武漢理工大學

（申请工学硕士学位论文）

基于标签关联分析的出版资源交互式推荐方法研究

培养单位：计算机科学与技术学院

学科专业：计算机科学与技术

研 究 生：熊峰

指导教师：刘永坚

2018 年 5 月

分类号_____ 密　　级_____

UDC _____ 学校代码__**10497**__

武漢理工大學

学　位　论　文

题　目 基于标签关联分析的出版资源交互式推荐方法研究

英　文 Research on Interactive Recommendation Method for

题　目 Publishing Resource Based on Tag Association Analysis

研究生姓名 熊峰

指导教师 **姓名** 刘永坚 **职称** 教授 **学位** 硕士

单位名称 计算机科学与技术学院 **邮编** 430070

申请学位级别 硕士 **学科专业名称** 计算机科学与技术

论文提交日期 2018 年 3 月 **论文答辩日期** 2018 年 5 月

学位授予单位 武汉理工大学 **学位授予日期** _____

答辩委员会主席 熊盛武 **评阅人** 教育部学位中心盲审

教育部学位中心盲审

2018 年 5 月

独 创 性 声 明

本人声明，所呈交的论文是本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，也不包含为获得武汉理工大学或其他教育机构的学位或证书而使用过的材料。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。

签 名：熊峰 日 期：2018.05.17

学位论文使用授权书

本人完全了解武汉理工大学有关保留、使用学位论文的规定，即学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权武汉理工大学可以将本学位论文的全部内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存或汇编本学位论文。同时授权经武汉理工大学认可的国家有关机构或论文数据库使用或收录本学位论文，并向社会公众提供信息服务。

（保密的论文在解密后应遵守此规定）

研究生（签名）：熊峰 导师（签名）：刘贵德 日期 2018.05.17

摘 要

在互联网技术飞速发展的时代，对具有大量资源的出版行业来说，如何满足用户个性化的知识服务需求成为出版行业亟需解决的重要问题。本文根据出版行业资源数据的特点，设计了一个出版资源个性化推荐算法，并针对传统推荐算法中存在的评分矩阵稀疏、推荐性能随资源增加而下降、冷启动等问题进行研究。本文的主要研究工作如下：

（1）用户特征提取方法改进。本文对传统推荐算法中的用户特征提取方法进行了改进，将充分考虑用户的标签特征、行为特征和时间特征，构建用户兴趣特征向量。在定义的规范化标签体系下，通过用户对资源的行为反馈和交互操作获取用户的标签特征，并根据用户不同的行为反馈定义权重大小，同时通过时间遗忘曲线考虑用户标签特征偏移的现象。

（2）概率矩阵分解算法改进。本文针对传统概率矩阵分解算法中忽略用户和资源间的影响关系的问题进行了改进。首先通过规范化的标签对用户和资源进行特征提取，寻找用户和资源的近邻，接着将相似邻居集合融入到概率矩阵分解算法中，降低评分矩阵的稀疏性并提高算法的准确率。通过大量实验证明改进后的概率矩阵分解算法的有效性和准确性。

（3）交互式推荐框架。本文提出了一种基于标签关联分析的交互式方法，并设计了交互式推荐框架，通过用户与标签的交互缩小资源备选集，对资源进行定位，提高推荐算法的效率。在用户交互过程中，通过资源-标签矩阵对标签进行关联分析，提供较优的标签属性供用户选择，使资源备选集的划分得到优化。通过交互式实验，验证了交互式方法提高了推荐性能。

（4）冷启动问题研究。为解决推荐过程的新用户冷启动问题，本文对用户的四个基本信息属性：年龄、性别、阅读时间、阅读地方进行分析。根据非新用户的基本信息创建决策树分类器，当新用户进入系统时，通过决策树分类模型进行匹配，初始化推荐主题资源，同时在交互式框架下，通过标签交互快速获取用户的需求。在实际数据集上的实验证明了该算法的有效性和准确性。

最后，本文对基于标签关联分析的交互式推荐方法进行系统性验证，进行了需求分析，模块设计，整体框架设计和系统实现。

关键字：标签特征；概率矩阵分解；关联分析；交互式推荐；冷启动

Abstract

With the rapid development of Internet technology, how to satisfy users' personalized knowledge service needs is an urgent problem to be solved in publishing industry which has abundant resource. In this thesis, a personalized recommendation algorithm for publishing resources is particularly designed based on the resource characteristics of the publishing industry, at the same time we have researched the problems in the recommendation algorithm, such as score matrix sparsity, recommendation performance decrease with the increase of resources, cold start and so on.The main contents of this thesis are as follows:

(1) Improvement of user feature extraction method. In this thesis, we improve the users' feature extraction method in traditional recommendation algorithm, with full consideration of users' tag characteristics, behavior characteristics and time characteristics, and build the users' interest characteristic vector. Based on the standardized tag, we get users' tag characteristics through users' action feedback and interaction operation on the resources, and define the tag weights by different users' behavior feedback, meanwhile through the time forgetting curve consider users' tag feature offset.

(2) Improvement of probabilistic matrix factorization algorithm. In this thesis, the problem of neglecting the relationship between users and resources in the traditional probabilistic matrix factorization algorithm is improved. First, we get the user and resource feature by using standardized tag and find the neighbor of users and resources, then we integrate the relationship of the similar neighbor sets into the probabilistic matrix factorization to reduce the sparsity of the scoring matrix and improve the accuracy of the algorithm. The validity and accuracy of the improved probabilistic matrix factorization are proved by a large number of experiments.

(3) Interactive recommendation framework. In this thesis, an interactive method based on tag association analysis is proposed, and we have designed an interactive recommendation framework, which can reduce the resource selection set by user interaction, locate the recommendation results, and improve the efficiency of the recommendation algorithm. In the process of user interaction, we analyze the tag

through the resource-tag matrix, provide better tag attributes for users to choose and optimize the dividion of resource alternatives set. Interactive experiments demonstrate that the interactive approach improves the performance of the recommendation.

(4) Research on cold start problem. In order to solve the cold start problem of new users in recommendation process, in this thesis we analyze four basic information attributes of users: age, gender, reading time and reading place. A decision tree classifier is created for the basic information of users. When new users enter the system, matching the decision tree classification model, initializing the recommended topic resources. Meanwhile, in the interactive framework, we can quickly get users' needs through tag interaction. The experiment based on the actual dataset proves the validity and accuracy of the algorithm.

Finally, an interactive recommendation method based on tag association analysis is verified in this thesis, at the same time, we have carried out the requirement analysis, the module design, the frame design and the system implementation.

Keywords: Tag Feature; Probabilistic Matrix Factorization; Association Analysis; Interactive Recommendation; Cold Start

目 录

第 1 章 绪论

1.1 研究背景及意义

1.1.1 研究背景

互联网的飞速发展在不断地改变我们的生活方式，同时也带来了网络信息数据的剧增，使人们需要投入大量的时间和精力获取自己所需的资源。从 2010 年到 2016 年，数字出版行业在我国蓬勃发展，市场总产值也在逐年增多。电子书产品由最初的不到 100 万种，在五年内已增长了三倍左右，同时各种出版资源也都在高速的增长。大众已经慢慢开始接受数字阅读方式，到 2016 年，数字阅读率已经领先于纸质阅读率 15.2%[1]。由于人们的阅读方式逐渐从纸质书籍转向数字化，传统的阅读方式已经不能满足广大读者的需求。在这个科技日新月异的时代，传统出版行业面临着巨大的挑战，激励着传统出版行业向数字出版行业的转型。

数字出版建立在计算机和通信等新兴技术的平台上，它将传统的出版内容和新兴的出版资源进行了有效的结合。本课题研究内容中的出版资源是对出版行业所具有电子资源的统称，其资源一般包含视频、音频、图片、PDF、TXT 文档等，它的特点是种类繁多和特征不明显，因此无法采用一般的特征提取方法来获取资源的关键信息。规范化标签则是出版资源的一大特点，通过资源创建者或者资源审核者给资源定义合适的标签，规范化标签可以很好的描述资源的特征。

在海量信息中获取用户所需信息成为难题时，个性化推荐服务逐渐进入人们的视野，受到众多学者的重视。在 2000 年 ACM（Association for Computing Machinery）国际计算机学会为个性化推荐研究专门建刊，标志着推荐算法研究进入高速发展阶段。目前个性化研究在实际应用中取得了巨大成功，比如 Amazon 在其电子商务上应用个性化推荐算法，据统计为其带来了 10%到 30%的附加利润[2]。另外淘宝、京东等大型电商也都通过推荐算法给用户推荐潜在需求的商品，提高平台收益。在大量的出版资源中，如何寻找用户所需要的信息成为一个难题，如何提升用户粘性，减少客户流失的现状，本文根据出版行业的资源特点，设计了一个出版资源个性化推荐算法，为广大用户推荐出版资源。

1.1.2 研究意义

出版资源的个性化推荐研究具有重要的理论意义和实际应用意义。在理论研究方面，如今信息处于高速增长阶段，用户每年产生的数据在现有的基础上还要翻倍，信息过载的现象越来越严重，并且硬件的发展速度也已经跟不上信息增长的速度，如何在大量出版资源中选择适合读者的资源已成为亟需解决的问题。用户通过利用移动设备和媒体通道，可以选择合适的资源进行浏览或分享，但与此同时会产生用户粘性不高，信息的使用率较低的问题。社交化、数据化、个性化的互联网形式已开始成为互联网数据公司的发展方向，与此同时用户也急需有效的个性化推荐服务来满足自己的需求，所以推荐算法是我们需要持续不断研究的一个课题。

在实际应用方面，目前个性化推荐算法已在各领域被广泛的使用。个性化推荐最早出现在 B2C 电子商务领域，它的原始目的是将用户潜在需求商品推荐给用户，提高用户的购买欲，从而增加商户的实际收益。出版行业的主要收入来源于用户的文化消费，通过为用户提供知识服务获取利益回报，传统的出版行业已经很难满足用户对知识的快速准确的获取需求。今日头条的成功为出版行业提供了很好的发展方向，通过实现出版资源的个性化推荐提高用户粘性和企业收益，因此出版行业的个性化推荐研究具有很重要的应用意义。

本文基于出版资源具有规范化的标签特征，进行了基于标签关联分析的出版资源交互式推荐方法研究，旨在为出版行业设计一个出版资源个性化推荐算法，并针对推荐算法中存在的评分矩阵稀疏、推荐性能随系统资源增加而下降、冷启动等问题进行研究。

1.2 国内外研究现状

1995 年 3 月，卡耐基梅隆大学首次在美国人工智能协会提出了个性化导航系统。在此以后个性化推荐技术受到了国内外研究人员的广泛关注，并逐渐应用到各个行业。在电子政务领域，Meo 等人[3]介绍了一种多代理的电子政务系统，在复杂的电子政务场景中构建新的系统帮助政府机构的决策者制定新的服务。在线上教育领域，Capuano 采用了一种自适应的学习推荐系统[4]，它根据用户个人资料以及学习目标自动生成个性化的学习方法。在线上资源领域方面，Nguyen 等人[5]为了提供更好的网页推荐，通过增强语义、整合领域和网址知识来构建新

的模型。国外很多学者对现有推荐算法的优缺点都进行了深入研究，也提出了相关建议。

国内学者和互联网企业对于个性化的推荐研究也都取得了不错的成绩。在电子商务领域中以京东和阿里等公司为代表，其中京东的推荐系统采用混合推荐模型算法，将用户模型、商品模型等进行综合考虑，进而为用户进行商品推荐。阿里则将推荐算法进行商业化，建立混合推荐算法并推出了阿里云推荐系统，为广大中小企业提供推荐服务，企业只需要导入相应的数据模型就可以为用户进行个性化推荐。在新闻资源推荐方面，今日头条提供的推荐算法在效率和准确率上得到很大提升。今日头条采用的推荐系统能在 0.1 秒内计算出推荐结果，而完成文章特征提取和分类只需 3 秒，推荐高效准确。

1.2.1 传统推荐算法

目前推荐算法大致可以分为：基于关联规则的推荐算法[6-7]、基于内容的推荐算法[8]、协同过滤推荐算法[9]、混合式推荐算法[10]。基于关联规则的推荐算法通过分析用户和项目信息，挖掘出用户的相关模型，并在此基础上联合用户的历史信息构建推荐规则，最后完成对目标用户的推荐。最早的关联分析方法，是在 1993 年由 Imielienskin 等人[11]首次提出。随着对关联分析的研究，它被应用在各领域，著名的经典案例“啤酒和尿布”就是通过发现项目与项目之间的关系，从而提高了商城的销量。基于内容的推荐算法是以产品信息为研究对象，不依赖用户对项目的显示评价，因此避免了推荐中常见用户冷启动问题和数据稀疏的问题，但存在对种类繁多的资源进行特征提取困难。混合推荐算法主要是为了解决单一推荐在实际应用过程中的缺陷，目前已在国内外得到了广泛的使用。比如文俊浩等[12]提出的基于社交网络用户信任度的混合推荐方法，解决基于用户信任推荐算法中覆盖率较低的问题。黄璐等人[13]提出的融合主题模型和协同过滤的多样化移动应用推荐，解决了推荐过程中长尾应用被淹没的问题。

本文在协同过滤算法基础上进行了改进，这里详细介绍协同过滤算法的研究现状。Zhang J 等人[14]为解决推荐算法中数据稀疏问题，提出基于用户偏好聚类的协同过滤算法。Zhang Z 等人[15]针对协同过滤算法中用户邻居选取的方法进行改进，通过该方法可以在稀疏数据集上取得较好效果。印鉴等人[16]通过利用大规模隐式反馈数据进行个性化推荐，通过将推荐任务转换成选择行为发生概率的优化问题，解决了隐式反馈中只有正反馈而缺少负反馈的问题。黄正华等

人[17]认为项目间的排序不应该只是根据项目的评分大小进行推荐，他们将排序方法用于推荐算法中，解决了传统推荐算法中普遍存在的排序问题，提高了算法性能。于洪等人[18]详细介绍了用户时间权重的概念，综合考虑用户时间与项目时间等因素，解决了项目冷启动的问题。

协同过滤算法虽然获取了巨大的成功，但仍然存在很多问题。比如忽略了用户与资源之间的结构关系以及新用户冷启动问题等[19-22]。如何有效的运用用户和资源的结构关系来丰富单个用户和资源的信息，从而准确地识别用户的个人兴趣。许多学者[23-24]以此为出发点，提出了基于用户的社交关系，用户的信任关系，对象间关联关系等方法来寻找用户或资源间的结构关系。这些方法主要通过用户对资源的评价训练出推荐模型，然后根据模型进行推荐。

根据协同过滤算法中使用模型的不同，可以分为：聚类模型[25]，图模型[26]，贝叶斯层次模型[27-28]等。由 Salakhutdinov 等人[29]提出了利用低维近似矩阵分解模型进行推荐的推荐算法(probabilistic matrix factorization，简称 PMF)。该算法将用户和资源信息映射到低维的特征空间中，通过用户-资源的评分矩阵获取用户和资源的特征向量，然后重构用户-资源评分矩阵，最后为用户推荐相关资源。但该算法中只利用了用户的评分矩阵信息，没有考虑用户和资源之间的影响关系，因此存在着精度有待提升的问题。孙光福等人[30]在 PMF 算法的基础上进行了改进，对用户购买资源时的时序行为进行考虑，通过时序行为寻找用户和资源近邻关系，并与 PMF 算法结合，从而提高推荐算法的准确率。杨阳等人[31]为解决协同过滤算法中的矩阵稀疏和新用户冷启动问题，提出了基于矩阵分解和近邻用户模型的推荐方法。该算法在构建用户邻居模型时引入了奇异值矩阵分解，解决了矩阵稀疏问题，同时对新用户构建了用户邻居，解决了新用户冷启动问题。本文在概率矩阵分解模型上通过规范化标签考虑用户和资源的影响关系，通过标签信息进一步挖掘用户和资源的结构关系，然后将获取的邻居关系与概率矩阵分解算法结合，对其进行了改进。

1.2.2 基于标签的推荐算法

随着 Web2.0 的发展，大量的资源被用户发布到网络上从而产生了大量的标签信息，比如豆瓣和 MovieLens。基于标签的推荐算法通过标签挖掘用户和资源的联系，将用户和资源的关系转换成了用户-标签-资源的三维关系。基于标签的推荐算法根据推荐目标的不同可以分为三种形式：为用户推荐标签、根据标签

特征为用户推荐资源、根据标签寻找相似用户进而推荐用户。

从目前基于标签推荐算法的发展趋势来看，它具有很好的研究前景。Song等人[32]为解决标签推荐的问题，从机器学习的角度进行考虑，改善了推荐的质量。Hotho 等人[33]将用户、资源、标签三者之间的联系看成一个无向图，根据图对标签进行排序，用户通过推荐的标签得到符合自己的资源。宋伟伟等人[34]在标签特征的基础上考虑时间权重，考虑用户兴趣的长期和近期影响，提高了算法的准确率。针对社会化标签推荐中忽略用户兴趣变化和反复性的问题，张艳梅[35]等人在分析反复出现的早期数据时考虑近期信息的影响程度，并提出了遗忘权重和时间窗口结合的算法。该算法首先需要建立基准标签集，由数据偏移后的标签向量选择目标用户的最近邻居，接着通过该用户时间窗内的资源获取其它所有资源的推荐权重向量，最后综合考虑推荐权重和资源相似度给出推荐结果。孔欣欣等人[36]采用了标签权重模型进行资源推荐，分析用户对资源的需求，通过对标签权重数据处理完成后进行推荐。Hao 等人[37]针对推荐算法中数据稀疏问题，提出了基于标签的跨域推荐模型，通过此模型在公共数据集的实验结果证明了方法的可行性。基于出版资源具有规范化的标签特点，该标签更能准确的描述资源特征，因此本文将基于标签进行推荐算法研究。

1.2.3 交互式推荐算法

Ali等人[38]提出了一种交互式方法，算法是通过整合偏好配置文件、标签信息和用户的交互式信息来实现的，它利用推荐结果与新用户进行交互，交互的过程可以采用显式或隐式的交互，通过交互获取信息为新用户提供推荐列表，使算法的效率和精度得到了提升。Nepal 等人[39]提出以用户隐式交互为基础的内容推荐模型，通过用户隐式交互确定信任关系，进而为用户进行推荐，在实验中也取得了较好的推荐结果。Kuramoto 等人[40]提出基于用户模糊意图的交互式推荐模型，它提供一个选择意见的界面，一旦用户选择了任何建议，那么系统将会给出相关的推荐。Liu 等人[41]提出了利用隐式交互信息构建社会用户图形的个性化推荐算法，首先衡量用户之间隐式交互信息的相似度，然后将用户划分归类，最后对用户进行推荐。学者们虽然针对不同的应用提出了交互式推荐，但是没有对交互的过程和划分方法给出详细的说明。本文将基于标签关联分析进行交互，并详细的描述交互过程和框架，通过对资源备选集进行合理划分，提高推荐效率和质量。

