。同时，分析发现计算后的年龄一栏的值存在误差值，对于年龄范围不在 15-80 这个范围区间的值采用该范围内的随机值代替。经采集并处理后的用户基本数据如图 3-7。

用户编号	昵称	性别	年龄	会员标识	代理商标识	积分余额	注册日期
1b6a64a5de02472da9400afb8c845225	_99	0	37	0	0	100	2018-01-01 00:05:06
cd8e2bf850ed4837866de1bfac2c9c9b	香桂(张洁)	1	29	0	0	100	2018-01-01 00:58:55
fa99c3a900044610b4502a1bedfb59e7	&	0	33	1	1	669	2018-01-01 05:19:16
539a6a73af9c48c5b90f22772f279619	心若初晴	0	28	1	1	0	2018-01-01 07:23:59
de3b6b43a7e04db3879ca1087de9c762	jing	0	42	1	1	500	2018-01-01 07:49:05
579c6e33eaa14dc4ba7cf00cfefe9cd8	福	1	24	0	0	100	2018-01-01 08:02:05
df86d77884a94779997ba333e28ce231	刘小瑞	1	30	1	1	162472	2018-01-01 08:05:09
e654f6feeb634a6289385c363c8a0301	游走四方	0	44	0	0	100	2018-01-01 08:17:55
b6ee1c04084749209f77cc30fb76f033	李立文	0	31	0	0	100	2018-01-01 08:20:55
2e249ad9b9e74e05af44303732078f28	微微初晨	0	20	0	0	100	2018-01-01 08:22:44
bb9cc273c36a4d3da1be2e6d456bb11a	李金良----康乐时尚营养俱乐部	0	39	0	0	100	2018-01-01 08:35:56
5cf7046fda6449a6a6ce66ba5004b5e3	AAA河北惠普戴尔-孔	1	30	0	0	100	2018-01-01 09:10:05
111c2e751787495980ed035654906b01	问心无愧	0	36	1	1	900	2018-01-01 09:20:56
e73dc66a14fb448aa99d68ab9b8a5951	Ida　Z	0	35	0	0	100	2018-01-01 09:27:40
a70d572153b24fdfbf8855129d8cb2b7	三韵茶庄18832622774	0	29	1	1	10097	2018-01-01 09:30:43

图 3-7　处理后的用户基本数据

处理后的用户静态数据共 50 万条。存放在用户信息表中。

对商品信息进行处理时，只需要去掉所有不需要的字段即可。处理后的商品数据共 6 万条。

对动态数据进行处理时，首先，删除所有不需要的属性；然后，把所用订单状态为未完成的订单全部从数据中除去；最后，删除详情表中的对应数据。经处理后的用户行为数据共 25 万条。其中，评论表增加一列分析结果属性，用于保存后续情感分析结果，如图 3-8。

评论编号	订单编号	商品编号	用户编号	评论文本	评论日期	分析结果
0008411afdbe4d209d74b	20180611102737279	1142060	11107729	上面的姜末真的是姜末啊，i	2019-02-28 17:07:36	(Null)
0008b82b623b42fba14cf4	20181128223628265	1655004	a6374550a0b14de4b13ff	质量不错	2018-12-08 19:35:54	(Null)
001277a1394c43879e701	20181031181652205	a690a180266141598a6a8	db52e092b78b42b5b200	好	2018-11-14 12:58:41	(Null)
001425117cdc4dba8820c	20190114212334775	1451008	bc02fef01eef4f9a8eae84f	实物比图片更好看，质量很	2019-01-16 18:42:07	(Null)
001b51b859e347e8ab974	20181009091045351	a9847995f3744ecb814f25	b2b8bd530f6147389b12c	家里人用，超棒，内存大	2019-01-02 21:59:24	(Null)
001dd52b843a42bb8491	20181117164214740	ba0a9d5720d0409e814ec	248fd42fb03445b3948c0	不错	2018-11-26 14:23:46	(Null)
001f2d876fd841e39ec847	20190131171631046	10caa656c3ef419a9d7d3	ee595ba0ae7840aea7b4(	确实很实惠	2019-04-15 22:48:19	(Null)
002b3af8687b475ab6a4e	20190126160515358	26b7876b89cd4597b7755	0127c1a4558e4fcb92b21	挺好	2019-01-26 17:06:22	(Null)
003348f388054e61877da	20181007235725749	2a511fb105c547bf8d506l	dfa264cddfc64eebbfde3l	质量好，外观时尚，电池耐	2018-10-24 14:19:07	(Null)
0035765403174358abce5	20181007201134591	7c39bb886bb44b3983e9!	e0d98c6cbdd2406c9b94(	好好好	2018-10-10 22:42:26	(Null)
00372bc60043490595c7c	20190221202541948	9868e6d122184885b66a(	77d63f1fd2334806b1700!	非常划算的套餐，以后会给	2019-02-21 20:40:32	(Null)
00375ee410124d5bbb42!	20181002194552937	794dff6fd0c6431aa07e21	facf01c9793e491aafc90b(	好	2018-10-09 18:54:12	(Null)
0039b3670486474faf8d3(	20180920124729675	e7092e3092e64f8eac43e(	eefca1d23c794e429c850(	非常不错，挺实惠	2018-09-20 13:07:50	(Null)
003b874fab5f4c0cb306c7	20180816155314129	9b4eed6b57e7408caf28d	5ad6403132904073818a9	很快	2018-08-16 15:53:49	(Null)
00405366e48449de9b2cf	20190203142259226	8a5a424a3e6f4fb988516!	2e93619cf1154494892674	这个平台真垃圾	2019-02-19 14:03:22	(Null)
004246dcc57645009c069	20180426104145642	2e31a7c805054e478b6e2	11112547	携带方便 再也不担心手机没	2018-06-08 10:08:40	(Null)
0045c6a07b894fb9814ab	20190714201007769	5b8bd58e9e7f4ed1b065:	517f6771446a466094d07	迪信通买手机放心实惠质量	2019-07-14 20:13:54	(Null)
004995757f8043ecbd1d2	20190209121903740	8483e69c5bfe47f093dca5	d585f0f90f0447b0862fb1	迪信通的优品好，手机价格	2019-02-09 14:39:06	(Null)
004dec8f9ce941eb91f648	20181118110821882	da19798078884b389dc47	4367730dcc124650969bc	超大屏幕，超级清晰	2018-11-18 19:41:30	(Null)

图 3-8　处理后的评论信息表

（2）标签确立：对于静态数据，直接确定两个标签，分别为性别和年龄；对于动态数据，采用 Elkan k-means 聚类算法对用户打上消费水平标签。Elkan k-means 算法是基于 k-means 算法的一种改进算法。k-means 算法的主要思想是：计算样本点之间的距离，将距离较近的划分为一类。算法伪代码如下：

1. 给定样本集合 $D=\{x_1,x_2,\ldots,x_n\}$，聚类簇数 k
2. 从 D 中随机选择 k 个样本作为初始均值向量 $\{\mu_1,\mu_2,\ldots,\mu_k\}$
3. Repeat
4. 令 $Cp=\varnothing(1\le p\le k)$
5. for $q=1,2,\ldots,n$ do
6. 计算样本 x_q 与各均值向量 $\mu_p(1\le p\le k)$ 的距离 $d_{pq}=\left|x_q-\mu_p\right|^2$
7. 选取距离最近的均值向量 x_q 确定的簇标记 $\lambda_q=\arg\min_{p\in(1,2,\ldots,k)}d_{pq}$
8. 将样本 x_q 划入相应的簇 $C_{\lambda_q}=C_{\lambda_q}\cup\{x_q\}$
9. end for
10. for $p=1,2,\ldots,k$ do
11. 计算新均值向量 $\mu'_p=\dfrac{1}{|C_p|}\sum_{x\subset C_p}x$
12. if $\mu'_p\neq\mu_p$
13. 将当前均值向量 μ_p 更新为 μ'_p
14. else
15. 保持当前均值向量不变
16. end if
17. end for

18.Until 当前均值向量均未更新

而 Elkan k-means 算法则在 k-means 算法的基础上做了如下优化：

1.对于给定的样本点 x 和两个均值向量 μ_p ， μ_q ，两个均值向量之间的距离 $D(\mu_p,\mu_q)$，若存在 $D(x,\mu_p)\le\frac{1}{2}D(\mu_p,\mu_q)$ 则有 $D(x,\mu_p)\le D(x,\mu_q)$。

2.对于给定的样本点 x 和两个均值向量 μ_p ， μ_q ，两个均值向量之间的距离 $D(\mu_p,\mu_q)$，若存在 $D(x,\mu_p)\le|D(x,\mu_p)-D(\mu_p,\mu_q)|$ 则有 $D(x,\mu_p)\le D(x,\mu_q)$。