1.3 研究内容

本文旨在对出版资源进行个性化推荐研究，并针对概率矩阵分解算法中忽略用户和资源的影响关系、冷启动问题、资源量增加导致推荐算法效率下降等问题进行研究。本文的主要工作如下：

（1）用户和资源特征提取方法改进

通过对出版资源特点的分析，给出了规范化标签的定义。通过用户-标签-资源的三维关系对用户和资源的标签特征表示，构建了用户-资源评分矩阵、用户-标签偏好矩阵、资源-标签特征矩阵。本文对传统的用户特征提取方法进行改进，采用 TF-IDF 算法获取用户的标签特征，然后将用户标签特征、行为特征、时间特征进行综合考虑，构建用户兴趣特征向量。

（2）概率矩阵分解算法改进

传统的概率矩阵分解模型只考虑了用户-资源评分矩阵，而忽略了用户和资源间的影响关系。本文通过规范化的标签寻找用户的兴趣特征，进而寻找出用户和资源的近邻，将近邻关系应用到概率矩阵分解模型中，充分考虑自身特征和邻居的影响，完成对评分矩阵重构，提高推荐算法的准确率。通过在真实数据集上的大量实验证明了改进后算法的有效性和准确性。

（3）交互式推荐框架

随着出版资源的增加，推荐算法的效率往往会降低，针对该问题提出了基于标签关联分析的交互式方法。通过资源-标签矩阵对标签进行关联分析，提供较优的标签属性供用户选择，然后通过用户与标签的交互获取用户需求，缩小资源备选集，完成对资源的定位，从而提高推荐算法的效率。通过在实际数据集上的实验证明了交互式方法提高了推荐效率。

（4）推荐算法冷启动问题研究

本文针对协同过滤算法中存在的冷启动问题进行了分析，并给出了新用户冷启动问题的解决策略。通过对用户的四个基本信息属性进行分析，构建非新用户的决策树分类器，当新用户进入系统时，通过决策树分类模型进行匹配，推荐用户感兴趣的主题资源。同时用户在交互式框架下，可以快速获取所需资源。在实际数据集上的实验证明了本文冷启动解决策略的有效性和准确性。

（5）出版资源推荐可行性验证

本文根据出版资源的特点提出了基于标签关联分析的交互式推荐算法，分别针对改进的概率矩阵分解算法、冷启动算法和交互式推荐进行了实验验证。

同时给出了系统验证，对交互式系统进行了需求分析和设计，并进行了系统的展示。

如图 1-1 所示为本文的主要工作路线图。

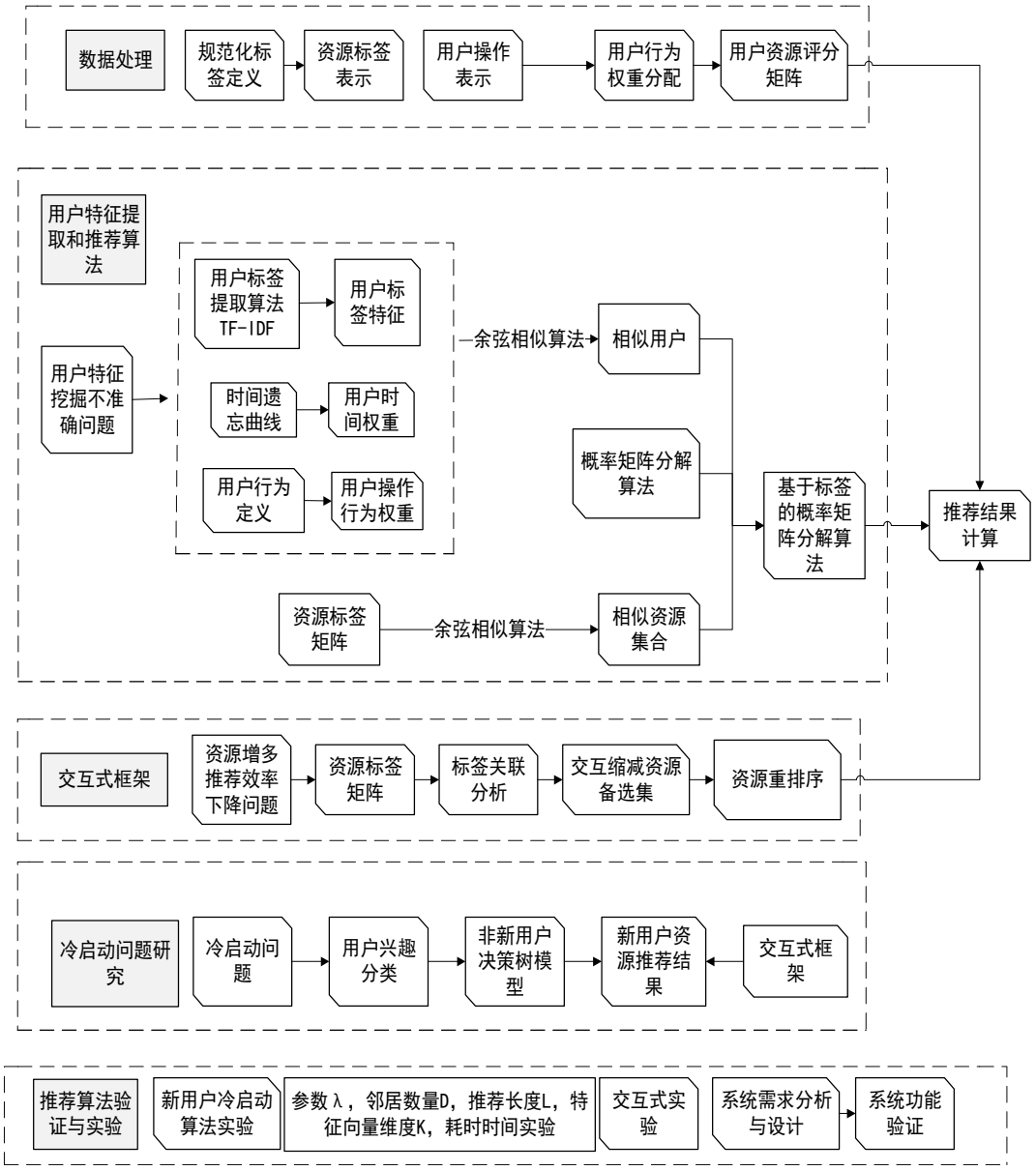


图 1-1 工作路线图

1.4 论文结构安排

基于以上研究内容，本文分为六个章节，具体章节内容如下：

第 1 章 绪论。首先介绍本文的研究背景和研究意义，对出版资源特点进行了说明，然后分析本文相关的国内外研究现状，具体分析了传统推荐算法、基于标签的推荐算法和交互式推荐算法的研究现状，最后介绍本文的研究内容。

第 2 章 基于标签关联分析的交互式推荐方法分析与设计。首先对本文的项目背景和出版资源规范化标签进行定义，然后设计了本文基于标签关联分析的交互式推荐的整体框架，对交互式推荐的两个主要研究重点进行分析，本文的后续章节也将围绕推荐算法和交互式两部分进行展开研究。最后对整个推荐系统的数据进行处理和表示，将用户、资源、标签间的关系通过二维矩阵进行表示，通过用户对资源的操作权重构建了隐式的用户-资源评分矩阵，接着构建了资源-标签矩阵和表示用户喜好的用户-标签矩阵。

第 3 章 基于标签的概率矩阵分解算法改进。首先针对传统推荐算法中特征提取方法进行改进，采用 TF-IDF 算法获取用户的标签特征，然后通过用户的行为特征和时间属性综合考虑用户兴趣特征。接着介绍了传统概率矩阵分解算法，并分析了其优缺点。针对概率矩阵分解算法忽略用户和资源的影响关系，本文进行了改进，通过改进的特征提取算法挖掘用户和资源的近邻，并将近邻关系融入到概率矩阵分解模型中，然后进行了公式推导和算法的流程介绍。最后进行了实验设计与结果分析，通过调节参数和对比实验证明了基于标签的概率矩阵分解推荐算法的有效性和准确度。

第 4 章 基于标签关联分析的交互式方法研究。首先对交互式方法进行详细的描述，然后基于资源-标签矩阵对标签进行关联分析，提供较优的标签属性供用户选择，使资源备选集的划分得到优化。同时针对概率矩阵分解过程中用户冷启动问题进行了分析，提出了决策树分类的交互式解决方案。最后对新用户冷启动和交互操作进行了实验验证，结果表明本文提出的新用户冷启动解决策略和交互方法的有效性。

第 5 章 基于标签的交互式推荐系统。首先基于本文的项目背景进行了需求分析，并对整个整体框架和推荐流程进行设计。然后分析了推荐系统各模块的具体功能，最后对系统主要界面进行展示和说明。

第 6 章 总结与展望。对本文的研究工作进行了全面的总结，并分析成果中需要进一步完善的内容。

第 2 章 基于标签关联分析的交互式推荐方法分析与设计

本文针对出版行业的资源特点，提出了基于标签关联分析的交互式推荐方法。为了更好的对交互式推荐方法进行说明，本章首先对交互式推荐方法的整体框架进行设计，分析了交互式推荐的两个主要步骤，在后续章节中，本文将围绕推荐算法和交互过程这两点进行详细的论述和研究。然后对交互式系统中的数据进行处理和表示，主要包括用户评分处理、资源特征表示和用户喜好度。

2.1 交互式推荐方法框架设计

本文的研究内容是依托于国家文化产业专项资金项目：数字内容资源知识服务技术支撑平台，该项目旨在帮助出版行业加快数字化转型进程。为了更充分的利用出版资源，加速出版行业数字化转型，提高出版行业的知识服务，本文采用基于标签关联分析的交互式推荐算法为用户推荐出版资源，为其提供更好的个性化知识服务。

本文研究的出版资源是对出版社所拥有的电子资源的简称，其资源一般包含视频、音频、图片、PDF、TXT 文档等，它的特点是种类繁多和特征不明显，因此无法采用一般的特征提取方法来获取资源的关键信息。规范化标签则是出版资源的一大特点，通过资源创建者或者资源审核者给资源定义合适的标签。规范化标签可以更好的描述资源的特征，同时相比于普通用户在网络上自由添加的社会化标签具有较高的准确性和严谨性。

在基于标签的推荐算法中，通常使用的是社会化标签。但传统的社会化标签中存在语义相当、拼写错误、数据冗余和语义表达不清等问题。例如“成龙”和“成龙大哥”语义相当；“笔记本”和“苹果”的语义表达不清导致理解可能出现偏差。因此本课题研究的推荐算法采用的是规范化的标签体系，它能有效的避免基于社会化标签推荐中常见的问题，如资源标签稀疏、标签语义模糊、标签冗余等。在出版社所拥有的资源集中，每个资源所具有的规范化标签相对于传统的标签数据显得更加密集，因此本文基于规范化的标签进行研究。

随着出版资源数量的剧增，如何满足用户的知识服务需求，提高传统推荐算法的效率和质量。基于出版行业资源规范化标签的特点，当用户对推荐的资

源不满意时，本文通过系统与用户的交互过程来解决此问题。因此本文的研究重点将分为推荐算法和交互式方法，首先通过推荐算法给用户推荐资源，当推荐结果不满足用户需求时通过用户的交互，系统缩减资源备选集定位用户的需求，同时在用户交互的过程中完善用户的特征模型。基于此思路，本节对基于标签关联分析的交互式系统进行整体框架设计，如图 2-1 给出了整个算法的整体框架。

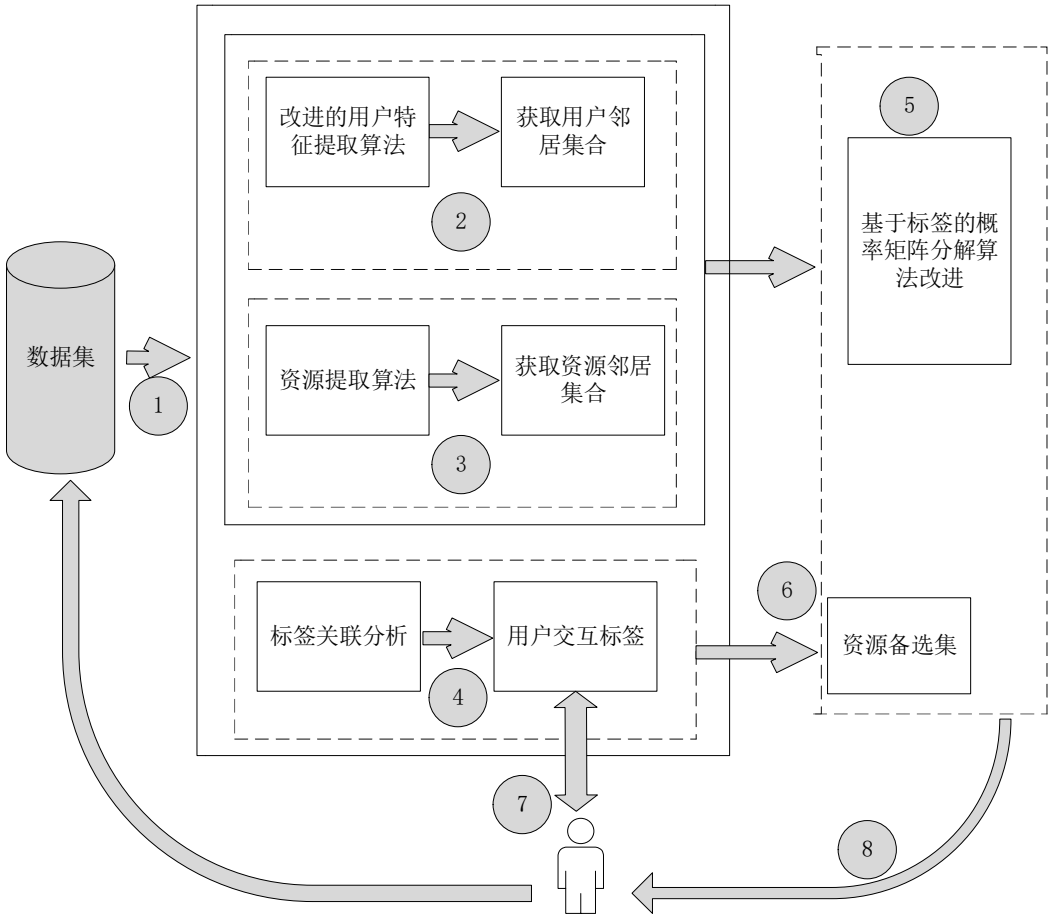


图 2-1 基于标签关联分析的交互式推荐框架

通过本文设计的基于标签关联分析的交互式框架，可以发现本文研究的交互式推荐方法的重点是资源推荐和交互过程两部分，其中推荐算法的优劣决定了推荐结果的质量，交互式则对资源备选集进行划分，快速获取用户的需求，提高整体推荐的效率。围绕推荐算法和交互式两大模块，整个算法细分为八个步骤，具体步骤如下：

（1）读入数据，其中数据包含用户操作行为、标签信息、资源标签信息和用户资源评分信息，然后进行数据处理。

（2）根据处理后的数据，采用改进的用户特征提取算法，构建用户特征模型，然后计算用户的相似邻居。

（3）通过对资源-标签矩阵的分析，构建资源特征模型，然后计算资源的相似邻居。

（4）根据资源-标签矩阵分析标签的关联关系，寻找满足最小支持度和置信度的标签关系。

（5）将用户和资源邻居关系融入到概率矩阵分解模型中，改进概率矩阵分解算法，通过基于标签的概率矩阵分解算法给出推荐资源列表。

（6）根据标签的关联分析对资源备选集进行划分，当用户对标签进行交互后，通过标签的关联分析缩减用户的资源备选集。

（7）用户通过系统提供的标签进行交互，交互过程系统根据标签关联分析提供标签选项同时系统获取用户的标签反馈。

（8）若用户没有进行交互操作，通过基于标签的概率矩阵分解算法获取推荐资源列表。若用户采用交互方式，则将对资源备选集进行划分，然后进行重排序获取推荐列表。

本节对基于标签关联分析的交互式推荐方法进行整体框架设计，并分析了整个框架的算法步骤，本文后续章节将对交互式推荐方法的两个关键部分资源推荐和交互式方法进行详细的研究，至于算法具体的八个步骤也将在这两部分中进行详细说明。

2.2 评分处理和数据表示

通过对基于标签关联分析的交互式框架设计和分析后，本节将对算法中的数据进行处理和表示，构建用户-资源评分矩阵、资源-标签特征矩阵、用户-标签的交互操作矩阵和用户喜好度的用户-标签矩阵，为下文推荐算法做准备。我们假设规范化的标签库为$T=\{t_1,t_2,\ldots,t_l\}$，标签的大小为 L。推荐系统中用户集合表示为$U=\{u_1,u_2,\ldots,u_N\}$，N 表示用户总数。资源集合表示为$I=\{i_1,i_2,\ldots,i_M\}$，M 表示资源总数。本文定义用户的操作集合表示为$O=\{o_1,o_2,\cdots,o_j,\cdots,o_F\}$，其中 F 表示操作种类大小，对每个操作o_i分配一个权重$w_i\left(1\le i\le F\right)$来表示操作的重要程度，也是本文定义的评分权重大小。操作权重越大表示用户对该资源的

喜爱程度越大，对资源的评分也越高。

2.2.1 用户评分处理

在推荐算法中用户-资源的评分高低代表用户的偏好兴趣，然而用户对于评分操作往往感到厌烦，导致评分矩阵过于稀疏或出现评分质量较低等问题。本文将通过用户的隐式评分来解决此问题，将用户对资源的行为操作权重构成用户-资源评分矩阵。用户的隐式反馈能较为直接的反映出用户的偏好，同时解决了用户进行具体打分的繁琐操作。本文数据集中的用户行为来源于用户在 RAYS 蓝海系统上的行为操作，系统记录有用户阅读时长超过一分钟、收藏、评论和购买等操作。系统将会收集用户的历史信息，并对其进行处理，收集的信息主要有用户标识，资源标识，操作时间，操作类型等。例如表 2-1 记录了某一时间段，用户标识为 135272 的用户阅读了资源标识为 256 的资源，并对资源表示为 368 的资源进行了购买，同时记录用户对标签标识为 297 的标签交互选择。

表 2-1 用户的部分操作记录

userId	itemId	tagId	createTime	type
135272	256		2017-03-17	read
135272	368		2017-03-25	buy
135272		297	2017-04-01	select
135272	1352		2017-04-01	favorite

因为用户对资源进行的是隐式操作，需要判断哪些具体的操作是有效的，以及操作权重的大小，因此需要定义明确的操作集合，如表 2-2 所示。

表 2-2 用户的行为和权重分布

用户行为	行为描述	权重值
浏览超过 1 分钟	o_1	w_1
评论	o_2	w_2
交互操作	o_3	w_3
收藏	o_4	w_4
分享	o_5	w_5
购买	o_6	w_6

用户的行为权重定义完成后，将对收集的用户数据进行处理，获取用户-资源评分矩阵，将其表示为一个 $N \times M$ 矩阵。

$$R_{N\times M}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{pmatrix} \tag{2-1}$$

用户的行为操作权重往往能代表用户的兴趣偏好，其隐式评分对用户乱评分和评分稀疏的问题进行了有效解决。其中，r_{ij} 表示用户 u_i 对资源 i_j 的操作权重即评分。如果用户 u_i 没有对资源 i_j 做出过操作，则 $r_{ij}=0$。在这种情况下，评分越高表示用户越喜欢该资源。

本文对用户和规范化标签的交互进行处理，用户对标签的交互通过用户-标签交互矩阵表示，矩阵大小为 $N \times L$。

$$S_{N\times L}=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1l} \\ s_{21} & s_{22} & \cdots & s_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nl} \end{pmatrix} \tag{2-2}$$

其中，s_{ij} 大小表示用户 u_i 与标签 t_j 的交互选择次数，若没有交互则 $s_{ij}=0$。

2.2.2 资源标签表示

本文的资源即是出版企业所拥有的电子资源，其种类较多并且特征难以提取。规范化的标签能准确的表示资源的特征属性，因此本文的资源采用基于规范化标签进行描述。将任意资源通过一维的标签向量进行表示，例如资源 i_j 的标签特征可以表示为 $i_j=\left(p_{j1},p_{j2},\cdots,p_{jl}\right)$，通过该方法可以很好的表示资源的特征，同时方便寻找相似资源。系统通过对所有资源进行数据处理，获得资源-标签矩阵，将其表示为一个 $M \times L$ 矩阵，具体矩阵如下：

$$p_{M\times L}=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1l} \\ p_{21} & p_{22} & \cdots & p_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{ml} \end{pmatrix} \tag{2-3}$$

其中，若 $p_{m\times l}=1$ 则表示资源 i_m 上标注有标签 t_l，若 $p_{m\times l}=0$ 表示资源 i_m 没有被标签 t_l 标注。通过资源-标签矩阵可以较为方便的发现资源的特征，同时标签

的关联关系也将通过资源-标签矩阵进行分析。

2.2.3 用户喜好度表示

用户一般只会对感兴趣的内容进行操作，用户对标签的交互也说明了用户的偏好，因此本文在考虑用户兴趣偏好时，综合了用户标签特征、时间权重和行为操作。规范化的标签对资源描述较为准确，因此用户获取的标签也更能描述用户的特征，同时根据用户的操作行为对标签进行加权，综合考虑用户标签特征。已知资源集合为 I，用户的操作集合为 F 和标签集合 T，则用户 u 的兴趣模型可以通过对资源的操作和标签的交互获取，如图 2-2 所示。

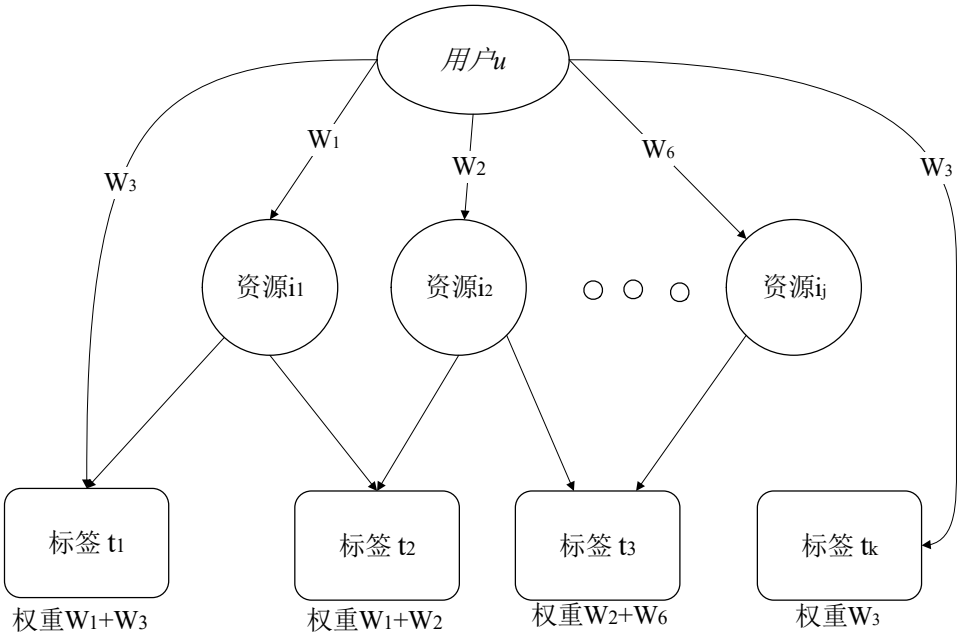


图 2-2 用户兴趣模型

通过图可以形象的表示用户 u 的具体操作，表示出用户的兴趣偏好。用户 u 对资源 i_1 的操作为 o_1，通过上文定义的权重值可以表示为 w_1，资源 i_1 具有的标签为 t_1 和 t_2，同时用户对标签 t_1 进行了交互操作，因此用户对标签 t_1 的喜好度可以表示为 $w_1 + w_3$。

用户对资源和标签的操作可以表示为用户的偏好，因此将根据式(2-1)、式(2-2)和式(2-3)的 $R_{N\times M}$、$S_{N\times L}$ 和 $p_{M\times L}$ 构建用户对标签的偏好关系。用户对标签的偏好是通过用户对资源的操作加权与用户交互标签加权的和，反馈出用户对该标签的喜爱程度，更能表达用户的标签特征，建立用户-标签偏好特征矩阵 $F_{N\times L}$ 如

式(2-4)所示。

$$F_{N\times L}=\begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1l} \\ g_{21} & g_{22} & \cdots & g_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nl} \end{pmatrix} \tag{2-4}$$

其中，N 为用户数量，L 为标签库数量，g_{ik} 表示用户 u_i 对标签 t_k 的喜好值，按照如下公式(2-5)进行计算。

$$g_{ik}=\sum_{j=1}^{m} r_{ij}p_{jk}+s_{ik}w_3 \tag{2-5}$$

其中，$\sum_{j=1}^{m} r_{ij}p_{jk}$ 是用户 u_i 在相关资源的标签 t_k 上进行累加的结果，它表示用户在操作不同资源时都会被隐式的贴上相同的标签 t_k，通过该标签可以反映出用户的标签特征，这里采用加权是为了更准确的表示用户的偏好，使标签的区分度更强。$s_{ik}w_3$ 表示用户在系统中与标签的交互权重，其结果是选择权重和选择次数的加权，也侧面反映出用户的兴趣特征。因此本文将两者进行求和，突出用户的标签偏好特征。

2.3 本章小结

本章首先对出版资源的特点进行了介绍，然后对基于标签关联分析的交互式推荐方法进行整体框架设计，分析了算法两个关键部分，资源推荐部分和交互式方法部分，本文后面章节也将对这两部分进行展开研究。

然后对本文的数据进行处理表示，为后文的推荐算法准备。将用户、资源、标签间的关系通过矩阵关系进行表示，首先根据用户对资源的操作权重构建了隐式的用户-资源评分矩阵，接着通过规范化的标签构建了资源-标签矩阵，最后通过分析用户对资源的操作构建了用户-标签偏好矩阵。

第 3 章 基于标签的概率矩阵分解算法改进

上一章对基于标签关联分析的交互式推荐方法进行了分析和设计，本章将针对其中的一个关键问题，推荐算法进行详细研究。推荐算法中用户特征提取的准确性会对推荐结果产生较大影响，本章首先对传统推荐算法中的特征提取方法进行改进，充分考虑用户的标签特征、行为操作、时间权重等因素，构建用户特征向量。然后针对概率矩阵分解算法中忽略用户和资源间关系问题进行研究，通过改进的特征提取算法获取用户和资源邻居集合，将用户和资源的近邻关系融入到概率矩阵分解模型中，对概率矩阵分解算法进行改进，并给出了公式推导过程和算法流程，提高推荐算法的质量。最后通过实验验证了基于标签的概率矩阵分解算法改进的有效性和准确性。

3.1 基于标签的特征提取方法改进

推荐算法的目的是给目标用户推荐感兴趣的个性化资源，因此获取用户的特征和相似用户就显得尤为重要了。传统的基于标签的协同过滤算法中往往只简单考虑用户的标签特征，忽略了用户的行为操作和时间权重，导致用户特征提取不够精确。本文将从用户的标签特征，操作行为和时间权重综合考虑，更加准确的描述出用户的兴趣特征，解决传统推荐算法中用户兴趣挖掘不准确的问题。针对出版资源资源特征提取困难的问题，本文则通过定义的规范化标签有效的解决了该问题。

3.1.1 用户特征提取方法改进

传统的基于标签推荐算法只简单通过标签进行用户特征提取，导致用户兴趣特征描述不够准确。本文将从用户的标签特征、用户时间特征、用户行为特征三方面进行综合考虑，对用户进行特征提取具体方法如下：

（1）用户的标签特征

本文将通过 TF-IDF 算法对用户标签特征进行提取。TF-IDF 算法是数据挖掘中常用的算法，它的核心是在一段文字中寻找出频繁出现的关键字，而寻找的关键字在又很少在其它文章中出现。通过该关键字可以很好的区分该文章和其它文章。因此本文将通过 TF-IDF 算法寻找出用户的关键标签特性，它也能充分

表明该标签对用户的影响程度。

假设一篇文章出现较多的三个关键字分别是：“学习”、“的”、“态度”，并且它们在 1000 字的文章中出现的次数分别是 20、120、45 次。根据普遍的相关性计算方法，计算出关键字“的”的频率为 0.12，关键字“的”的作用最为明显，但是在实际应用中关键字“的”不能表示出文章的特征，因此可以看出一般根据总频率计算的相关性并不准确。在实际应用中，如果一个词只出在特定的文章中出现，那么这个单词对于该文章就具有很强的说明性，通过该关键字可以明显的说明文章特征，因此对于用户标签特征的提取也应反馈出用户的特有的特征。例如词 k_i 出现在 W 篇文章中，当 W 越大的时候，关键词 k_i 的权重将越小，当所有文章都出现了关键词 k_i，则关键字将不能代表该文章特有的特征。

因此本文将采用 TF-IDF 算法采对用户的标签特征进行提取，提取出用户特有的标签特征，如公式(3-1)所示。

$$F_{ut_k}^{t}=\frac{c_{ut_k}}{\sum_j c_{ut_j}}\log\frac{n}{c_{t_k u}} \tag{3-1}$$

其中，c_{ut_k} 表示用户 u 被标记标签 t_k 的次数，$\sum_j c_{ut_j}$ 表示用户 u 被标记的标签总数，n 表示用户的总数，$c_{t_k u}$ 表示被标记了标签 t_k 的用户总数。通过 TF-IDF 算法对标签处理后，$F_{ut_k}^{t}$ 可以表示标签 t_k 对用户 u 的影响程度。

（2）用户时间特征

推荐算法中存在用户兴趣转移的问题，用户的兴趣会随着时间的改变发生变化，因此本文对用户进行特征提取时要考虑时间因素。用户的兴趣通常分为长期兴趣和近期兴趣，其中近期兴趣对于推荐算法较为重要，它是用户目前较为关注的特征，因此用户最近的操作信息往往更能突出用户目前的兴趣特征。Cheng 等人[42]提出了一种基于自适应指数衰减函数来处理标签的时间信息的用户兴趣模型。本文主要采用自适应的衰减函数定义用户的时间操作权重，近期的特征属性权重较大。用户 u 对标签 t_k 的时间权重如式(3-2)所示。

$$F_{ut_k}^{time}=e^{-\left(t_{now}-t_{ut_k}\right)} \tag{3-2}$$

其中，t_{now} 表示目前时间，而 t_{ut_k} 表示用户最后一次被标注标签 t_k 的时间。当被标注的时间越久，标签的时间权重越小，该标签的特征影响程度就越小。

（3）用户行为特征

用户对不同资源可能会有不同的操作权重，本文构建用户兴趣特征向量时也将考虑操作行为。操作权重越大，用户表现出的兴趣越明显。上一章对用户

-标签的偏好进行加权，接下来将基于此关系进行用户行为特征分析。

$$F_{ut_k}^{o} = e^{\frac{g_{ik}}{c_{ut_k}}} \tag{3-3}$$

其中，g_{ik}是式(2-5)计算的结果，表示用户对标签的加权喜好度。c_{ut_k}是用户u被标记标签t_k的次数，表示用户的平均喜好度。$F_{ut_k}^{o}$表示用户对标签的行为特征，权重越大表明用户对该特征的资源喜爱程度越大。

（4）用户兴趣特征

通过上述的用户标签特征、时间权重、操作特征综合考虑获取用户的兴趣特征，具体如式(3-4)所示。

$$F_{ut_k} = F_{ut_k}^{t} \cdot F_{ut_k}^{time} \cdot F_{ut_k}^{o} \tag{3-4}$$

F_{ut_k}是通过综合用户标签、时间、操作获取的兴趣特征，因此它能更准确的描述用户的兴趣偏好。然后对用户的兴趣特征进行归一化处理，获取单一用户的特征向量。用户u的特征向量可以表示为：$F_u = \left\{F_{ut_1}, F_{ut_2}, \cdots, F_{ut_l}\right\}$。通过获取的用户兴趣特征向量，可以计算用户的相似邻居。

3.1.2 基于标签的资源特征提取

本文推荐算法中的资源具有规范化的标签属性，该标签是由资源提供者和资源的审核者为资源进行标注的，通过该标签可以更加准确的表现出资源的特征。在传统的基于标签的推荐算法中，采用的是社会化标签，社会化标签往往具有数据冗余、稀疏等问题，通过采用规范化的标签能够一定程度解决该问题。上一章对资源-标签的二维矩阵进行了描述，因此对单一资源可以通过标签特征向量表示，因此单一的资源特征向量可以直接表示为$p_i = \left\{p_{it_1}, p_{it_2}, \ldots, p_{it_l}\right\}$。其中，$p_{it_k} = 1$表示资源$i$被标签$t_k$进行了标注，如果没有被标记则$p_{it_k} = 0$。通过资源的特征向量，可以计算出资源的相似邻居。

3.2 概率矩阵分解算法改进

3.2.1 概率矩阵分解算法

由于近年大数据的火爆，信息呈爆发性的增长，利用低维矩阵分解模型进行推荐的概率矩阵分解算法得到了广泛应用。它假设每个用户兴趣特征只受到少数的因数影响，首先将用户和资源映射到低维的特征空间中，通过用户-资源

的评分获取用户和资源的特征向量，最后预测评分信息，给出推荐列表。图 3-1 对算法的预测过程进行说明。

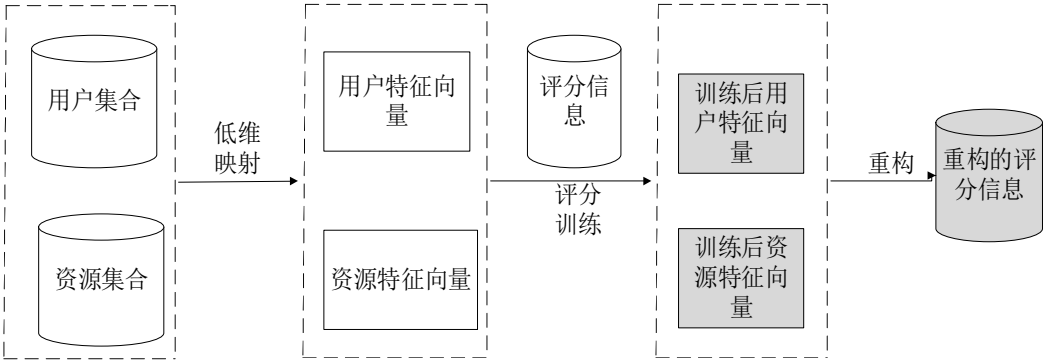


图 3-1 概率矩阵分解算法预测过程

这里假设 $U \in R^{k \times N}$ 和 $V \in R^{K \times M}$ 分别表示用户和资源的特征矩阵，U_u 表示特定用户 u 的 K 维特征向量，V_i 表示特定资源 i 的 K 维特征向量。对已有评分数据的条件概率定义如式(3-5)所示。

$$p(R \mid U, V, \sigma^2) = \prod_{u=1}^{N} \prod_{i=1}^{M} [N(R_{u,i} \mid g(U_u^T V_i), \sigma_R^2)]^{I_{u,i}^R} \tag{3-5}$$

其中，$N(x \mid \mu, \sigma^2)$ 表示平均值为 μ、方差为 σ^2 的正态分布。$I_{u,i}^R$ 是一个 0-1 函数，当用户 u 对资源 i 有评分时值为 1，否则为 0。$g(x)$ 是归一化函数，其中 $g(x)=1/(1+e^x)$，主要是将 $U_u^T V_i$ 的值映射到[0-1]的区间中。

在概率矩阵分解算法中，假设用户和资源的特征向量都服从均值为 0 的高斯先验，以防止过拟合的问题，如式(3-6)所示。

$$p(U \mid \sigma_U^2) = \prod_{u=1}^{N} N(U_u \mid 0, \sigma_U^2 I)$$

$$p(V \mid \sigma_V^2) = \prod_{i=1}^{M} N(V_i \mid 0, \sigma_V^2 I) \tag{3-6}$$

然后经过贝叶斯推断得到用户和资源特征的联合后验概率分布。

$$p\left(U,V\mid R,\sigma_R^2,\sigma_U^2,\sigma_V^2\right)\propto p\left(R\mid U,V,\sigma_R^2\right)p\left(U\mid\sigma_U^2\right)P\left(V\mid\sigma_V^2\right)$$
$$=\prod_{u=1}^{N}\prod_{i=1}^{M}[N(R_{u,i}\mid g(U_u^TV_i),\sigma_R^2)]^{I_{u,i}^R}\times \tag{3-7}$$
$$\prod_{u=1}^{N}N(U_u\mid 0,\sigma_U^2I)\times\prod_{i=1}^{M}N(V_i\mid 0,\sigma_V^2I)$$