Elkan k-means 算法每次迭代不用遍历所有样本来计算样本和每个均值向量的距离，迭代速度相比于 k-means 算法有较大提高，算法效率有较大提升。

将用户 2017 和 2018 年的年消费总额相加，然后将消费总额进行聚类分析，选取 k 值为 3，聚 3 类。将聚类后的总额较低一类打上低消费标签，较高一类打上高消费标签，其余用户打上中消费标签。将 2017 和 2018 年的消费每个用户分别计算用户 4 个季度的消费额分别占总额的比例，选取最高的值为用户打上偏好季度标签。分析用户所有的购物类别，为用户打上 5 个偏好标签，分别是电子产品爱好者、穿搭达人、美食达人、居家能手以及办公狂人。

（3）权重计算：对于性别、年龄、消费水平和偏好季度四个标签，权重全部为 1。对于电子产品爱好者、穿搭达人、美食达人、居家能手以及办公狂人 5 个标签，采用 TF-IDF 计算。TF-IDF 常用于事物检索，核心是衡量个体在整体中的重要程度，个体通常指字或者词。其中，词频（term frequency， TF）指的是个体在某个整体分集中出现的次数占比，

采用式（3-8）计算。

$$TF_w=\frac{\text{在某一类中词条}w\text{出现的次数}}{\text{该类中所有的词条数目}} \tag{3-8}$$

逆向文件频率 IDF 的主要思想是计算当前个体在整个整体集中的特殊程度。采用式（3-9）计算。

$$IDF=\log\left(\frac{\text{语料库的文档总数}}{\text{包含词含}w\text{的文档数}+1}\right) \tag{3-9}$$

其中分母加一是为了防止分母为 0。

TF-IDF 采用式（3-10）计算。

$$TF-IDF=TF\times IDF \tag{3-10}$$

根据某用户所购买的商品所属的某标签分类 x 的总订单数，除以该用户所有订单总数得到 TFx，再用 log(总订单数/(该标签 x 的订单总数量+1))得到此标签的 IDFx，最后用 TFx*IDFx 得到最终的标签权重。初步计算后的标签权重基本标

签部分如图 3-9（a）所示，偏好标签部分如图 3-9（b）所示。

用户编号	男	女	青年	中年	老年	低消费	中等消费	高消费	春季消费	夏季消费	秋季消费	冬季消费
ad3c100b27754486a6d0	0	1	1	0	0	1	0	0	0	1	0	0
11108133	0	1	1	0	0	1	0	0	0	1	0	0
8910571991f34485a10d4	1	0	1	0	0	1	0	0	0	1	0	0
11107202	0	1	1	0	0	1	0	0	0	1	0	0
34385faffc2a434e8e91ad	1	0	0	1	0	1	0	0	0	0	1	0
3bd55332a678427cae013	0	1	0	1	0	1	0	0	0	1	0	0
719251f7cc164598b7422	0	1	0	0	1	1	0	0	0	0	0	1
bf85cd6d75b24a2199439	1	0	1	0	0	1	0	0	0	0	0	1
3459b87d311b4835bae0	0	1	1	0	0	1	0	0	0	0	1	0
97e00c325b544c8aac7c9	0	1	0	1	0	1	0	0	0	1	0	0
b3d4dcaa9af14c82a053a	1	0	1	0	0	1	0	0	0	0	0	1
b7a95184d9c64db8b903	0	1	1	0	0	1	0	0	0	0	0	1
dce82456aa754f1d9a411	0	1	0	0	1	1	0	0	0	1	0	0
0a7a8769b19b4ad18bb1	1	0	0	1	0	1	0	0	0	0	1	0
2536964f731b4933a04a8	0	1	0	1	0	1	0	0	0	1	0	0

图 3-9（a）　初始用户画像基本标签权重

电子产品爱好者	穿搭达人	美食达人	居家能手	办公狂人
0.0062595012748838785	0	2.143621885942518	0	0
0.008346001699845172	0	2.1171574182148327	0	0
0.0097370019831527	0	0	2.01751277095637	0.26004962591809194
0.009828002001686839	0.08926381905045015	1.9529293194187287	0.042424333968708705	0.08749333208459167
0.010954127231046789	0	1.5283230112738324	0.567425466831479	0
0.011684402379783241	0	2.074814269850536	0	0
0.011684402379783241	0.21224952529773702	1.7784122313004596	0.15131345782172773	0
0.011684402379783241	0.21224952529773702	1.9266132505754978	0	0
0.012519002549767757	0	0	2.1075803053740647	0
0.012519002549767757	0.2274102056761468	1.7466548700272368	0.16212156195185112	0
0.013482002745903741	0.24490329842046582	0.8550058804329133	0.8729622566638139	0.3600687128096658
0.013482002745903741	0.7347098952613975	1.1970082326060785	0.34918490266552554	0
0.013482002745903741	0	1.8810129369524091	0.17459245133276277	0
0.013482002745903741	0.7347098952613975	1.0260070565194959	0	1.0802061384289974
0.013482002745903741	0	1.8810129369524091	0.17459245133276277	0

图 3-9（b）　初始用户画像偏好标签权重

3.3.2 评论情感分析调优用户画像

用户的喜好通过用户的行为外在表现。对于电商系统而言，用户的喜好通常表现为对于商品的点击、收藏和搜索等行为，用户购买某一商品后对于这一商品的评论更是直观的表现了用户对于此商品是否喜爱。因此，对用户的商品评论文本进行情感分析，得到用户对商品是否满意再根据此结果调整用户画像可以使用户画像更加精准，更能代表用户本身。评论文本情感分析调优用户画像的步骤如下：

（1）将评论文本送入本文的 BERT+Bi-LSTM+Attention 模型进行情感分析，将对应的评论文本打上标记，结果如图 3-10 所示。

订单编号	商品编号	用户编号	评论文本	评论日期	分析结果
20180611102737279	1142060	11107729	上面的姜末真的是姜末啊，i	2019-02-28 17:07:36	1
20181128223628265	1655004	a6374550a0b14de4b13ff	质量不错	2018-12-08 19:35:54	1
20181031181652205	a690a180266141598a6a8	db52e092b78b42b5b200	好	2018-11-14 12:58:41	1
20190114212334775	1451008	bc02fef01eef4f9a8eae84f	实物比图片更好看，质量很	2019-01-16 18:42:07	1
20181009091045351	a9847995f3744ecb814f25	b2b8bd530f6147389b12c	家里人用，超棒，内存大	2019-01-02 21:59:24	1
20181117164214740	ba0a9d5720d0409e814ec	248fd42fb03445b3948c0	不错	2018-11-26 14:23:46	1
20190131171631046	10caa656c3ef419a9d7d3	ee595ba0ae7840aea7b4f	确实很实惠	2019-04-15 22:48:19	1
20190126160515358	26b7876b89cd4597b775	0127c1a4558e4fcb92b21	挺好	2019-01-26 17:06:22	1
20181007235725749	2a511fb105c547bf8d506	dfa264cddfc64eebbfde3	质量好，外观时尚，电池耐	2018-10-24 14:19:07	1
20181007201134591	7c39bb886bb44b3983e9	e0d98c6cbdd2406c9b94c	好好好	2018-10-10 22:42:26	1
20190221202541948	9868e6d122184885b66ac	77d63f1fd2334806b1700	非常划算的套餐，以后会给	2019-02-21 20:40:32	1
20181002194552937	794dff6fd0c6431aa07e21	facf01c9793e491aafc90bc	好	2018-10-09 18:54:12	1
20180920124729675	e7092e3092e64f8eac43ec	eefca1d23c794e429c850c	非常不错，挺实惠	2018-09-20 13:07:50	1
20180816155314129	9b4eed6b57e7408caf28d	5ad6403132904073818a9	很快	2018-08-16 15:53:49	1
20190203142259226	8a5a424a3e6f4fb988516	2e93619cf115449489267	这个平台真垃圾	2019-02-19 14:03:22	0
20180426104145642	2e31a7c805054e478b6e2	11112547	携带方便 再也不担心手机没	2018-06-08 10:08:40	1
20190714201007769	5b8bd58e9e7f4ed1b065	517f6771446a466094d07	迪信通买手机放心实惠质量	2019-07-14 20:13:54	1
20190209121903740	8483e69c5bfe47f093dca9	d585f0f90f0447b0862fb1	迪信通的优品好，手机价格	2019-02-09 14:39:06	1

图 3-10　评论情感分析结果

其中满意为 1，不满意为 0。

（2）调整标签权重：取某一标签的某用户的全部评论，当好评数大于差评数，取式（3-11）调整权重；当好评数小于差评数，取式（3-12）调整权重。

$$W_n = W_o \cdot \left(1 + \left|\frac{P_g - P_b}{P_a}\right|\right) \tag{3-11}$$

$$W_n = W_o \cdot \left(1 - \left|\frac{P_g - P_b}{P_a}\right|\right) \tag{3-12}$$