接着对式(3-7)进行对数处理，最大化后验概率，相当于最小化对数处理结果，得到特征向量。其中，概率矩阵分解模型如图 3-2 所示。

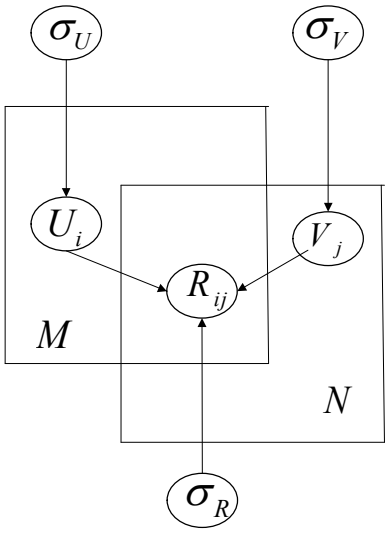


图 3-2 概率矩阵分解模型

算法的结果是通过目标函数使预测的评分值尽量逼近真实的已知评分值，获取用户-资源评分矩阵。通过对概率矩阵分解算法的描述，该算法在矩阵稀疏的条件下具有较好的推荐准确率，同时适用于大数据环境，目前已成为主流的推荐算法。但该算法只简单的对用户-资源评分矩阵进行分解，忽略了用户和资源的关联关系对算法的影响，因此算法的准确率有待提高。本文将通过改进的特征提取算法获取用户和资源的近邻，并将近邻影响关系与概率矩阵算法结合，重构评分信息，提高算法准确率。

3.2.2 相似邻居计算

推荐算法往往需要获取用户或资源的近邻，根据近邻的影响给用户推荐感兴趣的资源列表。为了获取用户和资源的最近邻居，需要进行相似度计算，然

后寻找单一用户和资源的近邻。本章对用户特征提取算法进行了改进，对于用户 u 的兴趣特征模型可以表示为 $F_u=\{F_{ut_1}, F_{ut_2},\cdots, F_{ut_l}\}$，该模型是由用户操作行为权重、用户的标签特征、时间权重的综合考虑的结果。对于单一资源 i 的标签特征模型则表示为：$p_i=\{p_{it_1},p_{it_2},\ldots,p_{it_l}\}$，通过规范化的标签进行了详细的描述。

相似性算法是协同过滤算法中较为关键的技术，若相似度计算不准确则会导致推荐结果出现较大的偏差，目前常用的相似度计算方法如下。

（1）余弦相似度(Cosine Similarity)[43]。通常计算向量间的相似度，将两个向量的余弦值作为度量两个个体间的差异大小，如式(3-8)所示。

$$sim(a,b)=\frac{\overline{a}\cdot\overline{b}}{\|\overline{a}\|\times\|\overline{b}\|} \tag{3-8}$$

（2）Pearson 相关系数(Pearson Correlation Coefficient) [43]是一种线性相关系数。主要是计算两个数值对自身总体进行规范化处理后它们之间余弦夹角的度量，同时是用来反映两个变量线性相关程度的统计量，如式(3-9)所示。

$$sim(a,b)=\frac{\sum_{k=1}^{m}(v_{ak}-\overline{v}_a)(v_{bk}-\overline{v}_b)}{\sqrt{\sum_{k=1}^{m}(v_{ak}-\overline{v}_a)^2}\sqrt{\sum_{k=1}^{m}(v_{bk}-\overline{v}_b)^2}} \tag{3-9}$$

其中，m 表示的资源总数，a 和 b 表示用户，$\overline{v}_a=1/m\sum_{k=1}^{m}v_{ak}$ 表示用户对资源的平均评分。

（3）Jaccard 相似系数(Jaccard Coefficient)[44]。计算符号度量或者布尔值度量的个体间的相似度，如式(3-10)所示。

$$sim(a,b)=\frac{\{I_a\cap I_b\}}{\{I_a\cup I_b\}} \tag{3-10}$$

此外计算相似度的方法还有曼哈顿距离、改进的余弦相似度、在计算相似度的基础上采取相关加权策略等方法。

本文已将用户和资源用一维向量进行表示，因此对于用户和资源的相似邻居将通过余弦算法进行计算。将计算的用户和资源的相似大小进行排序，获取用户和资源的近邻。用户间的相似度计算如式(3-11)所示，资源间的相似度计算如式(3-12)所示。

$$simU=\cos(F_u,F_v)=\frac{\overline{F}_u\cdot\overline{F}_v}{\|F_u\|*\|F_v\|} \tag{3-11}$$

$$simV\left(i,j\right)=\cos\left(P_i,P_j\right)=\frac{\overline{P}_i\cdot\overline{P}_j}{\left\|P_i\right\|*\left\|P_j\right\|} \tag{3-12}$$

其中，$simU\left(u,v\right)$表示用户u和用户v的相似度，$\mathrm{sim}V(i,j)$表示资源i和资源j的相似度。F_u是前面介绍的用户u的兴趣特征向量，P_i表示资源i的标签特征向量。

3.2.3 基于标签的概率矩阵分解算法

随着信息的爆炸式增长，概率矩阵分解算法得到了越来越多的关注。但传统的概率矩阵分解算法只考虑了用户和资源自身的特征而忽略了用户和资源间的影响关系，因此算法准确率有待提升。本课题根据出版资源的特点，通过规范化标签挖掘用户和资源隐藏规律，获取用户和资源近邻关系，并与概率矩阵分解模型进行结合，重构用户-资源评分信息，提高算法的准确率。将改进的算法称为基于标签的概率矩阵分解算法写作 TagMF，算法模型如图 3-3 所示。

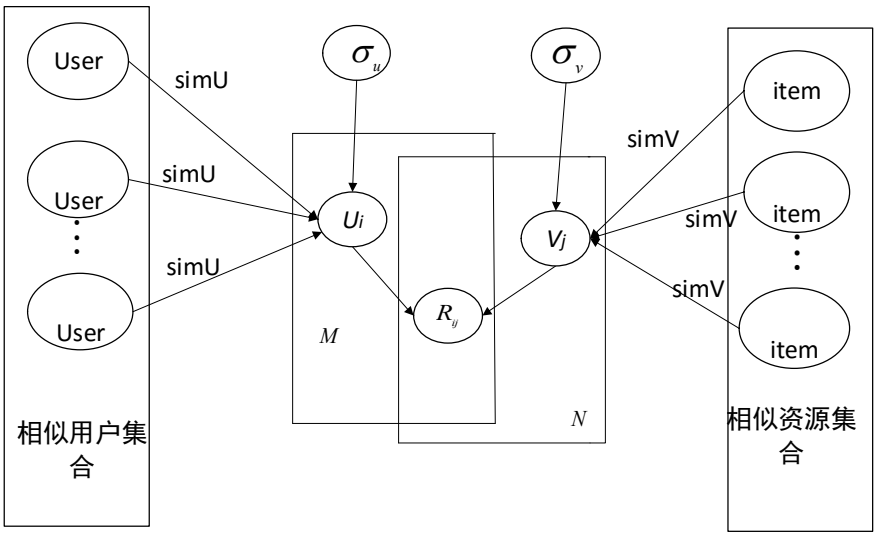


图 3-3 基于标签的概率矩阵分解模型

通过对用户和资源构建特征模型，采用余弦相似度算法寻找其邻居集合，在将用户和资源的邻居其融入到概率矩阵分解模型中，充分考虑了用户和资源邻居的影响关系，因此相似用户和资源具有相似的特征向量。

$$\widetilde{V}_{\mathrm{i}}=\sum_{j\in N_i}\mathrm{sim}V(i,j)*V_j\ ,\ \ \tilde{U}_u=\sum_{v\in N_u}simU\left(u,v\right)*U_v \tag{3-13}$$

其中，$\tilde{V}_j$和$\tilde{U}_u$分别表示相似的特征向量。N_u表示用户u的邻居集合，N_i代

表资源i的邻居集合。本文在考虑了用户和资源自身特性的同时，考虑了相似用户和资源的特征也即近邻的影响。基于 Liu 等人[45]的研究工作，本文将用户和资源的特征向量进行处理，得到公式(3-14)和公式(3-15)。

$$\begin{aligned} p\left(U\left|S,\sigma_U^2,\sigma_S^2\right.\right) &\propto p\left(U\left|\sigma_U^2\right.\right)\times p\left(U\left|S,\sigma_S^2\right.\right) \\ &=\prod_{u=1}^{N}N\left(U_u\left|0,\sigma_U^2I\right.\right)\times\prod_{u=1}^{N}N\left(U_u\left|\sum_{v\in N_u}simU\left(u,v\right)*U_v,\sigma_S^2I\right.\right) \end{aligned} \tag{3-14}$$

$$\begin{aligned} p\left(V\left|T,\sigma_V^2,\sigma_T^2\right.\right) &\propto p\left(V\left|\sigma_V^2\right.\right)\times p\left(V\left|T,\sigma_T^2\right.\right) \\ &=\prod_{i=1}^{M}N\left(V_i\left|0,\sigma_V^2I\right.\right)\times\prod_{i=1}^{M}N\left(V_i\left|\sum_{j\in N_i}simV\left(i,j\right)*V_j,\sigma_V^2I\right.\right) \end{aligned} \tag{3-15}$$

特征向量U和V服从平均值为 0 的高斯先验，与公式(3-7)相似，通过贝叶斯推断得到后验概率，如公式(3-16)所示。

$$\begin{aligned} p\left(U,V\left|R,S,T,\sigma_R^2,\sigma_u^2,\sigma_v^2\right.\right) &\propto p\left(R\left|U,V,\sigma_R^2\right.\right)\times p\left(V\left|T,\sigma_V^2,\sigma_T^2\right.\right)\times p\left(U\left|S,\sigma_U^2,\sigma_S^2\right.\right) \\ &=\prod_{u=1}^{N}\prod_{i=1}^{M}\left[N\left(R_{u,v}\left|g\left(U_u^TV_i\right),\sigma_R^2\right.\right)\right]^{I_{u,j}^R}\times \\ &\prod_{i=1}^{M}N\left(V_i\left|\sum_{j\in N_i}simV\left(i,j\right)*V_j,\sigma_V^2I\right.\right)\times \\ &\prod_{u=1}^{N}N\left(U_u\left|\sum_{v\in N_u}simU\left(u,v\right)*U_v,\sigma_S^2I\right.\right)\times \\ &\prod_{u=1}^{N}N\left(U_u\left|0,\sigma_U^2I\right.\right)\times\prod_{i=1}^{M}N\left(V_i\left|0,\sigma_V^2I\right.\right) \end{aligned} \tag{3-16}$$

根据对数函数的性质，为简化计算，将上式进行对数化。

$$\begin{aligned} &\ln p\left(U,V\left|R,T,S,\sigma_R^2,\sigma_U^2,\sigma_V^2,\sigma_T^2,\sigma_S^2\right.\right)=-\frac{1}{2}\left(\sum_{u=1}^{N}\sum_{i=1}^{M}I_{u,i}^R\right)\ln\sigma_R^2- \\ &\frac{1}{2\sigma_U^2}\sum_{u=1}^{N}U_u^TU_u-\frac{1}{2\sigma_V^2}\sum_{i=1}^{M}V_i^TV_i-\frac{1}{2\sigma_R^2}\sum_{u=1}^{N}\sum_{i=1}^{M}I_{u,i}^R\left(R_{u,i}-g\left(U_u^TV_i\right)\right)^2- \\ &\frac{1}{2\sigma_{\mathrm{S}}^2}\sum_{u=1}^{N}\left\|U_u-\sum_{v\in N_u}simU\left(u,v\right)*U_v\right\|_F^2-\frac{1}{2\sigma_T^2}\sum_{i=1}^{M}\left\|V_i-\sum_{j\in N_i}simV\left(i,j\right)*V_j\right\|_F^2- \\ &\frac{1}{2}\left(\left(N\times K\right)\ln\sigma_U^2+\left(M\times K\right)\ln\sigma_V^2+\left(N\times K\right)\ln\sigma_S^2+\left(M\times K\right)\ln\sigma_T^2\right)+C \end{aligned} \tag{3-17}$$

概率矩阵分解算法最终需要得到用户和资源的潜在特征向量U和V，以实

现最终的评分预测。机器学习是从一类数据中挖掘潜在规律，然后通过规律预测未知数据。因此，本文将概率矩阵分解的求解问题转化为数学中求目标函数的最小值，即相当于最小化公式(3-18)。

$$\begin{aligned} E\left(R,T,S,U,V\right)=&\frac{1}{2}\sum_{u=1}^{N}\sum_{i=1}^{M}I_{u,i}^{R}\left(R_{u,i}-g\left(U_u^{T}V_i\right)\right)^2+\\ &\frac{\lambda_S}{2}\sum_{u=1}^{N}\left\|U_u-\sum_{v\in N_u}simU\left(u,v\right)*U_v\right\|_F^2+\frac{\lambda}{2}\sum_{u=1}^{N}U_u^{T}U_u+\\ &\frac{\lambda_T}{2}\sum_{i=1}^{M}\left\|V_i-\sum_{j\in N_i}simV\left(i,j\right)*V_j\right\|_F^2+\frac{\lambda}{2}\sum_{i=1}^{M}V_i^{T}V_i \end{aligned} \tag{3-18}$$

其中，$\lambda_U=\sigma_R^2/\sigma_U^2$，$\lambda_S=\sigma_R^2/\sigma_S^2$，$\lambda_v=\sigma_R^2/\sigma_V^2$，$\lambda_T=\sigma_R^2/\sigma_T^2$。利用梯度下降算法，可以得到用户和资源的特征向量，计算方式如下。

$$\begin{aligned} \frac{\partial E}{\partial U_u}=&\sum_{i=1}^{M}I_{u,i}^{R}\left(-V_i\right)g'\left(U_u^{T}V_i\right)\left(R_{u,i}-g\left(U_u^{T}V_i\right)\right)+\lambda_u U_u+\\ &\lambda_S\left(U_u-\sum_{v\in N_u}simU\left(u,v\right)*U_v\right)-\\ &\lambda_S\sum_{u\in N_v}simU\left(v,u\right)\left(U_v-\sum_{w\in N_v}simU\left(w,v\right)*U_w\right) \end{aligned} \tag{3-19}$$

$$\begin{aligned} \frac{\partial E}{\partial V_i}=&\sum_{u=1}^{N}I_{u,i}^{R}\left(-U_u\right)g'\left(U_u^{T}V_i\right)\left(R_{u,i}-g\left(U_u^{T}V_i\right)\right)+\lambda_V V_i+\\ &\lambda_V\left(V_i-\sum_{j\in N_i}simV\left(i,j\right)*V_j\right)-\\ &\lambda_V\sum_{i\in N_j}simV\left(j,i\right)\left(V_j-\sum_{l\in N_j}simV\left(l,j\right)*V_l\right) \end{aligned} \tag{3-20}$$

其中，$g'\left(x\right)$是$g\left(x\right)$的导数，$g'\left(x\right)=e^{-x}/\left(1+e^{-x}\right)$。

3.3 算法流程描述

通过对基于标签概率矩阵算法的描述和公式的推导，本节将给出算法的详细流程和算法伪码步骤。根据出版行业资源规范化标签特点，首先根据用户对资源的操作获取用户的行为操作权重，将用户的标签属性进行加权，并综合考虑时间权重，完成对传统的特征提取算法的改进，通过改进的特征提取算法寻

找用户和资源的影响关系，将用户和资源的近邻关系融入到概率矩阵分解算法中，对概率矩阵分解算法进行改进，提高了推荐算法的准确率。如图 3-4 给出了基于标签的概率矩阵分解算法的流程图。

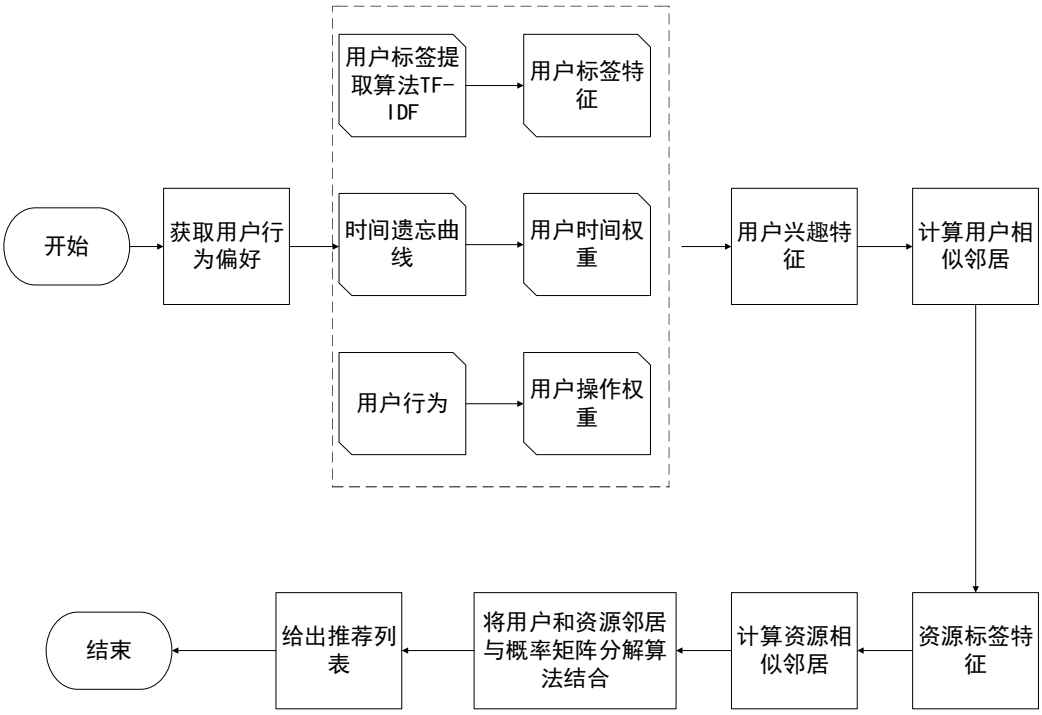


图 3-4 基于标签的概率矩阵分解算法流程图

通过对算法流程分析，整个算法主要包含四个步骤：第一步是处理数据，获取用户和资源的数据表示，已在上一章进行了处理；第二步是采用改进的特征提取算法挖掘出用户与用户，资源与资源的关系，找出邻居集合；第三步将用户和资源的邻居影响融入概率矩阵分解模型中，考虑自身特征的同时考虑邻居的影响，重构用户-资源评分矩阵；第四步根据评分矩阵对用户进行资源列表推荐。算法的时间复杂度主要集中在第二步和第三步，用户特征提取的时间复杂度为$O(N\times L)$，通过采用离线方式进行计算，提高推荐效率。对于算法第三步，其时间复杂度为$O\left(\left(r+N_u N+N_v M\right)K\right)$，其中 r 为用户所交互的资源总数，K 为特征向量维度，由于 K 的维度较少，所示该算法的总体时间复杂度并不高，可以有效的对大数据进行处理。通过流程图详细的描述了基于标签的概率矩阵分解算法，下面将对算法的详细步骤给出伪码表示，如表 3-1 所示。通过输入处理后的数据，算法最终将给用户推荐其感兴趣的资源列表。