其中 W_n表示计算后新的某类标签某用户的权重，W_o表示之前的权重，P_a表示全部评论数量，P_g表示好评总数，P_b表示差评总数。

3.3.3 兴趣遗忘曲线调优用户画像

用户的兴趣偏好随着时间会类似人类记忆逐渐消逝一样逐渐减弱。因此，采用拟合记忆遗忘曲线模拟得到的兴趣遗忘曲线对用户画像权重进行调优操作。兴趣遗忘曲线调优用户画像步骤如下：

（1）计算时间间隔：针对某一标签，遍历某一用户的该标签对应的所有订单，获取最近的一次订单时间与当前时间的差值，以月为单位，不足一月按 0 计算。

（2）调整标签权重：将某一用户的某一标签权重乘以兴趣遗忘曲线带入时间得到的遗忘率得到新的标签权重，见式（3-13）。

$$W_n = W_o \cdot \left(\frac{0.75}{1+0.42t} + \frac{0.25}{1+0.0003t} \right) \tag{3-13}$$

其中 W_n表示新的某一用户某一标签的新的权重，W_o表示之前的权重，t 表示间隔时间，单位是月。

对经过评论情感分析和兴趣遗忘曲线调优后的用户画像标签的权重进行归一化处理，处理后的用户画像基本标签部分如图 3-11（a）所示，偏好标签部分如图 3-11（b）所示。

用户编号	男	女	青年	中年	老年	低消费	中等消费	高消费	春季消费	夏季消费	秋季消费	冬季消费
ad3c100b27754486a6d0	0	0.1	0.1	0	0	0.2	0	0	0	0.2	0	0
11108133	0	0.1	0.1	0	0	0.2	0	0	0	0.2	0	0
0a7a8769b19b4ad18bb:	0.1	0	0	0.1	0	0.2	0	0	0	0	0.2	0
34385faffc2a434e8e91a(	0.1	0	0	0.1	0	0.2	0	0	0	0	0.2	0
719251f7cc164598b742:	0	0.1	0	0	0.1	0.2	0	0	0	0	0	0.2
bf85cd6d75b24a219943	0.1	0	0.1	0	0	0.2	0	0	0	0	0	0.2
3bd55332a678427cae01	0	0.1	0	0.1	0	0.2	0	0	0	0.2	0	0
b3d4dcaa9af14c82a053.	0.1	0	0.1	0	0	0.2	0	0	0	0	0	0.2
97e00c325b544c8aac7c!	0	0.1	0	0.1	0	0.2	0	0	0	0.2	0	0
b7a95184d9c64db8b90:	0	0.1	0.1	0	0	0.2	0	0	0	0	0	0.2
3459b87d311b4835bae(	0	0.1	0.1	0	0	0.2	0	0	0	0	0.2	0
2536964f731b4933a04a:	0	0.1	0	0.1	0	0.2	0	0	0	0.2	0	0
dce82456aa754f1d9a41:	0	0.1	0	0	0.1	0.2	0	0	0	0.2	0	0
2f8ad69381274fcb8339c	0.1	0	0.1	0	0	0.2	0	0	0	0.2	0	0
b517fb27dda34d3a829!	0	0.1	0.1	0	0	0.2	0	0	0	0.2	0	0

图 3-11（a）　调优的用户画像基本标签权重

电子产品爱好者	穿搭达人	美食达人	居家能手	办公狂人
0.0011646226274809645	0	0.39883537737251906	0	0
0.0015706399945816505	0	0.3984293600054184	0	0
0.0018892907358069285	0.10295804153020068	0.14377875923098846	0	0.15137390850300397
0.002079862094117996	0	0.2901829631581661	0.1077371747477159	0
0.0021701483908856526	0.03942118311400301	0.3303051638110097	0.028103504684101627	0
0.0021732891978263847	0.03947823650540072	0.3583484742967729	0	0
0.0022400018816773927	0	0.39775999811832263	0	0
0.0022983081266497413	0.04174923055657283	0.1457548259237175	0.14881589082590335	0.06138174456715654
0.002330519791174998	0.04233436193790606	0.3251547978178329	0.03018032045308608	0
0.0023504342209967463	0.12808833471362455	0.20868480490311328	0.060876426162265455	0
0.0023619653103943224	0	0	0.39763803468960573	0
0.0026063670010932384	0	0.3636410806244497	0.03375255237445707	0
0.0026063670010932384	0	0.3636410806244497	0.03375255237445707	0
0.002610898687801315	0	0.3973891013121987	0	0
0.002846564002551661	0	0.3971534359974484	0	0

图 3-11（b）　调优的用户画像偏好标签权重

3.4 情感分析实验结果分析

本节将介绍采用本文方法进行文本情感分析的实验结果分析。

本文所使用的数据集为带标签的商品评论数据，标签信息为用户满意情况，共 21065 条数据。其中，满意评价数据 10673 条，不满意评价数据 10392 条。实

验将 BERT+Bi-LSTM+Attention 模型与自构建情感词典、Bi-LSTM+Attention 两种方法的准确率和召回率进行了对比，如表 3-2 所示。

表 3-2　情感分析实验结果对比

实验方法	准确率	召回率
自构建情感词典	0.7981	0.7212
BiLSTM+Attention	0.8972	0.9118
BERT+BiLSTM+Attention	0.9348	0.9373

在对比实验中，我们发现 Bi-LSTM+Attention 模型在第 5 个 epoch 得到了最高的准确率 89.72%。使用 BERT 字向量后，BERT+Bi-LSTM + Attention 模型达到了 93.48% 的最高的精度。

3.5 本章小结

本章对用户画像的构建进行了介绍。首先介绍了 BERT+Bi-LSTM+Attention 情感分析模型，然后介绍了兴趣遗忘曲线，接着介绍了本文的用户画像标签模型。最后对本文用户画像构建过程中所用到的文本情感分析模型的实验进行了描述和分析。从数据采集处理、标签建立、权重计算以及应用 BERT+Bi-LSTM+Attention 情感分析模型和兴趣遗忘曲线精准化用户画像 6 个方面介绍了本文用户画像标签模型的完整建立过程。

第四章　个性化推荐算法设计

4.1 基于 slope-one 的协同过滤推荐算法研究

slope-one 算法常用于评分预测。由于它也需要构造用户-项目评分矩阵，因此它属于基于项目的算法。它用简单线性模型的计算代替复杂的相似度的计算，因此，它计算速度上更快。并且在数据较少时也有较好的表现。

4.1.1 slope-one 评分预测算法

slope-one 算法是利用线性模型预测用户对项目评分的一种算法，其主要思想是利用以评分项的评分均值预测未评分项。

slope-one 算法主要分两步进行：

（1）计算项目间用户的评分差均值。计算对两个项目都进行了评分的差值的和再除以全部对两个项目都进行了评分的人数，最终得到两个项目间的评分差均值。采用式（4-1）计算出所有项目之间的评分差均值。

$$dev_{ij}=\frac{\sum_{x\in(I_i\cap I_j)}\left(g_{xi}-g_{xj}\right)}{I_i\cap I_j} \tag{4-1}$$

其中 dev_{ij}表示项目 i 和项目 j 的评分差均值，$I_i\cap I_j$ 表示对于项目 i 和项目 j 都进行了评分的用户的数目，g_{xi}和 g_{xj}分别表示用户 x 对项目 i 的评分和用户 x 对项目 j 的评分。

（2）对未评分项目进行评分预测：预测某一用户的某一未评分项目的评分，首先计算所有已评分项目的评分加上该项目与未评分项目的均差值的和，然后除以已评分项目的个数，得到未评分项目的预测评分。采用式（4-2）进行评分预测。

$$P_{ui}=\frac{\sum_{i\in I_x}\left(I_i\cap I_j\right)\times\left(g_{uj}+dev_{ij}\right)}{\sum_{i\in I_x}\left(I_i\cap I_j\right)} \tag{4-2}$$

其中 I_x表示用户 x 评过分的项目。

4.1.2 基于 slope-one 的协同过滤算法

在原用户-项目评分矩阵上使用 slope-one 算法进行空值填充。填充后根据

新的用户-项目评分矩阵计算用户余弦相似度。

根据计算的相似度，选择与当前被推荐用户最相似的前 x 名最相似用户。将这 x 名相似用户的所有项目评分按从高到低进行排序，选取前 n 项该用户未评分项作为此用户的推荐列表。

4.2 基于用户画像的推荐算法设计

4.2.1 构建用户-项目评分矩阵

构建用户-项目评分矩阵参考用户购买商品后对商品的评论星级，评论总共 5 星，类比为 5 分。遍历用户的订单信息，得到所有的用户对于商品的评分信息，如图 4-1。此评分信息用于后续模拟用户-项目评分矩阵。

用户编号	商品编号	评分
5a53693b38fb423d9d2a3a2f0b0dd45e	f5db869eec8f4af68959b535d8b041d1	5
b8eaaf40b6cb471b83b4236a99fcf1fb	7ed3d09e9196457da3a33f7c289475db	4
5aed25eb392c4252808ce8f674712711	28eb536491394326b7e9843f47d98157	4
2f7c978677b24b4c9dc65e845c3f7e32	a6fdfac6873d4f62a009b7db32fa37d0	4
aa4944939ad745c7bcb83def13ab7e4b	bdb5cd48f5aa4913a3d8bbeba8865db	5
561fb8b8857d40c78bec0560895744c8	110e1575afa64c3199a0dd42ebe3e82b	Null)
80604f2847214ba0804b03ee03539df6	14d636e3d75943a589545906f355e3cd	2
5845cb32918d449d81886b88ed3ec03	28eb536491394326b7e9843f47d98157	4
13015ee7ba8e47ee92481dff15269e1a	945d26fdfeed4d9093b655c76acefd8b	4
c9cdfb92c7294a2991f94480f34be7bf	1165005	5
70996646bcce480788f3005f66504413	448ff6a301d24dc6b252a0f30e1eb17c	5
b5697fd753424f3cacc6e16b4bcb1399	5cbefda27b764be9bd533431afe2b5fc	Null)
11110944	015bb144f77547a8aebb3d56ed7fca4b	Null)
70996646bcce480788f3005f66504413	56ecf3bfc7244d41b26d1d019fcf20a3	5
a45cc4871a014613adfc1d60701876aa	63ad7ff180324caba047c5e989d231e6	5

图 4-1　用户对项目的评分信息

4.2.2 采用 slope-one 进行评分预测

经过分析发现，模拟的评分矩阵里面存在大部分空值，造成这种情况有两种原因：一是因为用户在使用电商 APP 或者购物网站时，由于订单完成后的评价功能不是强制的，因此部分用户没有填写评价，于是便留下了空值；二是因为用户尚未购买过部分商品，因此便没有评分记录，留下了空值。采用 slope-one 算法进行评分预测，填补空值。

（1）计算项目间用户的评分差均值，将计算后的结果进行保存。

（2）根据项目评分差均值进行评分预测，并将结果填入评分信息表，如图4-2。

用户编号	商品编号	评分
5a53693b38fb423d9d2a3a2f0b0dd45e	f5db869eec8f4af68959b535d8b041d1	5
b8eaaf40b6cb471b83b4236a99fcf1fb	7ed3d09e9196457da3a33f7c289475db	4
5aed25eb392c4252808ce8f674712711	28eb536491394326b7e9843f47d98157	4
2f7c978677b24b4c9dc65e845c3f7e32	a6fdfac6873d4f62a009b7db32fa37d0	4
aa4944939ad745c7bcb83def13ab7e4b	bdb5cd48f5aa4913a3d8bbeba8865db0	5
561fb8b8857d40c78bec0560895744c8	110e1575afa64c3199a0dd42ebe3e82b	5
80604f2847214ba0804b03ee03539df6	14d636e3d75943a589545906f355e3cd	2
5845cb32918d449d81886b88ed3ec033	28eb536491394326b7e9843f47d98157	4
13015ee7ba8e47ee92481dff15269e1a	945d26fdfeed4d9093b655c76acefd8b	4
c9cdfb92c7294a2991f94480f34be7bf	1165005	5
70996646bcce480788f3005f66504413	448ff6a301d24dc6b252a0f30e1eb17c	5
b5697fd753424f3cacc6e16b4bcb1399	5cbefda27b764be9bd533431afe2b5fc	3
11110944	015bb144f77547a8aebb3d56ed7fca4b	5
70996646bcce480788f3005f66504413	56ecf3bfc7244d41b26d1d019fcf20a3	5
a45cc4871a014613adfc1d60701876aa	63ad7ff180324caba047c5e989d231e6	5

图 4-2　评分预测后用户-项目评分信息

4.2.3 根据用户画像计算用户相似度

根据用户-项目评分信息采用协同过滤算法计算用户的相似度，仅利用到了用户的行为信息，没有考虑用户的基本信息。仅有行为信息不能完整的表征用户的个体。而用户画像的生成充分利用了用户的基本信息和行为信息，因此，在计算用户相似度时，基于已经构建的用户画像计算相似度，可以更好地体现用户的完整个体。将用户画像的每个用户当作向量，仅考虑两个用户对同一标签都存在权重的情况，采用皮尔逊相关系数计算用户相似度，夹角越大则用户越相似。计算后保存相似度最高且相似用户存在购买行为的十条相似度信息，用于推荐时参考。

4.2.4 采用 Top-N 策略进行推荐

根据用户相似度从前往后遍历相似用户，获取当前用户与所有相似用户的评分信息。对用户画像进行分析，若当前季节为用户的偏好购物季节，则选取所有评分中最高的当前用户未购买过的前 15 项商品对用户进行推荐；若当前季节非用户的偏好购物季节，则选取所有评分中最高的当前用户未购买过的前 10 项商

品对用户进行推荐，推荐后查询用户 2019 年订单，判断推荐的商品是否被用户购买。并将推荐列表保存，如图 4-3。

用户编号	商品编号	是否购买
0015be2a3a164b8e92486dc83cca34fc	02f72aea25604cccadffad3888edec03	0
0015be2a3a164b8e92486dc83cca34fc	000c947d26fe40cbbc9e190bb8cae35a	0
0015be2a3a164b8e92486dc83cca34fc	006f629a78154094a773f6ee873e4b29	0
0015be2a3a164b8e92486dc83cca34fc	00f957a738c44d1282f8653f8a6b4dc4	0
0015be2a3a164b8e92486dc83cca34fc	0036ec07246e4dd494f97b283281facc	0
0015be2a3a164b8e92486dc83cca34fc	0072cdcc7d2046c79fbccc438d2ec2d8	1
0015be2a3a164b8e92486dc83cca34fc	00f99e64b0c84fdf932c899fe759998f	0
0015be2a3a164b8e92486dc83cca34fc	0048383db4fb49ad95d6833f8459e824	0
0015be2a3a164b8e92486dc83cca34fc	005c9b2cda584cf5bf484496619b0bcb	0
0015be2a3a164b8e92486dc83cca34fc	007e4022b8c44361a01b4bf247f68d3d	0
0019d82a793b4b738dd4e708f87218b7	0072cdcc7d2046c79fbccc438d2ec2d8	0
0019d82a793b4b738dd4e708f87218b7	00ceaa2c81ce4ffc9ac5d6a8365898d9	0
0019d82a793b4b738dd4e708f87218b7	00eaf55f8e5249609b86c0e5ca1ee0f9	0
0019d82a793b4b738dd4e708f87218b7	01730aba04284ffcbaeed1fd580db432	0
0019d82a793b4b738dd4e708f87218b7	01e3df35652449f8a86b576833d14a2b	0
0019d82a793b4b738dd4e708f87218b7	01f7a0eb6e6e4d15997b5434c5188c9c	0
0019d82a793b4b738dd4e708f87218b7	0240271a08684b9085912ba894cc6ca2	0
0019d82a793b4b738dd4e708f87218b7	025483410e3243e7a1c3445e4ede4fd2	0
0019d82a793b4b738dd4e708f87218b7	05b43372c1ce4a9495186324756898cd	0
0019d82a793b4b738dd4e708f87218b7	05eaf5e3fc50469ca92363c8d1e43034	0

图 4-3　推荐列表

4.3 冷启动问题研究

协同过滤算法必然存在冷启动问题，这是由于新用户刚使用 APP 或购物网站时可能没有留下任何行为信息，只是简单注册留下了基本信息。因此，对于新用户，根据用户画像中的基本信息计算皮尔逊相关系数，随后根据相似度最大的 10 名用户，获取他们的用户-项目评分矩阵，选取评分最高的 10 件商品对用户进行推荐。