表 3-1 算法伪码步骤

输入：用户-资源评分矩阵 $R_{N\times M}$、用户-标签交互矩阵 $S_{N\times L}$、资源-标签矩阵 $p_{M\times L}$、用户标签偏好 $F_{N\times L}$ 二维矩阵，用户集合 $U=\{u_1,u_2,\ldots,u_N\}$，资源集合 $I=\{i_1,i_2,\ldots,i_M\}$

输出：Top-N List(推荐列表)

For $u=1$ to N

利用式(3-1) 计算用户标签特征 $F_{ut_k}^{t}$

利用式(3-2) 计算用户时间权重特征 $F_{ut_k}^{time}$

利用式(3-3) 计算用户行为操作特征 $F_{ut_k}^{o}$

利用式(3-4) 计算用户兴趣特征向量 F_u

End for

//同理计算资源特征向量 P_i

For $u=1$ to N

For $u=1$ to N

$simU(u,v)$ //利用式(3-11)计算相似用户

End for

End for

//同理计算相似资源 $simV(i,j)$

While 目标函数 $E:E\left(R,T,S,U,V\right)$ 最小

For $u=1$ to N

计算 $\frac{\partial E}{\partial U_u}$ 利用式(3-19) 更新 U_u

End for

For $i=1$ to M

计算 $\frac{\partial E}{\partial V_i}$ 利用式(3-20) 更新 V_i

End for

Until E 收敛

End

计算新的评分矩阵 $\hat{R}_{N\times M}$

根据评分高低组成推荐集 Top-N List。

返回 Top-N List

3.4 实验设计

3.4.1 实验数据与环境

本文实验环境分为硬件和软件两部分，其中硬件环境：CPU 为 Intel(R) Xeon(R) E3-1226 v3 @3.30GHz 3.30GHz ，内存为 16G。软件环境：操作系统 centos 6.5、java、MySQL、redis 等。

本实验使用的数据集是出版集团真实数据集，其中包含用户数量为 6323954，出版资源(包含文章、书籍、视频、音频、PDF、图片等)数量为 64542，标签数量为 9859，其中出版资源都被标注了标签。本实验选取时间在 2017 年 2 月到 2017 年 8 月间的所有数据进行试验。本实验选取数据集中较为活跃的 2000 名用户，并获取其用户的所有相关数据，将数据集按 4:1 的比例分为训练集和测试集，其中测试集是随机选取连续时间段的历史数据，其余数据作为训练集进行试验。

上文介绍了用户的操作行为和操作权重，操作权重越大表示用户对该资源或者标签的喜爱程度越大。用户的行为权重如表 3-2 所示。

表 3-2 用户的行为权重

用户行为	权重值	具体评分
浏览超过 1 分钟	w_1	1
评论	w_2	2
交互操作	w_3	2
收藏	w_4	3
分享	w_5	4
购买	w_6	5

3.4.2 实验评价指标

本文实验采用了较为常用的准确率（precision）、召回率（recall）和均方根误差（RMSE）来评价推荐结果的好坏。准确率和召回率自从 1998 年被首次使用后，目前已经得到了广泛的应用。RMSE 是评价预测评分准确性的标准，反映了算法的预测评分和用户实际操作评分的贴近程度。其中 RMSE 值越小，意味着推荐质量越高。准确率是评估推荐项目列表中命中测试数据集中资源的比例，召回率是测试集中推荐的项目个数在用户时间感兴趣项目的比例。同时本文将

采用 F 指标来衡量算法，F 值综合了准确率和召回率。

假设 C 是测试样例的数量，算法对资源 C 的实际评分为 $\{p_1, p_2, p_3, \ldots, p_c\}$，对应的预测评分为 $\{r_1, r_2, r_3, \ldots, r_c\}$，则对应的均方根误差如公式(3-17)所示。准确率、召回率和 F 值的定义如公式(3-18)、(3-19)和(3-20)所示。

$$RMSE=\sqrt{\frac{\sum_{i=1}^{c}\left(p_i-r_i\right)^2}{C}} \tag{3-21}$$

$$precision=\frac{\sum_{u\in U}\left|R(u)\cap T(u)\right|}{\sum_{u\in U}\left|R(u)\right|} \tag{3-22}$$

$$recall=\frac{\sum_{u\in U}\left|R(u)\cap T(u)\right|}{\sum_{u\in U}\left|T(u)\right|} \tag{3-23}$$

$$F=\frac{2\times precision\times recall}{precision+recall} \tag{3-24}$$

其中，$R(u)$ 是通过训练集训练后给出的推荐资源列表，$T(u)$ 是用户在测试集上具体的操作资源列表。根据上面的描述，在实验中 RMSE 的值越小，同时准确率、召回率和 F 越高这样说明推荐结果越优。准确率和召回率在某些情况下会出现矛盾，因此本文也将为综合考虑 F 值，F 值越高推荐算法越好。本文将通过多组对比实验和参数调节实验进行实验验证。

3.4.3 对比算法与参数设定

本次实验将选取 3 种方法作为对比实验，并将基于标签的概率矩阵分解算法写作 TagMF。

（1）概率矩阵分解算法（PMF）：对用户-资源评分矩阵进行概率矩阵分解，只考虑了用户和资源自身的特征。

（2）基于社交关系的推荐算法（SocailMF）[46]：将用户社交网络关系考虑到概率矩阵分解模型中，忽略了资源间的影响关系。

（3）基于对象间关联关系算法（Probabilistic Matrix Factorization with User and Item relation，PMFUI）[47]：结合推荐对象间关联关系进行推荐，将用户和资源的关系都进行考虑。

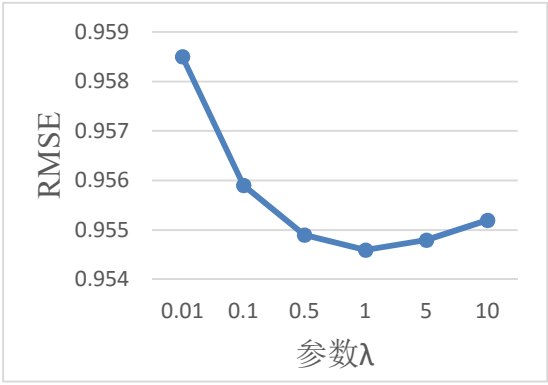
在实验中为了降低模型的复杂度，本文假设 $\lambda_U=\lambda_V=0.001$，并且假设

$\lambda_S = \lambda_T = \lambda$。其中用户和资源的初始特征向量$U$和$V$服从均值为0的正态分布。此后每一次迭代运算中，特征向量U和V是根据前一代的值进行迭代更新，直至收敛。相似邻居均选取Top-20最相似的用户和资源。

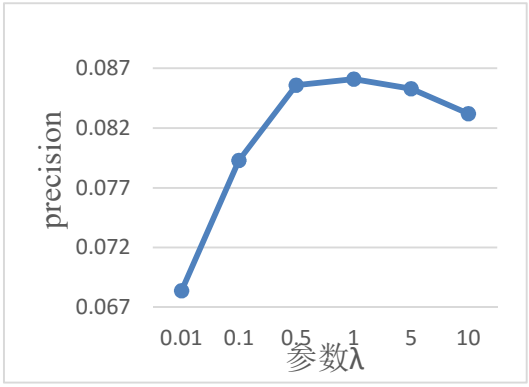
3.5 实验结果分析

3.5.1 参数λ的影响实验

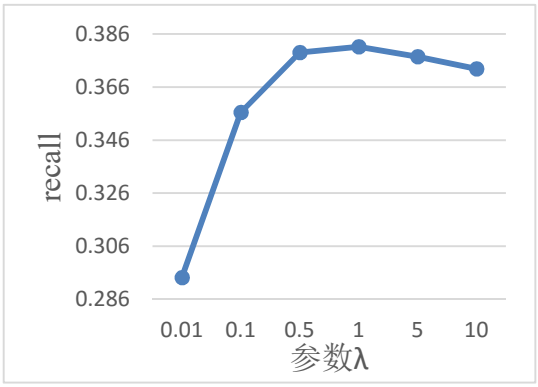
参数λ控制着用户和资源关系对算法的影响程度。本实验中为了降低算法复杂度假设$\lambda_S = \lambda_T = \lambda$，特征向量维度$K=10$，推荐列表长度L=10。通过改变$\lambda$的数值研究对基于标签概率矩阵分解算法的影响程度。参数λ的影响实验如图3-5所示。



（a）参数λ对RMSE的影响



（b）参数λ对precision的影响



（c）参数λ对recall的影响

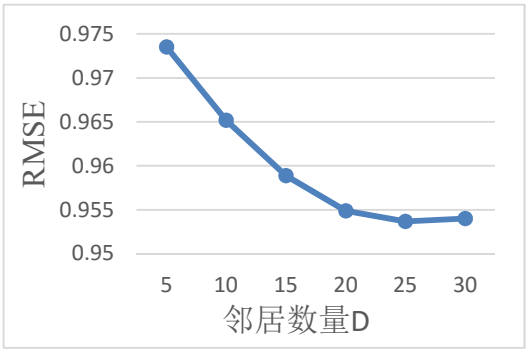
图3-5 参数λ的影响实验

如图3-5所示，参数λ改变后对评分矩阵的预测精度RMSE参数影响，同时推荐算法中的准确率和召回率也都发生了变化，说明了λ所控制的用户和资源

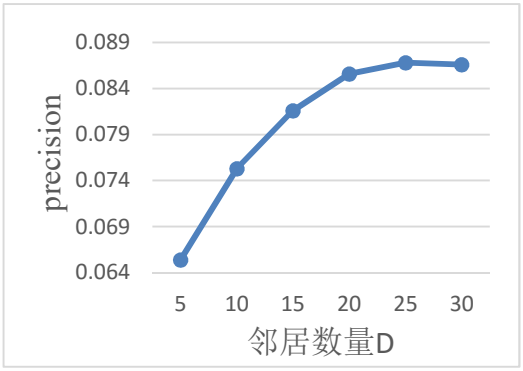
关系在整个算法中所占的比重会对算法产生影响，也说明本文基于标签的概率矩阵算法的合理性。当λ从 0 增加时，RMSE 将会降低，准确率和召回率提升，说明算法的精度有所提升，当其超过某个阀值时，RMSE 出现增大精度降低的现象，说明了用户的资源的影响关系将对推荐产生影响，当继续增大会出现过拟合问题，推荐过程中控制λ将有效提高推荐的质量。对于算法的准确率和召回率，随着λ的增大而增加，当超过阀值后，开始逐渐降低。通过实验发现当λ由 0-0.5 区间进行变化时，算法的变化较大，当继续增大时算法的变化率变少，因此在后面的实验中将设定λ=0.5，使其具有较好的推荐质量。通过本次实验证明了基于标签的概率矩阵分解算法的有效性。

3.5.2 邻居数量 D 的影响实验

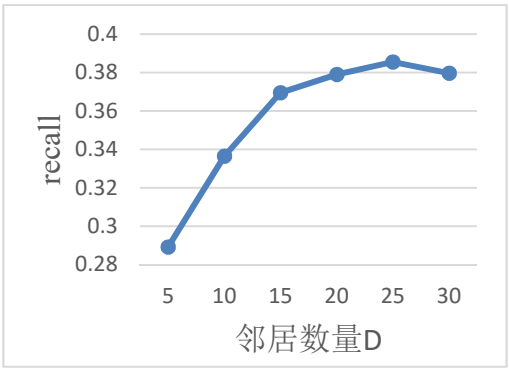
在基于标签的概率矩阵分解算法中，将用户和资源的邻居集合的影响融入到概率矩阵分解算法中。本实验为研究邻居数量对算法的影响关系，其它参数保持最优。邻居数量 D 的影响实验如图 3-6 所示。



（a）邻居数量 D 对 RMSE 的影响



（b）邻居数量 D 对 precision 的影响



（c）邻居数量 D 对 recall 的影响

图 3-6 邻居数量 D 的影响实验

对图 3-6 展示的实验结果进行分析，当邻居数量增加时，算法的 RMSE 指标降低，准确率和召回率都有所提升，说明相似邻居能影响最终的推荐结果，也证明了本文基于标签概率矩阵算法的有效性。当邻居数量在 5-20 进行增加时，算法的准确率等指标增长较快，说明邻居对算法的影响程度较大。当邻居数量超过 20 后，邻居的影响关系有所下降，推荐算法的准确率和召回率虽然有所增加但基本保持不变，同时算法的效率将随着邻居数量的增加有所降低，因此在对比实验中将选取相似邻居为 Top-20 的用户和资源，在保证推荐算法质量的同时保证整个算法的效率。

3.5.3 特征向量维度 K 的对比实验

不同的特征向量维度会使构建的用户和资源的特征向量发生变化，导致重构的用户-资源评分矩阵的不同。本实验将比较了各个算法在不同特征向量维度下的结果，研究维度对算法的影响和算法的对比效果。在实验中，我们设定特征向量维度分表为 $K=5$，$K=10$。其中 λ=0.5，推荐列表长度 L=10。表 3-3 给出了不同算法下的实验结果。

表 3-3 不同特征向量维度 K 的实验比较

方法	$K=5$			$K=10$		
	RMSE	*precision*	*recall*	*RMSE*	*precision*	*recall*
PMF	1.0137	0.0512	0.2103	1.0121	0.0524	0.2121
SocialMF	0.9743	0.0622	0.2405	0.9728	0.0638	0.2436
PMFUI	0.9687	0.0726	0.3105	0.9652	0.0733	0.3123
TagMF	0.9552	0.0835	0.3782	0.9549	0.0856	0.3791

通过以上对比可以看出：（1）随着维度 K 的增大算法的精度有一定的提升，所有算法的准确率和召回率也都有一定的提升，但是随着 K 的增大会加大算法模型的时间复杂度。（2）SocialMF，PMFUI，TagMF 与 PMF 相比算法的精度、准确率和召回率都有较大的提升，这说明用户间的关系会对推荐结果产生影响，充分考虑其关系可以提高算法精度、准确率和召回率。（3）PMFUI 算法较 SocialMF 算法有所提升，是因为其考虑了资源间的影响关系，说明推荐过程中在考虑用户关系的同时考虑推荐资源间的影响关系能提高推荐的准确率。（4）TagMF 算法较 PMFUI 有所提升，是因为本文通过用户的操作权重，用户标签，

时间权重等多方面综合考虑用户兴趣特征，说明本文基于标签综合因素考虑的用户特征更加准确，证明了 TagMF 算法的有效性。（5）TagMF 算法与概率矩阵算法 PMF 相比，评分的精度、准确率和召回率都有所提升，证明本文提出的基于标签的概率矩阵算法的有效性与合理性。

3.5.4 推荐长度 L 的对比实验

在推荐过程中，推荐列表的长度往往也能影响推荐的质量。实验过程中保证其它参数一致，通过改变推荐资源列表的长度对比本文 TagMF 算法与其它三种算法的推荐质量。

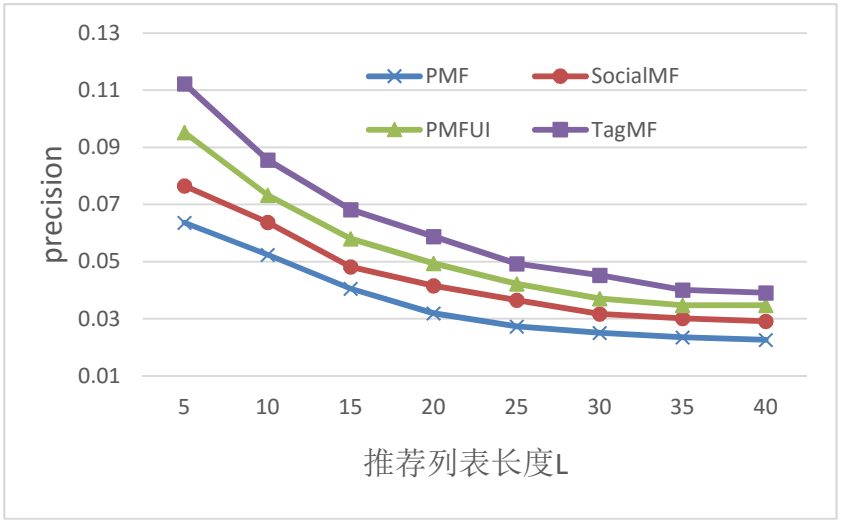


图 3-7 precision 对比实验

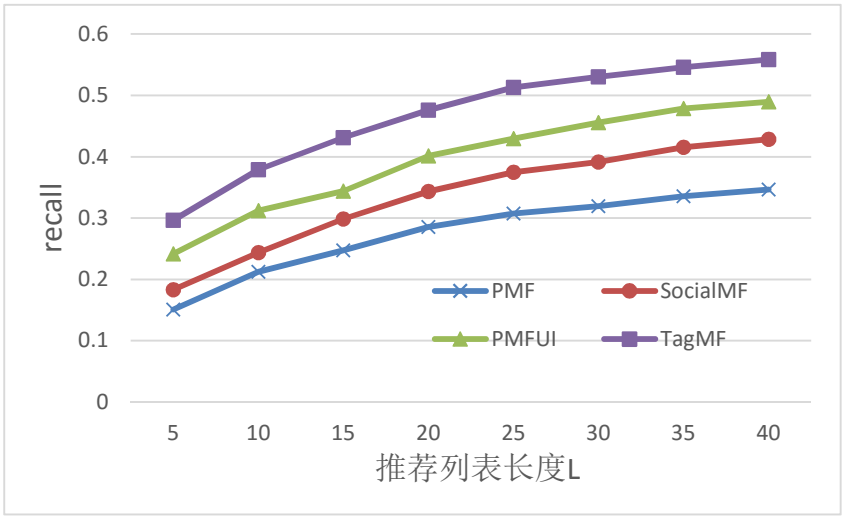


图 3-8 recall 对比实验

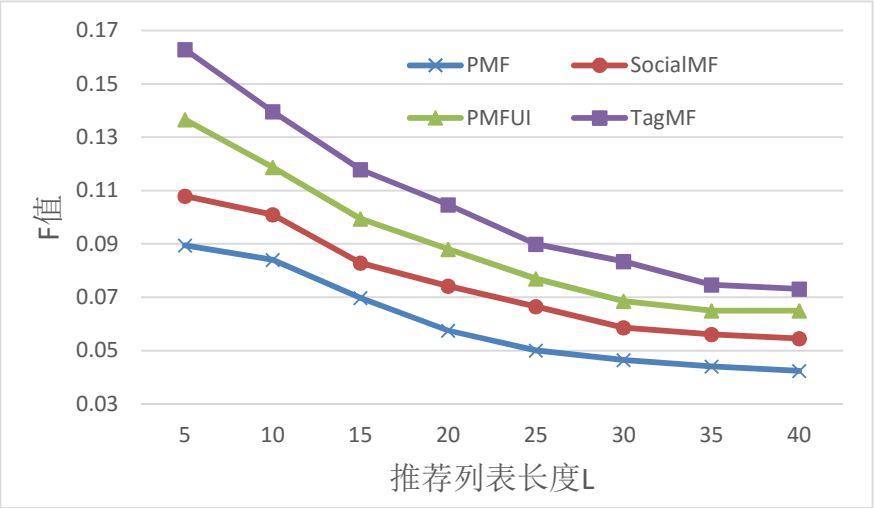


图 3-9 *F* 值对比实验

通过图 3-7、图 3-8 和图 3-9 对比实验发现，在给用户推荐相同数量的资源情况下，TagMF 算法较其它算法准确率、召回率以及 *F* 值均是最高，也说明了基于标签的概率矩阵分解算法较其它算法有所提升。由实验结果发现推荐资源数量在区间 5-15 时，TagMF 相比于其它的算法有提升幅度较大，说明推荐较少资源本文算法效果较优。当推荐数量继续增加时，各推荐算法的准确率都有所降低，但召回率有所提升，当推荐资源数量超过 30 后准确率、召回率和 *F* 值都趋于平稳，但 TagMF 算法的各项参数均高于其它算法。实验数据证明了在考虑用户-资源评分矩阵自身特性的同时，考虑用户和资源的影响关系有利于算法质量的提升，同时本文 TagMF 算法较 PMFUI 算法有所提升，证明了通过规范化标签寻找用户和资源关系更为准确。实验结果最终证明，在推荐资源列表长度相同的情况下，TagMF 算法的推荐质量高于其它算法。