4.4 实验结果分析

4.4.1 评分预测实验

由于用户-项目评分信息中存在的空值会对后续实验产生影响，因此需要进行评分预测对空值进行填充。评分预测实验主要分三个步骤进行：

（1）从用户-项目评分信息中随机取十分之一的值置空并标记为 0；

（2）分别采用均值填充、众数填充和 slope-one 评分预测算法对标记的位置进行评分预测；

（3）对比三种预测值与原值，并计算准确率，结果如表 4-1。

表 4-1　评分预测方法准确率

方法名称	准确率
均值填充	0.312
众数填充	0.464
slope-one 评分预测	0.732

从实验结果可看出 slope-one 算法预测准确率更高，因此，本实验采用 slope-one 算法进行评分预测得到的值会比均值填充和众数填充更贴近实际。

4.4.2 商品推荐实验

分别根据用户画像和模拟的评分矩阵对用户进行商品推荐实验。推荐主要分 3 步进行：

（1）将所有用户的购买数据分为两个部分，2019 年的购买数据作为验证集，其余数据作为测试集；

（2）根据初始的用户画像、调优的用户画像和模拟的评分矩阵对用户进行推荐，并将推荐结果保存在对应的推荐信息表中，表结构如图 4-3；

（3）分别统计 2019 年存在购物行为的用户于三种方法各自成功推荐的次数占各方法总推荐次数的比例，得到三种方法的推荐准确率、召回率，结果见表 4-2。

表 4-2　商品推荐方法对比

方法名称	准确率	召回率
模拟的评分矩阵	9.3%	19.01%

初始的用户画像	11.65%	24.47%
调优的用户画像	14.39%	25.45%

从实验结果可看出，因为用户画像包含了用户的基本信息和行为信息，更加全面地利用了用户的信息，所以基于用户画像的推荐算法具有更高的准确率。同时，采用评论文本情感分析结合兴趣遗忘曲线调优后的用户画像推荐的准确率高于初始用户画像推荐的准确率，证明了调优操作的合理性。

4.5 本章小结

本章对基于用户画像的协同过滤推荐算法进行了介绍。首先介绍了slope-one 评分预测算法，然后介绍了基于 slope-one 的协同过滤推荐算法。通过分析 slope-one 协同过滤推荐算法的原理，引入用户画像计算用户相似度，利用了用户更加全面的信息。并根据相似度对用户进行个性化物品推荐。接着，针对协同过滤算法存在的冷启动问题，提出了解决办法。最后，对本文个性化推荐算法所涉及的评分预测实验和商品推荐实验进行了描述和分析。

第五章　推荐系统的设计与实现

5.1 系统设计

5.1.1 功能模块设计

本文的推荐系统结构如图 5-1 所示。

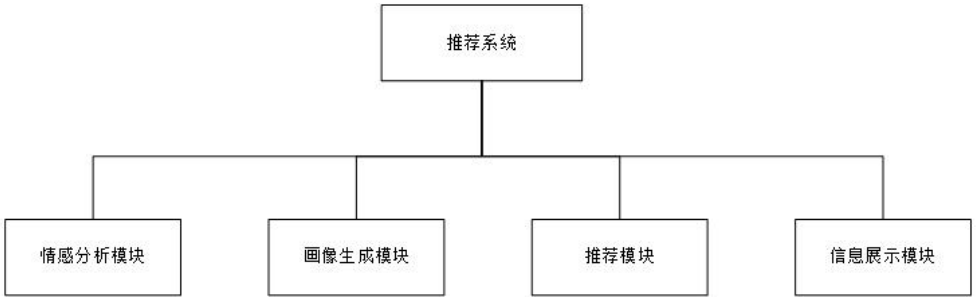

图 5-1　推荐系统功能模块结构

由图可知本系统主要由 4 大模块构成：

（1）情感分析模块：主要负责对用户的评论进行情感分析并对相应评论打上正/负标签；

（2）画像生成模块：主要负责根据用户数据计算得到用户画像；

（3）推荐模块：主要负责根据用户画像对用户进行商品推荐；

（4）信息展示模块：主要负责展示用户画像和推荐结果。

5.1.2 系统逻辑设计

系统逻辑如图 5-2 所示。

图 5-2　推荐系统运行流程图

从电商网站后台获取的数据首先进行数据预处理，经过处理后的评论数据送由情感分析模块进行分析，为评论打上标签。然后所有数据经过画像生成模块生成用户画像。用户画像生成后，数据进入推荐模块，根据画像对用户进行个性化推荐。最后信息通过信息展示模块进行展示。

5.1.3 数据库设计

通过分析整体业务，该系统总共的核心表共 12 张。其中，基本信息表 5 张，数据信息表 7 张。

基本信息表分别为：评论信息表、订单详情表、订单基本信息表、商品信息表和用户信息表。

评论信息表记录了所有已评论用户的信息，表结构如表 5-1 所示。

表 5-1　评论信息表

字段名称	字段含义	字段类型	是否可空	是否主键
id	评论编号	varchar(64)	否	是

order_id	订单编号	varchar(64)	否	否
goods_id	商品编号	varchar(64)	否	否
user_id	用户编号	varchar(64)	否	否
text	评论文本	varchar(512)	否	否
praise_sum	评分	int(11)	是	否
analysis_result	情感分析结果	int(11)	是	否

订单详情表记录了订单里商品的详细信息，表结构如表 5-2 所示。

表 5-2　订单详情表

字段名称	字段含义	字段类型	是否可空	是否主键
id	订单详情编号	varchar(64)	否	是
order_Id	订单编号	varchar(64)	否	否
goods_id	商品编号	varchar(64)	否	否
goods_name	商品名称	varchar(255)	否	否
goods_num	购买数量	int(11)	否	否
price	商品价格	decimal(10)	否	否
create_date	购买日期	datetime	否	否

订单基本信息表记录了订单时间、金额、购买用户等信息，表结构如表 5-3 所示。

表 5-3　订单基本信息表

字段名称	字段含义	字段类型	是否可空	是否主键
id	订单编号	varchar(64)	否	是
payment	支付金额	decimal(10)	否	否
payer_id	购买人	varchar(64)	否	否
pay_time	支付时间	datetime	否	否
status	订单状态	varchar(16)	否	否

商品信息表记录了所有的商品信息，表结构如表 5-4 所示。

表 5-4　商品信息表

字段名称	字段含义	字段类型	是否可空	是否主键
id	商品编号	varchar(64)	否	是
name	商品名称	varchar(255)	否	否

category	一级类目	varchar(255)	否	否
market_price	市场价	decimal（10）	否	否
member_price	会员价	decimal（10）	否	否
sales_volume	销量	int(32)	否	否
create_date	上架日期	datetime	否	否
second_cate	二级类目	varchar(255)	否	否

用户信息表记录了所有的用户信息，表结构如表 5-5 所示。

表 5-5　用户信息表

字段名称	字段含义	字段类型	是否可空	是否主键
id	用户编号	varchar(64)	否	是
nick_name	用户昵称	varchar(64)	否	否
sex	性别	int(11)	是	否
age	年龄	int（11）	是	否
member_flag	会员标识	char（1）	否	否
agent_flag	代理商标识	char(1)	否	否
online_integral	积分余额	int(10)	否	否
create_date	注册日期	datetime	否	否

数据信息表分别为消费总额表、用户画像表、评分矩阵表、评分矩阵推荐表、初始用户画像推荐表、调优用户画像推荐表和相似度表。

消费总额表统计了用户 2017 年和 2018 年的消费总额，用于聚类分析，表结构如表 5-6 所示。

表 5-6　消费总额表

字段名称	字段含义	字段类型	是否可空	是否主键
consumption_id	消费总额编号	int(11)	否	是
user_id	用户编号	varchar(64)	否	否
consumption_value	消费总额	double	否	否

用户画像表保存了用户的用户画像，表结构如表 5-7 所示。

表 5-7　用户画像表

字段名称	字段含义	字段类型	是否可空	是否主键
lable_id	标签编号	int(11)	否	是

user_id	用户编号	varchar(64)	否	否
sex1	男	double	否	否
sex2	女	double	否	否
age1	青年	double	否	否
age2	中年	double	否	否
age3	老年	double	否	否
consumption1	低消费	double	否	否
consumption2	中等消费	double	否	否
consumption3	高消费	double	否	否
season1	春季消费	double	否	否
season2	夏季消费	double	否	否
season3	秋季消费	double	否	否
season4	冬季消费	double	否	否
preference1	电子产品爱好者	double	否	否
preference2	穿搭达人	double	否	否
preference3	美食达人	double	否	否
preference4	居家能手	double	否	否
preference5	办公狂人	double	否	否

评分信息表保存了用户的评分信息，表结构如表 5-8 所示。

表 5-8　评分信息表

字段名称	字段含义	字段类型	是否可空	是否主键
rank_id	评分编号	int(11)	否	是
user_id	用户编号	varchar(64)	否	否
product_id	商品编号	varchar(64)	否	否
rank	评分	int(11)	是	否

评分矩阵推荐表保存了评分矩阵的推荐结果，初始用户画像推荐表保存了依据初始用户画像推荐的推荐列表，调优用户画像推荐表保存了根据文本情感分析和兴趣遗忘曲线调优后的用户画像进行推荐的推荐列表，三个表结构一直，表结构如表 5-9 所示。

表 5-9　商品推荐表

字段名称	字段含义	字段类型	是否可空	是否主键

recommendation_id	推荐编号	int(11)	否	是
user_id	用户编号	varchar(64)	否	否
goods_id	商品编号	varchar(64)	否	否
purchase	是否购买	int(11)	否	否
sort	排序	int(11)	否	否