3.5.5 算法耗时对比实验

对比 4 种推荐算法的运行时间，其中 PMFUI 耗时最久，根据表 3-4 可以看出算法的运行时间满足：$PMFUI > TagMF > SocialMF > PMF$ 。PMFUI 算法耗时较久，考虑了用户和资源间的关联关系，将关联关系应用到推荐算法中，其中关联分析耗时较长。本文的 TagMF 算法则可以通过用户和资源的规范化标签寻找用户和资源的影响关系，保证了推荐算法的质量的同时具有较好的效率。SocialMF 较 TagMF 算法的推荐时间有所降低，只考虑了用户间的影响关系而没有考虑资源间的影响关系，而 PMF 算法只是对用户-资源评分矩阵进行分析，因

此这两种算法的耗时较短，但是这两种算法的推荐质量则较低。

表 3-4 算法运行时间

算法	耗时(s)
PMF	9.5
SocialMF	15.2
PMFUI	22.6
TagMF	19.3

通过上诉实验得出的结论：基于标签概率矩阵分解算法通过用户的操作行为，用户标签特征和操作时间权重综合获取的用户特征向量较为准确，并将挖掘的用户和资源间的关联关系运用于概率矩阵分解中，通过调节参数可以达到较好的准确率和召回率。但因算法将用户和资源的影响关系融入概率矩阵分解算法中，使得算法的复杂度有所提升，本文通过基于标签的关联关系的交互式操作缩小资源备选集，提高推荐的运行效率，下文将详细介绍交互式方法。

3.6 本章小结

本章首先对传统特征提取方法进行改进，通过综合用户的标签特征、行为特征和时间权重考虑用户的兴趣特征。然后介绍了概率矩阵分解算法和其优缺点，针对传统概率矩阵分解算法仅考虑用户-资源评分矩阵，忽略了用户和资源的关系，进行了改进。首先根据改进的特征提取算法获取用户和资源的近邻关系，将近邻关系与概率矩阵分解模型融合，对改进的算法进行了公式推导和详细推荐流程，分析了算法的相关时间复杂度，并给出算法伪码步骤。

最后对改进的推荐算法进行实验验证。首先验证参数 λ 对算法的影响关系，通过调节参数 λ，算法的精度发生了改变，说明了融入用户和资源关系能改变推荐质量。接着验证相似邻居数量 D 对算法的影响关系，通过增加邻居数量算法的精度有所提升，证明了本文基于标签的概率矩阵分解算法的有效性。接着设定不同特征向量维度 K 进行对比实验，证明了本文算法在评分精度、准确率和召回率上较其它三种对比算法均有较好的效果。最后通过改变推荐列表长度，证明了本文改进的推荐算法在推荐准确度和召回率上优于对比的三种算法。

第 4 章 基于标签关联分析的交互式方法研究

在传统推荐算法中，随着系统资源的增多，推荐算法的性能和准确率往往会降低，同时推荐过程中存在新用户冷启动等问题。上一章对交互式推荐方法中的一个关键步骤推荐算法进行了研究，本章将对交互式的推荐机制进行研究。首先通过资源-标签矩阵进行分析获取标签的关联关系，然后提供标签供用户交互，交互完成后由标签的关联分析对资源备选集进行划分，通过不断缩减资源备选集来满足用户需求，提高推荐效率。接着针对推荐算法的冷启动问题，给出了解决方案，分析用户的特征构建决策树，新用户进入系统后进行匹配，给新用户推荐兴趣的主题资源，若用户不满意则在本文的交互式框架下获取资源，完善用户模型，解决了冷启动问题。最后通过在真实数据集上的实验验证了本文冷启动算法和交互式方法的有效性。

4.1 交互式方法概述

交互式推荐方法主要分为交互方法和推荐算法两部分，上一章对推荐算法进行了详细的介绍。本章将对交互推荐流程，交互式方法的选择过程、资源划分进行研究。当用户进入系统后，通过推荐算法会为用户推荐资源。但当用户进入系统后可能由于其信息不完整、较长时间未使用系统、用户兴趣发生较大改变都将导致推荐准确率不高，以至于用户对推荐的资源不满意。为满足用户的需求，系统将给出标签进行交互，通过用户的反馈定位其感兴趣的资源，交互推荐流程如图 4-1 所示。

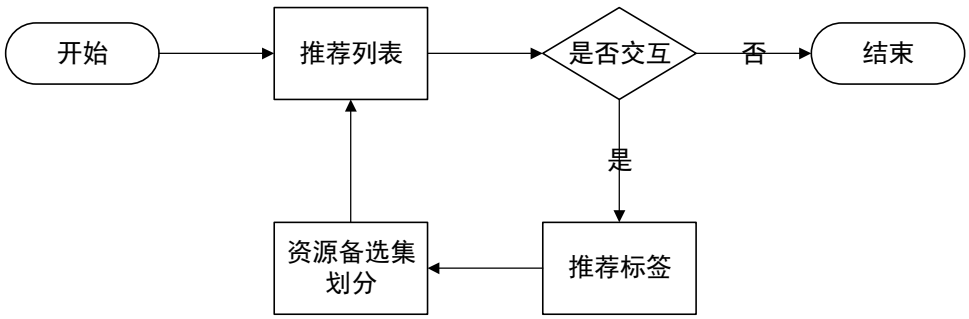


图 4-1 交互推荐流程

交互推荐流程的关键是判断用户是否需要进行交互操作和交互操作后资源备选集的划分。用户进入系统后，推荐系统分两种情况进行推荐，如果不是新用户，系统利用已有的用户模型，然后采用基于标签的概率矩阵分解算法给用户推荐感兴趣的资源。如果是新用户，会产生用户冷启动问题，本文将会采用决策树分类算法和交互方法相结合推荐用户感兴趣的资源，冷启动将在后文进行详细说明，用户推荐流程如图 4-2 所示。

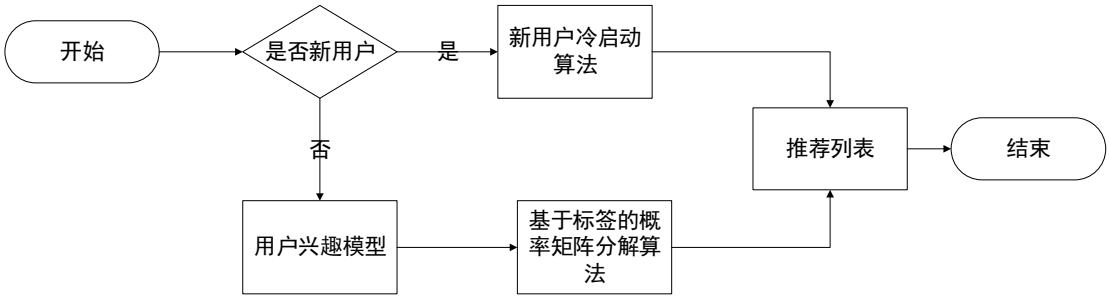


图 4-2 用户推荐流程

本文基于标签关联分析的交互式推荐方法，可以有效的解决资源急速增长的推荐效率问题，通过交互过程对资源备选集进行筛选，在保证推荐质量的同时提高算法的效率。如图 4-3 所示用户标签交互过程。

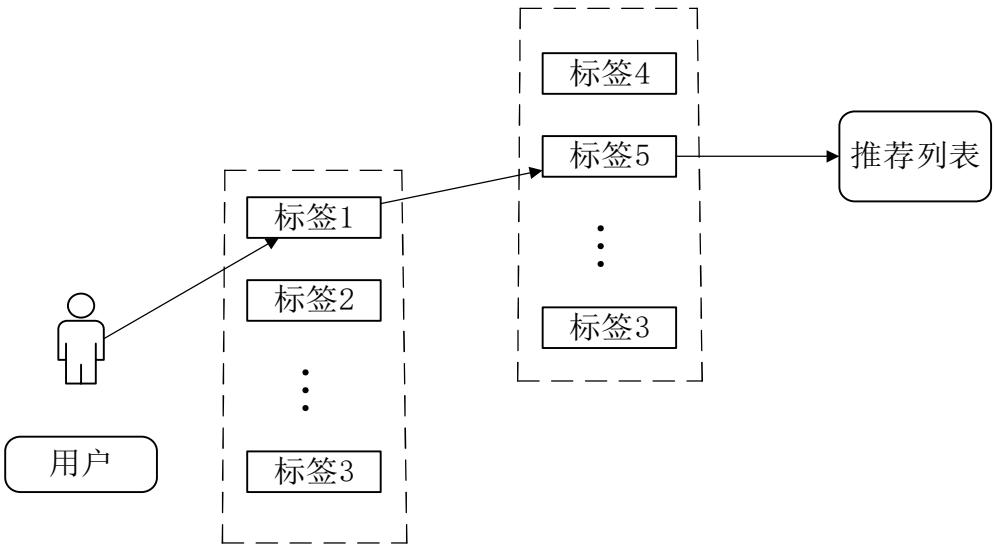


图 4-3 用户标签交互过程图

交互过程是对标签选择的过程，交互过程则不区分新用户和非新用户。在

交互过程中，用户对推荐资源列表不满意，将再次提供标签进行交互，用户的每一次交互都是对资源备选集的进一步缩减，若用户对推荐的资源较满意此时交互过程将停止。如图 4-3 所示，当用户通过交互选择标签 1 后，若用户对推荐的资源列表不满意，将再次提供标签供用户交互，同时根据分析后的标签划分资源备选集，重复此操作直至用户对交互后的资源列表满意。

交互式算法中用户是否对推荐资源满意，通过用户有效的操作进行证明。用户的有效操作通过用户浏览资源的时长，是否收藏，是否购买等有效操作进行说明。只有用户对资源进行了有效操作，说明用户对推荐资源比较满意，否则将再次提供标签供用户交互。用户有效的操作行为，也能进一步说明用户的兴趣爱好，系统也会将操作行为收集，构建完善的用户兴趣模型。

4.2 基于标签关联分析的资源划分

4.2.1 标签关联分析

本文研究目的是为了出版行业设计一个个性化推荐系统，提升用户粘性。根据出版行业的资源具有规范化标签这一特点采用交互式推荐算法，规范化的标签可以准确的描述资源的特征，不同的资源可能具有相同的标签，同时关联性强的标签可能经常出现在同一资源中，通过挖掘标签的关联关系可以在用户交互过程中对资源备选集的划分进行优化，满足用户的需求。

关联规则反映的是一个事物与其他事物之间的相互依赖性和关联性，常用于在线电商和大型商场的推荐中，通过对客户的购买记录进行分析，最终发现客户群体的购买习惯的内在共性。在传统的个性化推荐过程中，通过分析用户购买的关联关系进行推荐，本文则通过分析资源-标签中标签的关联关系，对用户交互过程中的资源备选集进行划分。关联规则分析中通常包含：支持度、置信度与提升度等指标，本文主要考虑支持度和置信度。

本文标签关联关系将通过如下关系进行表示：规范化的标签库为 $T=\{t_1,t_2,\ldots,t_l\}$，资源集合表示为 $I=\{i_1,i_2,\ldots,i_M\}$，M 表示资源总数，通过对资源-标签矩阵 $p_{M\times L}$ 进行分析。对于标签 $A\Rightarrow B$ 的支持度表示为：

$$support(A\Rightarrow B)=\frac{count\left(A\cap B\right)}{M} \tag{4-1}$$

其中，$count\left(A\cap B\right)$表示标签 A 和 B 共同出现在相同资源的次数。

$support(A \Rightarrow B)$ 支持度表示 A 与 B 同时出现的概率。对于标签 $A \Rightarrow B$ 的置信度表示为：

$$confidence(A \Rightarrow B) = \frac{count(A \cap B)}{count(A)} \tag{4-2}$$

其中，$count(A)$ 表示标签 A 在资源中出现的次数。$confidence(A \Rightarrow B)$ 表示交集部分在 A 中所占的比例，如果比例大说明具有标签 A 属性的资源很大程度会具有标签 B 的属性。

通过对整个资源-标签矩阵进行标签分析，指定最小支持度阀值 *minsup* 和最小置信度阀值 *minconf* ，挖掘满足需求的标签关联规则。通过标签的关联分析对资源备选集进行划分，每次资源划分时的标签关系满足最少支持阀值和最小置信度阀值。对于标签的关联分析，会随着资源-标签矩阵的改变而发生变化，但是对用户资源推荐的影响有限，为了提高推荐算法的总体效率，将标签的关联分析放于线下进行。线下分析后，将标签的关联规则进行存储，同时保留相关标签的支持度和置信度。当资源备选集进行划分时通过标签的关联分析，划分资源备选集，同时对划分的资源备选集采用标签关联权重对资源进行重排序。

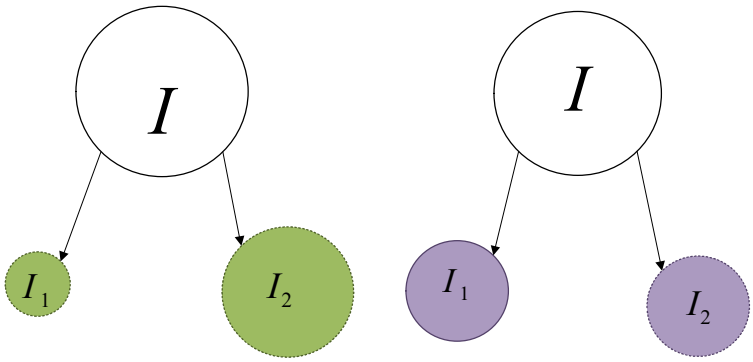
4.2.2 资源备选集划分

随着出版资源的增加，推荐算法的效率将有所降低，同时推荐的资源列表可能不被用户所接受，因此提供标签供用户进行交互，可以快速获取感兴趣的资源。对于集合资源 I ，集合中包含的信息量是固定的，如何快速的获取用户所需要的信息，是我们需要解决的问题。当我们猜测一个角色时，最有效的问题是：“这个角色是男的还是女的？”。从信息论的角度考虑，回答该问题可以提供最多的信息，并且可以有效的缩减搜索空间。本文将通过信息熵和标签的关联分析对资源备选集进行划分。本文资源通过规范化的标签特征向量进行描述，系统通过标签的交互了解用户的喜好，通过标签属性的划分使资源的区分信息尽可能大，这样系统才能对用户反馈的资源进行高效的筛选。

在用户标签交互过程中，发现不同的标签可能经常出现在同一资源上，出现的概率越高，该标签间的关系也越密切，用户选择其中一个标签时，可能也隐含的选择了另一个标签，通过标签间的联系对资源进行划分可以对资源备选集划分的进行优化。比如“机器人”，“科学小说”这两标签经常出现在同一资源上，当用户选择标签“机器人”时，用户可能也对“科学小说”的资源比较

感兴趣，因此划分资源备选集时不能仅通过“机器人”标签对资源进行大幅度的缩减。

如图 4-4 所示，标签属性 A 代表“机器人”属性，标签 A 关联属性包含标签“机器人”、“科技小说”、“未来”。仅通过单一属性 A 进行资源划分的后的两个后的两个资源备选集I_1和I_2所包含的资源数量区别很大，系统会将用户的资源备选集划分为I_1，这种结果在划分准确的情况下大幅度降低了资源的划分次数，提高了推荐效率。但实际可能出现用户可能感兴趣的资源有一部分在备选集I_2中，导致推荐质量有所下降。因此在实际资源划分中，应该考虑标签的关联关系，通过相关联的标签进行统一的划分，因此在实际应用中我们希望对资源划分后出现的结果是资源备I_1和I_2的数量基本持平，即图(b)的划分方式。采用该方式对资源进行划分，资源备选集的划分效率可能有所下降，但是可以通过用户再次的交互定位用户的兴趣资源。



（a）标签属性 A 划分资源　（b）标签 A 关联属性划分资源

图 4-4 资源分区

通过对出版行业的标签数据进行分析，出版社对于不同的资源进行了的归类划分，系统中每个标签还具有对应的分类属性，对于描述用户兴趣主题时可以使用该分类属性，同时通过标签也能快速的定位到其所属的分类。例如“音乐”标签属于“文化艺术”类，而“小学作文”属于“小学阅读”类。在数据集中规范化的标签存储格式如下。