相似度表保存了根据皮尔逊相关系数计算后的用户相似度,表结构如表5-10所示。

表 5-10　用户相似度表

字段名称	字段含义	字段类型	是否可空	是否主键
similarity_id	相似度编号	int(11)	否	是
user1_id	用户 1 编号	varchar(64)	否	否
user2_id	用户 2 编号	varchar(64)	否	否
similarity	相似度	double	否	否

5.1.4 实验环境和语言

课题研究用户画像在推荐系统的应用。实验依据某电商网站采集的脱敏数据，设计了一个 B/S 架构的推荐系统。编码所用的语言主要有 java、html、css、javascript。所采用的架构为 Spring+SpringMVC+Mybatis。实验所需环境主要包括 Tomcat 服务器，MySQL 数据库，Google 浏览器。具体信息见表 5-11。

表 5-11　实验环境信息表

名称	版本	用途
IDEA	1.8.0_92-b14 amd64	程序开发
Tomcat	7.0	服务器
MySQL	5.6	数据存储
Google	79.0.3945.88	访问网站

5.2 系统测试

软件测试是软件工程中不可或缺的重要环节之一。本节将按照软件工程的思想对系统进行单元测试。

单元测试按照模块划分，分别测试各个模块中单个方法的功能是否正确，测

试结果见表 5-12。

表 5-12　单元测试结果

模块名称	测试方法数量	通过	未通过
情感分析模块	3	3	0
画像生成模块	6	6	0
推荐模块	3	3	0
信息展示模块	12	12	0

5.3 结果展示与分析

5.3.1 结果展示

成功登录后进入推荐系统主界面。系统提供了基本的系统用户管理、系统管理及推荐中心管理。其中，推荐中心分为四个部分：用户信息、商品信息、评论信息和用户画像。推荐系统主界面模块结构如图 5-3。

图 5-3　推荐系统主界面模块结构

用户信息和商品信息均展示经数据清洗后的对应的基本信息。用户信息界面提供了用户的基本信息的分页查看及根据用户 id 和用户昵称查询对应用户的功能。以用户信息为例，如图 5-4。

序号	用户编号	用户昵称	性别	年龄	会员标识	代理商标识	积分余额	注册日期
1	000004f46f7d4917ad8d3d9341e1e6e5		女	23	否	否	0	2019-09-04
2	000044a3c7e34f7580e2106e0205ec90	<梅>琳彬	女	28	否	否	100	2018-08-03
3	000056e0e1d14468ad9417eac4ecbc70		男	46	否	否	0	2019-07-27
4	000069eb176842d1b49f5bea955067c5	多雨的季节	男	46	否	否	100	2018-09-29
5	00007311c9024cbe8066d5682a3370cd	张雷	男	27	是	是	600	2019-01-08
6	000095d5d6d34d83a55636c4611f6a6d	乌托邦	男	45	是	是	600	2018-09-09
7	0000a3fac0114325b949df7d693dbc1f	N1ce 尛美	女	63	是	是	600	2018-08-04
8	0000bdf0ef8a454d95547479954d412d	□猫九	女	26	是	是	600	2018-07-27
9	0000e7e8503b456aa73258252e7994d1	Mr.Qโคิดถึง	女	37	是	是	600	2018-08-30
10	00012ce67c4c4ed4ada6635cc52c1e95	小 雄	男	18	否	否	100	2018-01-19

共504709条数据，每页 10 条记录，当前第 1 页，共50471页

图 5-4　用户信息界面

用户评论信息界面提供了用户评论信息分页查看的功能，并展示了经文本情感分析后用户的真实情感。如图 5-5。

	评论	评分	评论日期	分析结果
28	非常划算的套餐，以后会给家里亲戚朋友推荐的。	5	2019-02-21	满意
3c	好	5	2018-10-09	满意
ef	非常不错，挺实惠	4	2018-09-20	满意
1f6	很快	5	2018-08-16	满意
5d	这个平台真垃圾	1	2019-02-19	不满意
	携带方便 再也不担心手机没电了	5	2018-06-08	满意
ab	迪信通买手机放心实惠质量	4	2019-07-14	满意
7	迪信通的优品好，手机价格实惠，比很多网上都便宜很多，而且都是正规保修，还有积分和钱赚，强烈推荐	4	2019-02-09	满意
0fb	超大屏幕，超级清晰 双手操作，极致体验	4	2018-11-18	满意
9b	到货速度快，好评	5	2019-07-04	满意

图 5-5　用户评论信息界面

用户画像界面提供了用户画像分页查看的功能，并可以根据用户 id、性别标签、年龄标签、消费水平标签、消费季度标签和用户的个性化标签筛选用户画像。由于用户画像的推荐列表数据量超过了 500 万条，因此用户的推荐列表不进行集中展示，由每个用户的用户画像点击对应按钮进入推荐列表界面。用户画像界面的基本标签如图 5-6（a）所示，个性标签如图 5-6（b）所示。

用户编号	男性	女性	青年	中年	老年	低消费	中等消费	高消费	春季消费	夏季消费	秋季消费	冬季消费
019401a2fbc2467a85461ce981df0440	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0
01d23c1014e14a598f1cb8db29b22882	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.2	0.0
11106977	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.2	0.0	0.0

图 5-6（a）　用户画像界面基本标签

电子产品爱好者	穿搭达人	美食达人	居家能手	办公狂人	推荐列表
0.15899393044345558	0.0	0.14828127895472837	0.030279080614457318	0.062445709987358745	普通推荐 初始画像 调优画像
0.03933531839797682	0.008028471801804988	0.173779700001495	0.13164111456069735	0.04721539523802582	普通推荐 初始画像 调优画像
0.053011931872390054	0.005177271139280786	0.06868453683342021	0.273126260154909	0.0	普通推荐 初始画像 调优画像

图 5-6（b）　用户画像界面个性化标签

推荐列表界面展示用户的推荐列表，推荐列表中包含被推荐商品是否已被用户购买的信息。如图 5-7。

商品编号	商品名称	是否被购买
1156073	冰箱贴开瓶器 风暴英雄 银色	否
1759142	网易云音乐卫衣C-HOODY Ear-wing耳朵私享家款-军绿色 XXL码	否
1293002	网易游戏热爱者胸包/斜挎包 鸽羽灰	否
1452008	2017暴雪嘉年华 暴雪文化 马克杯 暴雪文化	否
1390041	Q版史蒂夫 儿童短袜 我的世界 袜长18-20cm	否
1762088	洋葱小鱿T恤 白色 守望先锋 XXL	否
1326030	达摩异形抱枕 阴阳师 招福（红）	否
3635259	男式轻薄无感运动短裤 蓝色*S	否
1662037	星球大战 感受原力 主题短袖T恤 L*黑色	否
1260165	钻石史蒂夫 短袖T恤 我的世界 XXL	否

图 5-7　推荐列表界面

5.3.2 结果分析

（1）由用户-项目评分矩阵进行推荐的用户的推荐列表中不存在已购买商品，属于推荐失败。这些用户在经过用户画像进行推荐后，部分用户的推荐列表中出现了已购买的商品。说明采用用户画像进行推荐更为有效。以用户“0d7d60b60cfe4a7da0b6586fd0fa0566”为例，评分矩阵推荐列表如图 5-8（a）所示，初始用户画像推荐列表如图 5-8（b）所示。

商品名称	是否被购买
3包装丨日本进口 帮宝适一级帮纸尿裤 L52片/包 空气纸尿裤婴儿透气尿不湿	否
毛毛虫儿童运动鞋（断码清仓） 夏威夷粉 *105(2.0)，脚长11~12cm	否
月白林青·全棉提花四件套 1.5m床:适用2mx2.3m被芯	否
全棉色织磨毛四件套 红赤色*1.8M	否
adidas kids 阿迪达斯 男童 儿童鞋 D96858 30.5 一号黑/一号黑/白	否
极简保温随行杯 珠光黑	否
羊脂玉白紫金线茶具 6件装 【优惠组】羊脂玉白茶组+竹制干泡茶盘	否
儿童防蚊T恤 4-16岁 蓝色*110cm	否
绒感纯棉休闲卫衣（女童） 160CM*桐花粉	否
美国VISIONS康宁晶彩汤锅炖锅0.8升+晶彩2.25升+防烫手夹	否