表 4-1 标签格式

标签标识	分类标识	标签名称	描述	编码
Tag_id	Classification_id	Tag_name	Tag_desc	Tag_code

通过对资源-标签矩阵的分析，我们可以将整个资源备选集按类别进行划分。

由于某些标签分类可能无法划分整个资源集，因此我们只考虑可以覆盖资源项的类别。

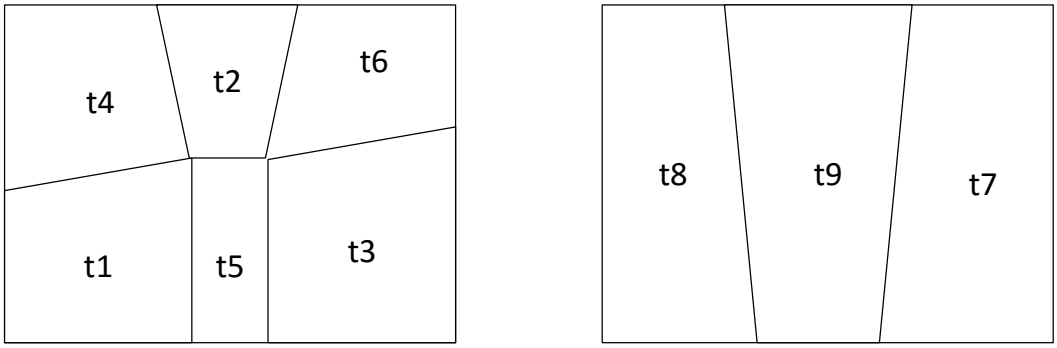


图 4-5 不同类别标签的资源集划分图

如图 4-5 所示，c1 包含标签$\{t_1,t_2,t_3,t_4,t_5,t_6\}$，c2 包含标签$\{t_7,t_8,t_9\}$，整个资源集可以根据不同的类别进行划分，多边形的面积代表资源数量。基于信息论，这两种不同标签的类别可以提供不同的信息熵。假设资源集合能被 c 类标签覆盖，标签描述为$t_1^c,t_2^c,\ldots t_k^c$。对于不同类别标签划分的信息熵计算方式如下：

$$E_c=-\sum_{i=1}^{k}p_i\log p_i \tag{4-3}$$

其中，$p_i=n_i/n$，n_i表示被标签t_i^c标记的资源数量。在交互过程中，系统首先选择具有最大值熵的标签类别，并根据标签的关联分析将该类别相关的标签提供给用户进行选择。当用户做出决定后，通过标签的关联分析将该标签关联的资源作为备选集，然后对备选集进行排序，推荐给用户。若用户找不到想要的资源，进行下一轮的交互。资源备选集划分详细步骤如下：

（1）计算对资源备选集进行划分的标签分类信息熵。

（2）选取熵最大的标签类别供用户交互。

（3）通过交互标签的关联分析，对资源备选集进行划分，同时更新标签选项，等待下次交互。

4.3 资源备选集重排序

交互式推荐算法中，资源备选集的划分和备选资源重排序是其中要解决的关键问题。通过概率矩阵分解算法对资源进行了评分，然后通过用户的交互操

作对资源备选集进行了划分，然后就是资源重排序。因为资源是基于标签关联分析进行划分的，因此将通过对标签的加权对资源进行重排序，更加准确的获取用户可能感兴趣的资源。

用户每次交互是对资源备选集的再一次划分，推荐资源列表也将进行更新，本文采用的是基于标签关联分析的交互方法，由于标签的关联分析可能出现资源集合扩大现象，因此将通过标签加权的方法对备选资源进行重新排序，更加显著的突出用户的交互选择权。如何对不同类别的标签或者相同的类别下的标签进行加权，本文将通过对首选标签进行加权操作，通过标签关联分析的置信度进行加权。由于标签具有分类属性，传统的加权方式可能会对同类标签或者不同类标签进行区分操作，本文将通过分析的关联度进行统一加权。例如标签 A 属性为“机器人”，标签 B 属性为“科技小说”，这两个标签都属于标签“高科技”类别下的子标签。标签 C 属性“科技电影”与标签 A 和 B 则属于不同的类别标签。

本文采用的是基于关联分析的资源划分方法，当用户选取标签属性 A 会对满足置信度和支持度的相关属性 B 和 C 同时进行考虑，对于相同类别和不同类别下的关联标签，资源加权可以分为 $A+B$ 、$A+C$ 、A 、B 、C 。其中，$A+B$ 是相同类别下的标签属性，表示同时含有标签 A 和 B 的资源。$A+C$ 是不同类别下的标签属性，也表示含有标签 A 和 C 的资源。然后对备选资源通过加权排序后，将资源评分较高的资源推荐给用户。如表 4-2 所示，部分标签关联分析结果。

表 4-2 标签关联分析

标签标识	关联标签标识	置信度
35	70	0.86
35	896	0.82
…	…	…

其中标签标识 35 表示的是“机器人”标签，标签标识 70 表示的是“科技小说”标签，而标签标识为 896 表示的是“科技电影”标签。通过对标签的关联分析后将资源进行重排序然后推荐给用户。对于新用户可能缺少评分信息，将获取用户对资源的平均评分，然后对平均评分的资源进行重排序。通过对标签的关联分析和资源的划分，图 4-6 给出了交互推荐的详细框架，将本文的交互式推荐框架和改进的概率矩阵推荐算法进行结合。首先是对数据的处理，接着

通过改进的特征提取算法获取用户和资源的近邻集合。然后将近邻关系与传统的概率矩阵分解算法结合，提出了基于标签的概率矩阵分解算法，提高了推荐算法的准确率。若用户对推荐资源不满意，则分析标签的关联关系，对资源备选集进行划分，进而提高推荐的效率。

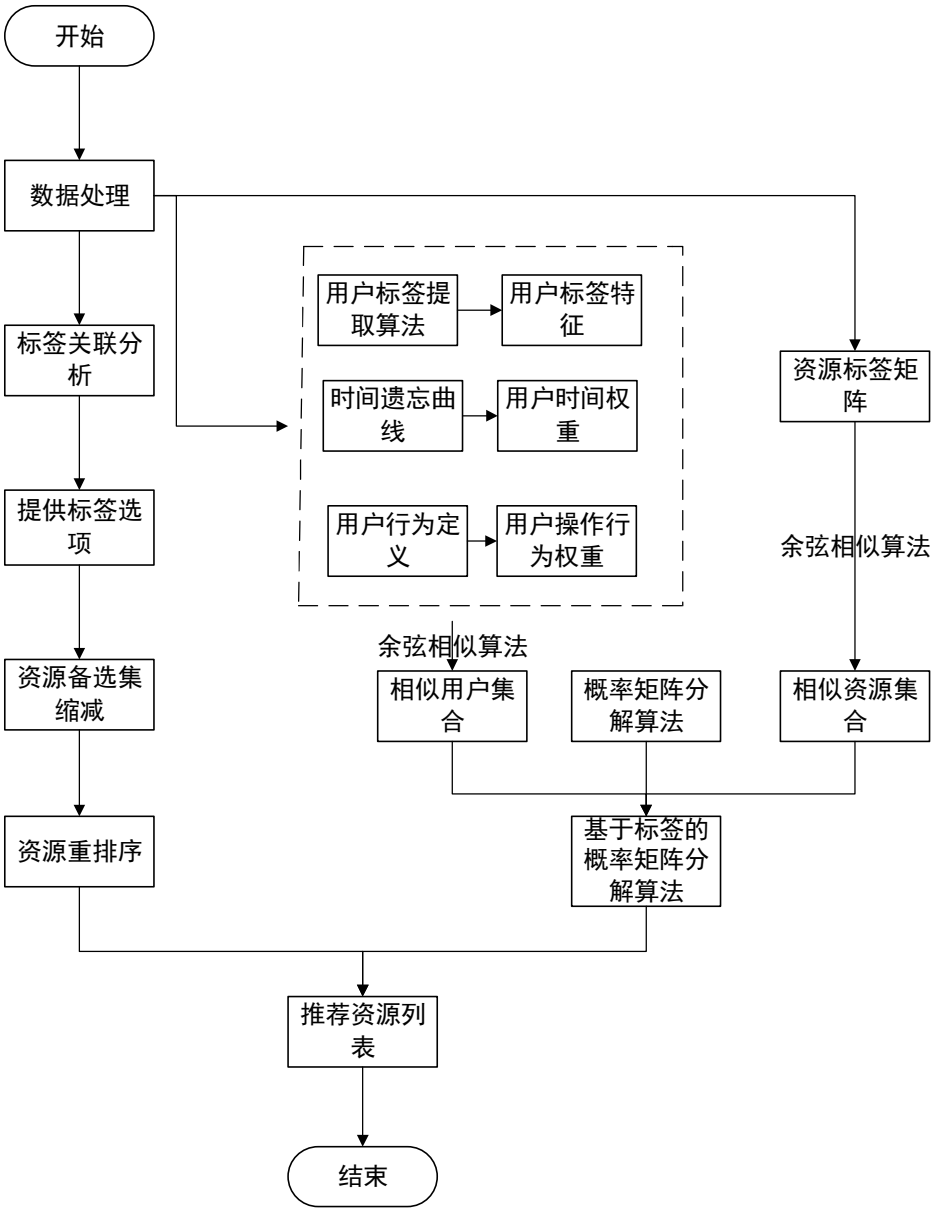


图 4-6 交互式详细推荐流程

表 4-3 将给出基于标签关联分析的交互式算法步骤。算法主要分为四步。第一步是对用户进行判断，非新用户采用基于标签的概率矩阵分解算法，新用户

采用决策树分类算法，这将在下文进行详细说明。然后是对用户是否交互的判断，用户启用交互操作后，根据标签的关联分析对备选资源进行划分，在考虑用户首选标签的同时考虑关联性强的标签属性进行划分，用户感兴趣资源的快速定位，因此提高了整个推荐过程的效率。

表 4-3 算法步骤

输入：备选集（包含评分信息的资源集合），Tag-list（标签列表，每次交互输入一个）
输出：Top-N List（推荐列表）
1、判断是否新用户，非新用户通过基于标签的概率矩阵分解算法给出推荐列表，若是新用户通过冷启动算法给出推荐列表。 2、用户是否接受推荐列表，如果接受推荐结束，否则给出标签进行交互（首次交互标签通过信息熵计算获取）。 3、通过用户选择的标签进行标签关联分析，同时对资源备选集进行划分，缩减资源备选集。 4、对划分后的资源备选集进行标签关联分析加权重排序，将排序的资源列表推荐给用户，回到第 2 步。
返回 Top-N List

4.4 冷启动问题解决策略

4.4.1 冷启动问题

本文主要研究的是基于模型的协同过滤算法，因此存在在冷启动问题，其主要原因是用户-资源评分矩阵中新用户和新项目的评分项都为空。新项目冷启动原因是无人对该资源进行评分，导致无法判断该资源是否被用户喜好，因此产生新项目冷启动问题。新用户冷启动问题则是缺失用户信息，无法构建用户的特征模型，导致无法寻找相似用户为其推荐个性化的资源。

对于新项目冷启动问题，传统的协同过滤算法中，无法寻找出新项目的特征，导致无法建立资源的特征模型获取相似邻居。本文研究的出版资源种类较多，采用传统的协同过滤算法无法提取特征模型，因此本文采用具有规范化标签的特征。通过标签特征进行新项目的预评分操作，首先根据该资源的标签特征寻找相似资源，本文采用寻找相似的 Top-10 资源，然后通过资源的平均评分

为资源进行评分，最后为用户推荐评分较高的资源。本文研究的出版资源具有规范化的标签集合，因此新项目冷启动问题可以得到有效的解决，本文也不将进行赘述。

对于新用户冷启动问题，往往是由于缺少用户的信息导致无法构建用户的兴趣特征，因此无法寻找出相似邻居，为新用户推荐感兴趣的资源，新用户冷启动问题也是推荐算法的重要研究方向。对于推荐算法的新用户冷启动问题，一般采用大众排名算法，即将大多数用户感兴趣的资源推荐给新用户，但是该方法忽略了用户的个性化特征，导致推荐结果准确率较低。本文通过决策数分类算法和交互操作解决新用户冷启动问题，具体算法处理过程如图 4-7 所示。

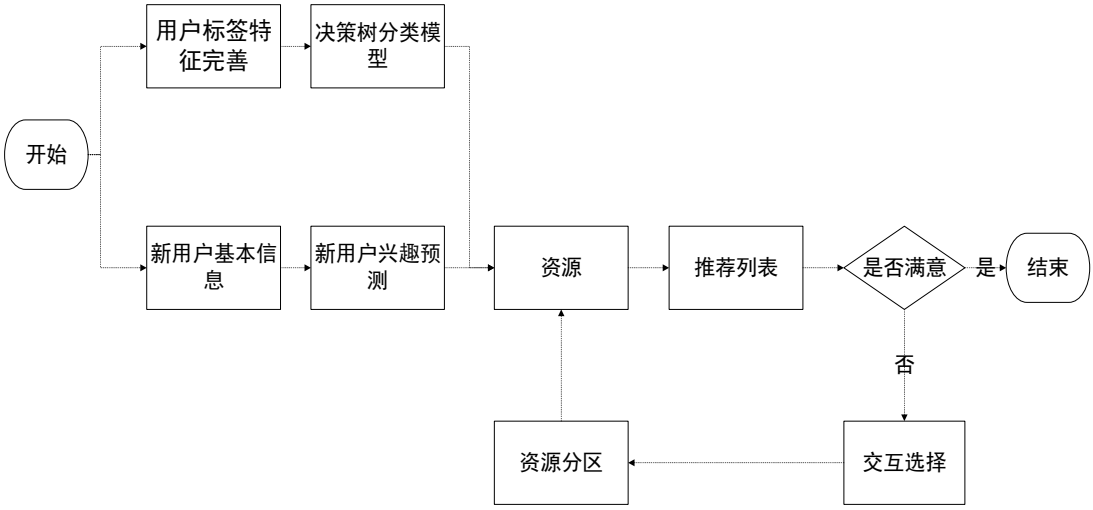


图 4-7 新用户推荐流程

如图 4-7 所示，对于新用户冷启动问题，本文给出了基于决策树分类和交互结合的解决策略。首先利用已有的用户信息构建决策树分类模型。当用户进入系统后，判断用户是否为新用户，若用户是新用户匹配决策树分类模型，根据匹配的信息为新用户推荐感兴趣的资源列表，若用户对推荐的资源不感兴趣，则系统提供标签供用户进行交互式选择，直至用户对推荐的资源满意为止。通过采用决策树分类模型和用户交互可以有效的解决新用户冷启动问题。

4.4.2 用户兴趣分类

新用户进入系统后首先进行决策树匹配，推荐感兴趣的资源列表。决策树构建过程中需要获取用户的兴趣主题，其中兴趣主题通过用户的标签特征进行

加权获取，选取分类属性权重大的分类标签作为用户的兴趣主题标签。

一般推荐系统获取用户的基本信息会采用问卷或者用户注册等方法。对用户而言注册所需的个人信息越多，用户的抵触就越强，造成的虚假信息也越多，因此本文推荐过程中只简要的收集用户性别和年龄信息，其它信息通过用户在系统操作时获取，降低过多收集用户信息而引起的反感。本文主要考虑用户的年龄、性别、阅读时间段和阅读地点，通过四个基本因素构建决策树分类模型。

阅读时间段主要是描述用户在哪个时间点进行阅读，用户可能在上午上班路上看一些时政信息，晚上在家可能会读一些文学休闲类文章，因此用户在不同时间阅读习惯也会有较大的区别。阅读地方主要是获取用户的地理位置，不同地理位置用户的阅读习惯也会有较大的差异，若一用户长久待在湖北，收集信息时检测到用户在海南阅读，说明该用户可能在旅行，因此可以根据当地的情况对用户进行推荐。通过对用户基本属性信息的分析，假设用户的年龄、性别、阅读时间和阅读地点相同，那么用户的兴趣主题也可能相同。

本文分析的四个因素都可以较为简单的获取，用户的阅读时间段和阅读地点可以通过移动端获取，年龄和性别在用户注册时获取。对于用户阅读时间段本文将用户一天阅读的时间划分为五个时间段：$t_1(08:00-11:00)$、$t_2(11:00-14:00)$、$t_3(14:00-17:00)$、$t_4(17:00-21:00)$、$t_5(21:00-08:00)$。对于年龄段也进行划分，年龄段划分为 8 阶段：$a_1(0-7)$、$a_2(8-11)$ $a_3(12-18)$、$a_4(19-25)$、$a_5(26-35)$、$a_6(36-45)$、$a_7(46-60)$、$a_8(61-)$。对于年龄在a_1的用户可能会对一些带图画的启蒙书籍比较高兴趣，而a_8阶段的用户可能就会对健康养生的资源或者预防疾病的资源比较感兴趣。因此本系统主要从这四个方面构建决策树分类模型。如表 4-4 是非新用户部分信息。

表 4-4 非新用户部分信息

用户标识	年龄段	性别	阅读时间段	阅读地方	兴趣主题
1001	a_4	男	t_1	湖北	科技
1005	a_4	女	t_2	北京	娱乐
1006	a_6	男	t_1	深圳	时政
……	……	……	……	……	……

通过对用户的基本属性分析后，获取用户的兴趣主题，构建决策树。首先分析用户的兴趣特征模型，获取用户每个标签的分类标识，根据标签的权重对

分类标签进行加权，最后获取用户的分类兴趣标签特征。

4.4.3 用户决策树构建

决策树是一种分类模型，它利用树形结构可以进行有效的分类。本文决策树采用 ID3 算法，对决策树各级节点通过计算的信息增益进行属性的选择。在决策树分类过程中，根节点往往表示的影响程度最大，而叶子节点是最终获取的样本类别值。

设 S 为一个数据样本集合，包含的数据样本数为 m 。其中 C_i 为类别， s_i 是类别 C_i 中的样本数，$i \in \{1,2,\ldots,m\}$ 。如果需要对样本中的对象进行分类，由式(4-4)获取其期望信息。

$$I\left(s_1,s_2,\cdots,s_m\right)=\sum_{i=1}^{m} p_i \log\left(p_i\right) \tag{4-4}$$

其中， p_i 是任意一个数据样本属于类别 C_i 的概率，可以通过 s_i / s 计算。设属性 A 具有 v 个不同的值。当前样本被属性 C_i 划分的信息熵如下。

$$E\left(A\right)=\sum_{j=1}^{v}\frac{s_{1j}+\ldots+s_{mj}}{s} I\left(s_{1j},\ldots,s_{mj}\right) \tag{4-5}$$

其中， s_{ij} 表示类别 C_i 所具有子集 S_j 的样本数。 $s_{1j}+\cdots+s_{mj}/s$ 是第 j 个子集的权值。子集划分的纯度由信息熵决定，信息熵越小划分越完美。

$$I(s_{1j},\cdots,s_{mj})=-\sum_{i=1}^{m} p_{ij}\log_2\left(p_{ij}\right) \tag{4-6}$$

其中， $p_{ij}=s_{ij}/\left|S_j\right|$ 。属性 A 对分支节点进行划分的信息增益如下。

$$gain(A)=I(s_1,s_2,\cdots,s_m)-E(A) \tag{4-7}$$

通过分析非新用户的标签特征和兴趣模型，获取非用户的兴趣主题，兴趣主题可以通过规范化标签的上级分类标签进行表示。

本文通过对非新用户构建决策树模型，对于新用户通过匹配决策树，分析新用户感兴趣的资源，其中决策树的构建成为关键步骤。在构建决策树前，对用户信息属性进行处理，使其具有离散性。然后通过决策树生成算法[48]，完成对决策树的构建。通过决策树算法对非新用户构建决策树，通过计算年龄的信息熵最大，也说明在本数据集中对于不同的年龄的用户其喜好的资源差异化较