图 5-8（a）　案例 1 评分矩阵推荐列表

商品名称	是否被购买
3包装丨日本进口 帮宝适一级帮纸尿裤 L52片/包 空气纸尿裤婴儿透气尿不湿	否
毛毛虫儿童运动鞋（断码清仓） 夏威夷粉 *105(2.0)，脚长11~12cm	否
月白林青·全棉提花四件套 1.5m床:适用2mx2.3m被芯	否
中国电信 一元免费打半年（130） 存200打200	是
adidas kids 阿迪达斯 男童 儿童鞋 D96858 30.5 一号黑/一号黑/白	否
极简保温随行杯 珠光黑	否
羊脂玉白紫金线茶具 6件装 【优惠组】羊脂玉白茶组+竹制干泡茶盘	否
儿童防蚊T恤 4-16岁 蓝色*110cm	否
绒感纯棉休闲卫衣（女童） 160CM*桐花粉	否
美国VISIONS康宁晶彩汤锅炖锅0.8升+晶彩2.25升+防烫手夹	否

图 5-8（b）　案例 1 初始用户画像推荐列表

（2）由初始用户画像进行推荐得到的推荐列表不存在已购买的商品，对同一个用户进行用户画像调优操作之后再进行推荐，发现推荐列表出现了已购买商品。说明了采用调优的用户画像进行推荐更为有效。以用户“4cc8dd74b609456186b80fdf5594c098”为例，初始用户画像推荐列表如图 5-9（a）所示，调优用户画像推荐列表如图 5-9（b）所示。

商品名称	是否被购买
Yessing女式优雅高领款卫衣 M(165/88A)*黑色	否
网易云音乐青春系列加厚套头卫衣NO.1 粉色*L	否
振金战服 主题卫衣 漫威黑豹 XL	否
爆笑星际主题T恤-灰色款 星际争霸II XL	否
Yessing男式中长款廓形运动羽绒服 浅军绿色*XXL(185/100A)	否
星球大战 我是你爸 主题短袖T恤 XL*白色	否
粘土人 多角色 初音未来 初音未来	否
云音乐 哆啦探索·哆啦A梦T恤 条纹款 蓝色*XL码	否
陶瓷马克杯 哈士奇 新倩女幽魂 哈士奇	否
网易云音乐潮流系列高领卫衣 宝蓝色*L	否

图 5-9（a）　案例 2 初始用户画像推荐列表

商品名称	是否被购买
Yessing女式优雅高领款卫衣 M(165/88A)*黑色	否
网易云音乐青春系列加厚套头卫衣NO.1 粉色*L	否
振金战服 主题卫衣 漫威黑豹 XL	否
爆笑星际主题T恤-灰色款 星际争霸II XL	否
Yessing男式中长款廓形运动羽绒服 浅军绿色 *XXL(185/100A)	否
无限极植雅牙膏 140g*2支	是
粘土人 多角色 初音未来 初音未来	否
云音乐 哆啦探索·哆啦A梦T恤 条纹款 蓝色*XL码	否
陶瓷马克杯 哈士奇 新倩女幽魂 哈士奇	否
网易云音乐潮流系列高领卫衣 宝蓝色*L	否

图 5-9（b）　案例 2 初始用户画像推荐列表

(3)当用户更为活跃产生了更多的消费信息用于生成用户画像时，用户画像更加丰富，推荐结果也更为准确。以用户“11714985b7d544a8a5a7417cfcbf1dcb”为例，此用户用户画像较为丰富，推荐列表中成功推荐数量为 4。用户画像基本标签如图 5-10（a）所示，用户画像个性化标签如图 5-10（b）所示，调优的用户画像推荐列表如图 5-10（c）所示。

序号	用户编号	男性	女性	青年	中年	老年	低消费	中等消费	高消费	春季消费	夏季消费	秋季消费	冬季消费
1	11714985b7d544a8a5a7417cfcbf1dcb	0.1	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.2	0.0

图 5-10（a）　案例 3 用户画像基本标签

电子产品爱好者	穿搭达人	美食达人	居家能手	办公狂人	推荐列表
0.061844650449882355	0.05617102729577309	0.03922083447433899	0.16017811785631725	0.08258536992368837	普通推荐 初始画像 调优画像

图 5-10（b）　案例 3 用户画像个性化标签

商品名称	是否被购买
中国电信 前半年69（130） 20G降速（0低消 0预存 无密码号码）	是
宾迪mini电源10000毫安 蓝色	否
男式都市休闲棉服夹克 黑色*XL（180/96A）	否
Adidas男子篮球短裤 BR1951 2XL	是
【可用券】GOO.N 大王 XL28片 光羽系列 短裤式纸尿裤	否
网易智造云感磁吸手机壳 深空黑*iPhone Xs Max	否
雪麸蛋糕 香蕉牛奶风味-180克	是
adidas kids 阿迪达斯 男小童黑/藏青蓝色LOGO图案针织长裤3-11岁 104厘米 黑色	否
女式蝴蝶结方头平底鞋 粉格子*38	否
OPPO Reno 极夜黑 6G+128GB 全网通	是

图 5-10（c）　案例 3 调优用户画像的推荐列表

5.4 本章小结

本章采用软件工程的思想，从系统的设计、实现、测试三大部分对推荐系统的开发进行了介绍，并对推荐的结果进行了展示和分析。首先，介绍了系统的功能模块设计和逻辑设计，并对模块进行了编码实现和测试；然后，对推荐结果进行了展示；最后，对推荐结果进行了分析。通过推荐结果，证明了采用用户画像进行推荐的效果高于采用用户-项目评分矩阵的推荐效果，并且调优后的用户画像推荐效果比调优前好，证明了调优操作的合理性。因此，采用本文的基于用户画像的个性化推荐算法对用户进行个性化推荐，将获得更高的用户满意度。

第六章　总结与展望

6.1 全文总结

随着科技的发展，网络购物越发普及，商品种类与数量越来越繁杂，用户规模飞速扩大。为了使用户花费更少时间接触自己心仪的商品，推荐系统成为了优秀的解决方案之一。首先，从概念、构成要素、构建流程和构建方法四个层面对用户画像进行了研究，并重点研究了构建方法，根据用户的基本信息和行为信息，对用户打标签并计算标签权重，初步得到用户画像；然后，考虑到用户的评价信息为用户兴趣偏好的体现，研究了文本情感分析常用的方法，并重点研究了基于深度学习的情感分析方法，使用 BERT+Bi-LSTM+Attention 的情感分析模型对用户的评论文本进行情感分析，并对用户标签权重进行调整，初步调优用户画像；其次，考虑到用户的兴趣会随时间发生改变，研究了记忆遗忘曲线，并模拟得到兴趣遗忘曲线对用户画像再次调优；最后，对常用的推荐算法进行了研究，采用基于用户画像的推荐算法对用户进行个性化推荐。实验结果显示，本文的基于用户画像的个性化推荐算法取得了不错的效果。

6.2 不足与展望

科技不断地革新，我们即将步入 5G 时代，电商的用户和数据规模将持续扩大，电商行业也将面临新的挑战。面对海量的数据和用户，本文的方法也将面临挑战。因此，根据本文提出的方法可能存在的不足之处，提出以下四点不足以及展望：

（1）用户画像的标签数量有限，且为人为定义，一定程度上欠缺专业知识的支撑，在接下来的工作中是否有更加科学的方法定义用户画像的标签将影响着最终推荐结果的准确性。

（2）当前用户画像的更新需要间隔一定周期。由于推荐时根据的用户画像是之前计算好的，可能有用户的新的个人信息或者行为信息尚未体现在用户画像之中，不能很好地体现用户当前状态的兴趣爱好，因此，由此产生的推荐的实时性有所欠缺，没有根据所有已存在的用户数据进行推荐。随着科技的发展，在接下来的工作中是否有新的方法动态或者增量地更新用户画像将极大程度地影响着推荐效果。

（3）目前的用户画像和推荐是分开计算的，且每次计算的主体都是用户和商品的全体数据。在接下来的工作中是否有方法可以更新单一的用户画像，并将用户画像和推荐结果关联更新以保证推荐结果的实时性。

（4）目前的推荐系统没有用于生产系统中验证，仅根据历史数据验证了采用用户画像推荐的合理性以及调优操作的合理性。所以真实的推荐效果与当前的推荐效果可能会存在一定的误差。

参考文献

[1] Salminen J，Jung S G，An J，et al. Findings of a User Study of Automatically Generated Personas[C]//Proceeding of the2018 CHI Conference on Human Factors in Computing Systems. ACM，2018:1-6.

[2] Chen R，Liu J. Personas:Powerful Tool for Designers[A]//Luchs M G，Swan K S，Griffin A. Design Thinking:New Product Development Essentials from the PDMA[M]. Wiley，2016:27-40.

[3] Zhiliang Z， Deyang L， Jie L， et al. A Dynamic Personalized News Recommendation System Based on BAP User Profiling Method[J]. IEEE Access，2018:1-1.

[4] Chader A，Haddadou H，Hidouci W K. All Friends are not Equal: Weight-Aware Egocentric Network-Based User Profiling[C]// IEEE/ACS International Conference on Computer Systems & Applications. IEEE Computer Society，2017.

[5] Nguyen Phong H，Henkin Rafael，Chen Siming，Andrienko Natalia，Andrienko Gennady， Thonnard Olivier， Turkay Cagatay. VASABI: Hierarchical User Profiles for Interactive Visual User Behaviour Analytics.[J]. IEEE transactions on visualization and computer graphics，2020，26(1).

[6] 过仕明. 数字图书馆用户画像及场景重构研究[J]. 情报科学，2019，37(12):11-18.

[7] 康存辉. 基于用户画像的高校智慧图书馆服务空间再造探索[J]. 图书馆工作与研究，2020(04):79-83.

[8] 谭浩，郭雅婷. 基于大数据的用户画像构建方法与运用[J]. 包装工程，2019，40(22):95-101.

[9] 张长浩，余志勇，周振，石瑞杰，王新勇. 基于国网商旅大数据融合背景的用户画像构建[J]. 电信科学，2019，35(12):148-154.

[10] 单晓红，张晓月，刘晓燕. 基于在线评论的用户画像研究——以携程酒店为例[J]. 情报理论与实践，2018，41(04):99-104+149.

[11] 陈泽宇，黄勃. 改进词向量模型的用户画像研究[J/OL]. 计算机工程与应用 :1-6[2019-12-13].http://kns.cnki.net/kcms/detail/11.2127.TP.20190924.1504.010.html.

[12]安璐，周亦文. 恐怖事件情境下微博信息与评论用户的画像及比较[J]. 情报科学，2020，38(04):9-16.

[13]Z. Xu，O. Tifrea-Marciuska，T. Lukasiewicz，M. V. Martinez，G. I. Simari and C. Chen，"Lightweight Tag-Aware Personalized Recommendation on the Social Web Using Ontological Similarity，" in IEEE Access，vol. 6，pp. 35590-35610，2018.

[14]X. Cai，J. Han，W. Li，R. Zhang，S. Pan and L. Yang，"A Three-Layered Mutually Reinforced Model for Personalized Citation Recommendation，" in IEEE Transactions on Neural Networks and Learning Systems，vol. 29，no. 12，pp. 6026-6037，Dec. 2018.

[15]M. He，X. Wu，J. Zhang and R. Dong，"UP-TreeRec: Building dynamic user profiles tree for news recommendation，" in China Communications，vol. 16，no. 4，pp. 219-233，April 2019.

[16]J. Zhang，Y. Yang，L. Zhuo，Q. Tian and X. Liang，"Personalized Recommendation of Social Images by Constructing a User Interest Tree With Deep Features and Tag Trees，" in IEEE Transactions on Multimedia，vol. 21，no. 11，pp. 2762-2775，Nov. 2019.

[17]Z. Wu，G. Li，Q. Liu，G. Xu and E. Chen，"Covering the Sensitive Subjects to Protect Personal Privacy in Personalized Recommendation，" in IEEE Transactions on Services Computing，vol. 11，no. 3，pp. 493-506，1 May-June 2018.

[18]张兰兰. 基于关联数据的图书个性化智能推荐系统设计[J]. 现代电子技术，2019，42(23):86-90.

[19]邹洋，赵应丁，姜允志. 基于多权重相似度的随机漫步推荐算法[J/OL]. 计算机应用研究:1-5[2019-12-26]. https://doi.org/10.19734/j.issn.1001-3695.2019.08.0275.

[20]王雅青，郭彩丽，楚云霏，周洪弘，冯春燕. 面向会话型推荐系统的个性化分层循环模型[J/OL]. 北京邮电大学学报:1-6[2019-12-26]. https://doi.org/10.13190/j.jbupt.2019-143.

[21]汪涛，潘芳，潘郁，朱晓峰. 一种融合时间权重的张量分解标签推荐模型[J]. 统计与决策，2019，35(21):80-82.

[22]刘晓飞，朱斐，伏玉琛，刘全. 基于用户偏好特征挖掘的个性化推荐算法[J]. 计算机科学，2020，47(04):50-53.

[23]邓凯，黄佳进，秦进. 基于物品的统一推荐模型[J/OL]. 计算机应

用 :1-6[2019-12-26].http://kns.cnki.net/kcms/detail/51.1307.TP.20191120.1103.018.html.

[24] Cooper A. About Face: The Essentials of User Interface Design[M]. 1995.

[25] 王宪朋. 基于视频大数据的用户画像构建[J]. 电视技术，2017，41(06):20-23.

[26] 余孟杰. 产品研发中用户画像的数据模建——从具象到抽象[J]. 设计艺术研究，2014，4(06):60-64.

[27] 曾建勋. 精准服务需要用户画像[J]. 数字图书馆论坛，2017(12):1.

[28] 李映坤. 大数据背景下用户画像的统计方法实践研究[D]. 首都经济贸易大学，2016.

[29] 刘海鸥，孙晶晶，苏妍嫄，张亚明. 基于用户画像的旅游情境化推荐服务研究[J]. 情报理论与实践，2018，41(10):87-92.

[30] 宋美琦，陈烨，张瑞. 用户画像研究述评[J]. 情报科学，2019，37(04):171-177.

[31] 高广尚. 用户画像构建方法研究综述[J]. 数据分析与知识发现，2019，3(03):25-35.

[32] 牛温佳，刘吉强，石川. 用户网络行为画像：大数据中的用户网络行为画像分析与内容推荐应用[M]，北京：电子工业出版社，2016.

[33] 曾鸿，吴苏倪. 基于微博的大数据用户画像与精准营销[J]. 现代经济信息，2016(16):306-308.

[34] 郭光明. 基于社交大数据的用户信用画像方法研究[D]. 中国科学技术大学，2017.

[35] 张小可，沈文明，杜翠凤. 贝叶斯网络在用户画像构建中的研究[J]. 移动通信，2016，40(22):22-26.

[36] 汪强兵，章成志. 融合内容与用户手势行为的用户画像构建系统设计与实现[J]. 数据分析与知识发现，2017，1(02):80-86.

[37] Adomavicius，G，Tuzhilin，A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions[J]. IEEE Transactions on Knowledge & Data Engineering，17(6):734-749.

[38] Devlin J，Chang M W，Lee K，et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.

[39] 杨飘，董文永. 基于 BERT 嵌入的中文命名实体识别方法[J/OL]. 计算机工程:1-7[2019-10-21]. https://doi.org/10.19678/j.issn.1000-3428.0054272.

[40] Hochreiter S，Schmidhuber，Jürgen. Long Short-Term Memory[J]. Neural Computation，1997，9(8):1735-1780.

[41] Ebbinghaus，Hermann. Memory: A Contribution to Experimental Psychology[J].

Annals of Neurosciences，2013，20(4).

在学期间的研究成果

一、发表论文

1.方英兰，孙吉祥，韩兵. 基于 **BERT** 的文本情感分析方法的研究**[J]**. 信息技术与信息化，**2020(02):108-111.**

二、专利或软件著作权申请及授权情况

1. 计算机软件著作权. 软件名称：快递箱智能维修服务平台软件 **v1.0** 著作权人：北方工业大学，孙吉祥，方英兰.

致　谢

三年时光弹指一挥间，在北方工业大学就读研究生的三年是我学习生涯里最难忘的三年，也是我的人生中重要的三年。在这三年的时光里，我增长了自己的见识，丰富了自己的学识，结识了生命中很重要的朋友、同学、老师，他们在我成长的道路上留下了不可磨灭的印记。

首先，我要感谢的就是我的指导老师方英兰老师。从入学之初到毕业之际，方老师无论是学业上还是生活上对我关爱有加。于学业上：入学之初，对学术知之甚少，方老师耐心引我入门，让我找到了自己喜欢的学习方向。每当我迷茫时，她总是循循善诱，引导我找到前行的方向；每当我放弃之际，她总是鼓励我，让我继续鼓起勇气向前；每当我成功之时，她总是训戒我，让我能够虚心向前。是她，让我懂得了“胜不骄，败不馁”。于生活上：她对我总是嘘寒问暖，让我即使身在他乡依然能够感受到如家人一般的温馨。她亦师亦友，是我的良师益友。

然后，我要感谢实验室所有的同学。是他们让我有幸处在一个学习氛围浓厚，同学之间关系融洽的实验室。遇到困难之事，我们迎难向前，绝不退缩；遇到高兴之事，我们彼此分享，谈笑风生；遇到失落之事，我们彼此安慰，携手前行。他们是我的同学，也是我的老师，更是我的朋友。他们的存在让我遇难事不放弃，失败也不气馁。我们彼此依靠，互帮互助，共同奋斗。他们是我生命中一道亮丽的风景线。

最后，我要感谢我的家人。家人在背后一直默默的付出，他们也一直在努力，让我没有后顾之忧，全心全力放在自己的学习上，才有了我今日的成就。我也必将一直奋斗下去，用我的行动，报答他们的恩情，书写属于我自己的亮丽人生。