大，因此也将年龄作为决策树的根节点。对于每个属性的信息增益，通过递归的方式对每个节点进行分类计算，同时按照子节点对父节点的信息增益选择样本属性值，直到没有多余的属性来划分训练样本。

本文将年龄喜好作为根节点，首先对于大于某一年龄的属性划分为左子树，其余划分为右子树。通过分析结果对非新用户构建决策树分类模型如图 4-8 所示。决策树构建完成后，新用户进入系统后通过进行匹配，系统将推荐其感兴趣的资源。若用户对推荐的资源列表不满意，则可以通过交互方法快速定位自己感兴趣的资源，通过此方法解决推荐过程中的新用户冷启动问题。例如新用户 u 进入系统后，判断该用户的年龄小于 a_3，将进入右子树，最终判断用户达到叶子节点 $class_k$，证明了该用户的兴趣主题属于 k，根据对非新用户的判断，推荐主题 k 的资源给用户。

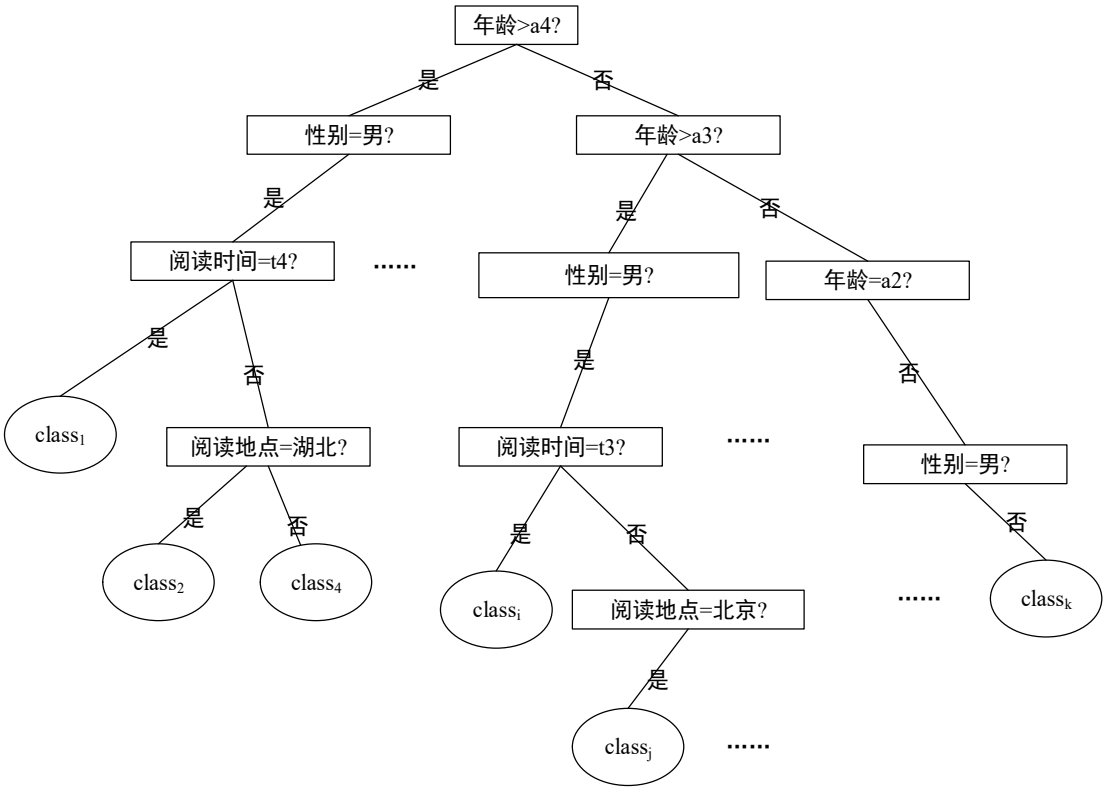


图 4-8 用户决策树模型

决策树分类器构建完成后，将对新用户冷启动问题进行解决。具体算法如表 4-5 所示。

表 4-5 新用户冷启动算法

输入：新用户信息（年龄、性别、阅读地方和时间），资源备选集，非用户信息（用户的标签特征、基本信息）
输出：Top-N List(推荐资源列表)
1、根据非新用户信息，获取用户的兴趣主题，构建决策树分类器。 2、新用户信息匹配决策树模型，根据匹配的兴趣主题给出推荐列表。 3、新用户对推荐的资源列表满意，冷启动算法结束。若用户对推荐的资源列表不感兴趣，提供标签列表供用户交互，启动交互式方法。 4、通过用户信息和交互的标签，对资源备选集进行缩减，推荐资源给用户。若用户对其满意则冷启动算法结束，否则进行第 3 步。
返回 Top-N List

第一步是构建决策树分类模型，获取用户的标签特征模型，通过用户的标签特征获取权重较大的分类标签作为其兴趣主题，通过用户基本信息和兴趣主题构建决策树分类模型。第二步是当用户进入系统后，判断用户是否为新用户。如果是新用户则将用户的基本信息与构建的决策树匹配，分析用户的兴趣主题，将兴趣主题中评分较高的资源推荐给用户。第三步判断用户是否对推荐资源满意，若用户对推荐的资源满意，则新用户冷启动推荐算法结束，同时记录用户的行为完善特征模型。若用户对推荐的资源不满意，则提供标签供用户进行交互操作，交互过程对资源备选集进行了缩减，资源的评分选取用户的大众评分的均值，接着进行资源重排序算法，推荐资源列表给用户。

4.5 实验结果分析

4.5.1 新用户冷启动推荐实验

为了证明本文冷启动解决方案的有效性，进行了用户冷启动实验。在实验数据集中随机选择 200 名用户，并将他们的行为记录全部删除，模拟新用户，以此对算法的推荐性能进行分析。新用户可能对推荐的结果不满意，进行交互操作，选择 200 名中的 50 名用户，添加 5 次标签，以模拟交互产生的标签特征。本实验对比算法是基于平均值算法(AVG，将用户对资源的评价取平均值)的大众

排名、基于人口统计学的推荐算法(DR，Demographic-based Recommendation)，基于文献[49]进行计算。本文冷启动算法称为 DT_IR。对比实验的准确率和召回率和 *F* 值与推荐列表长度的关系如图 4-9、图 4-10 和图 4-11 所示，横坐标表示推荐列表长度。

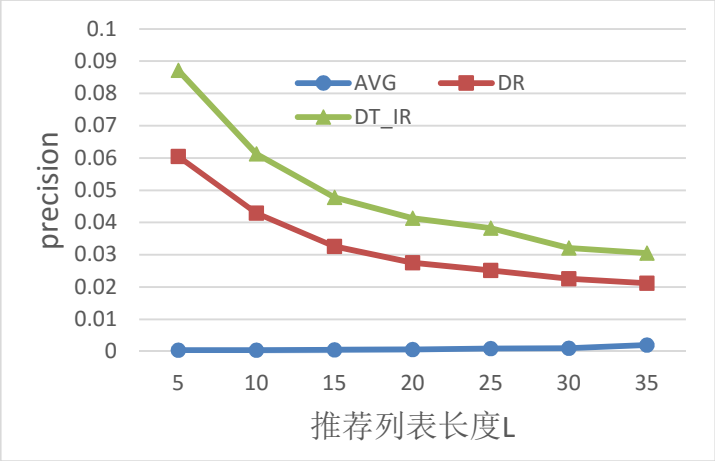


图 4-9 precision 对比实验

分析图 4-9 中的实验结果，可以观察到在相同推荐列表长度下 DT_IR 算法的准确率最高，而 AVG 算法的准确率最低。随着推荐资源数量的增加，DT_IR 算法和 DR 算法的准确率都呈下降趋势，但 DT_IR 整体推荐准确率最高。AVG 算法在推荐资源较少的情况下接近于 0，说明在推荐资源较少的情况下采用大众的评分排名算法忽略了用户的个性化特征导致推荐结果质量不佳，当推荐资源增大的情况，推荐准确率才有所上升但是总体准确率呈现较低水平。通过对准确率的分析，本文的算法对于新用户冷启动问题的推荐准确率有较大的提升。

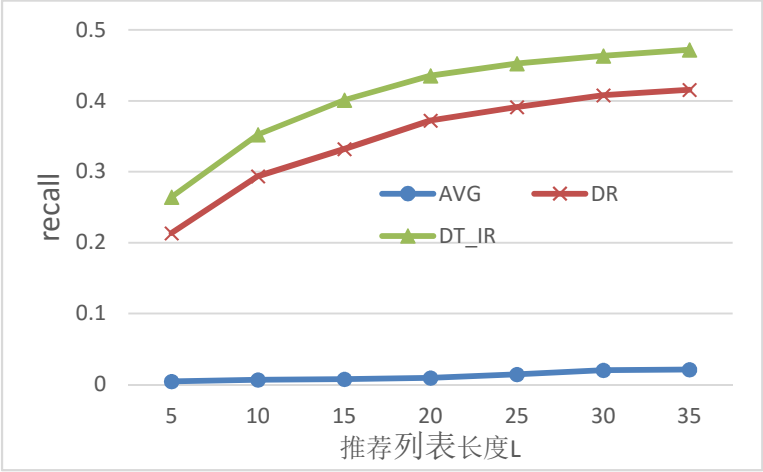


图 4-10 recall 对比实验

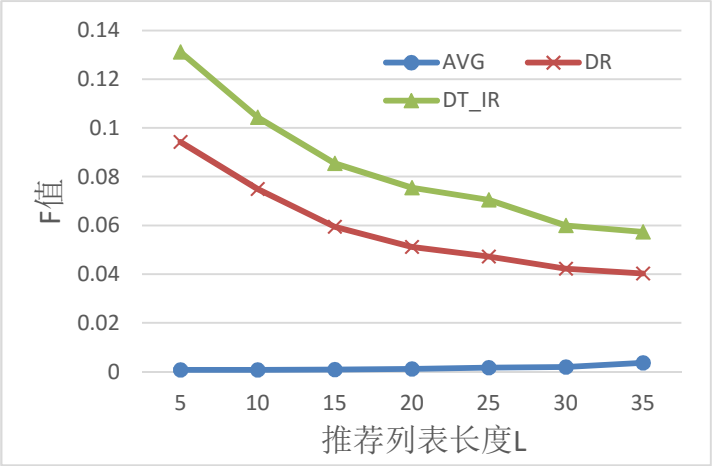


图 4-11 *F* 值对比实验

分析图 4-10 中的实验结果，随着推荐资源的数量增加，推荐算法的召回率都有所上升，其中 DT_IR 算法的召回率最高，AVG 算法的召回率随推荐资源数量的增加，提升的较少。这也反映在新用户冷启动问题中，传统的 AVG 算法的效果较差。DT_IR 和 DR 算法的召回率都随推荐资源数量的增加而增加，有趋于平稳的趋势。DT_IR 算法的召回率一直比 DR 和 AVG 算法的召回率高。分析图 4-11 中的实验结果，随着推荐资源数量的增加，DR 算法和 DT_IR 算法 *F* 值都有所下降，但本文 DT_IR 算法的 *F* 值相比其它两种算法均处于较高水平，AVG 算法 *F* 值虽然增加但整体维持较低水平。

在新用户冷启动问题上，通过对比实验可以发现 DT_IR 的准确率、召回率和 *F* 值最高，该算法相比于 DR 算法不仅考虑了用户基本信息还考虑了用户的兴趣特征，通过对系统中已有用户构建决策树分类模型获取其兴趣主题特征，当新用户进入系统对其进行决策树分析推荐相对应的主题资源，如果用户对推荐的资源不满意还可以通过交互的方式快速的获取推荐资源，其准确率和召回率也有较大的提升，有效的解决了用户冷启动问题。

4.5.2 交互式推荐实验

交互式推荐算法中交互次数是影响性能的关键因素。为了研究交互次数和推荐性能的关系，通过增加用户交互操作的标签来模拟用户交互，得到交互操作次数与算法的准确率和召回率的关系，推荐资源列表长度 L=10。具体实验的准确率和召回率与交互次数的关系如图 4-12、图 4-13 所示。

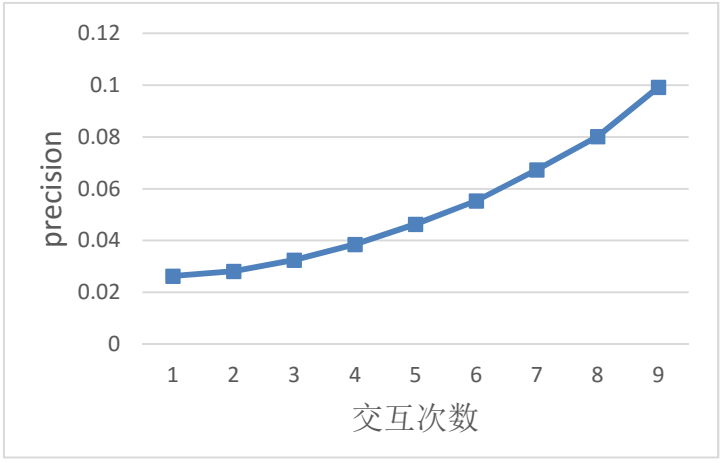


图 4-12 交互式推荐实验准确率

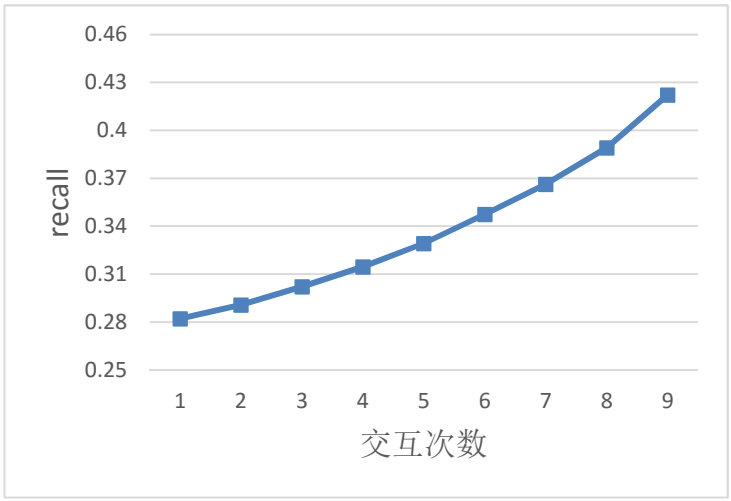


图 4-13 交互式推荐实验召回率

通过分析图 4-12 和图 4-13 中的实验结果，推荐准确率和召回率会随着交互的次数而发生变化，随着用户交互次数的增加推荐的准确率和召回率都有所提高，随着交互次数的增加准确率和召回率增加的越加明显。从实际应用的情况考虑，用户交互的次数越多，用户表达的想法也更加明确，系统对用户的需求了解也更加清楚。同时对资源备选集划分也越来越小，候选资源的范围也越来越小，因此准确率和召回率都有较好的提升。通过实验证明，交互式方法对推荐的准确率和召回率有所提升，当用户对推荐结果不满意时，通过交互方法和推荐的结合可以更快的满足用户获取知识服务的需求。

在交互式推荐实验中，根据交互次数的变化，推荐耗时也发生了变化。每次交互资源备选集将会缩减，从新获取推荐列表时间将降低，推荐耗时实验如

图 4-14 所示。

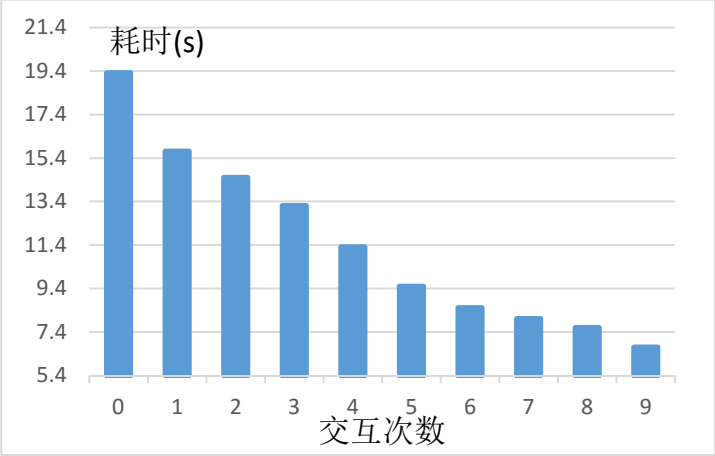


图 4-14 交互式推荐推荐耗时图

通过对交互式推荐推荐耗时图的观察，当交互次数是为零次时，即不采用交互式过程时，算法的耗时较多。随着交互次数的增加，资源备选集在一定程度上进行了缩小，推荐耗时也呈下降趋势，交　互次数越多推荐效率越高。由于推荐耗时不仅和算法的性能相关，同时也和实验的软硬件环境相关，所有这里也给出了交互耗时实验结果作为算法的评价参考。通过交互式推荐实验和交互耗时实验，说明本文提出的基于标签关联分析的交互式算法的有效性。

4.6 本章小结

本章首先对交互式方法进行了描述，给出了交互的具体流程。然后对标签的关联关系进行了研究，分析标签的信息熵和关联分析进而对资源备选集进行划分，以及备选资源的重排序，提高推荐效率。

针对基于模型的协同过滤算法中存在的冷启动问题进行了研究，本文通过规范化的标签属性对新项目采用资源相似性评分操作。对于新用户采用决策树分类与交互方法解决，首先分析非新用户的特征构建决策树分类器，新用户进入系统后进行匹配推荐资源，若用户不满意则通过交互快速获取资源。

最后进行了实验验证，首先对本文的新用户冷启动算法进行对比实验，在实际数据集上的结果证明了本文的新用户解决策略在准确率、召回率和 F 值上具有较好的结果。最后验证了交互式推荐算法中交互次数和推荐效率的实验，实验结果显示交互式方法提高了推荐结果质量，证明了本文提出的基于标签关联分析的交互式方法的有效性。

第 5 章 基于标签的交互式推荐系统

本文围绕基于标签关联分析的交互式推荐方法进行研究，并针对推荐算法和交互式方法进行了展开研究，通过实验证明了本文改进的概率矩阵分解算法、冷启动解决策略和交互式方法的有效性。基于本文的项目背景，本章将对出版资源交互式推荐系统进行设计与实现。首先对交互式推荐功能进行设计，进行了需求分析、详细设计和推荐流程设计，最后对系统的主要验证界面进行展示。

5.1 交互式推荐功能设计

5.1.1 需求分析

基于本文的项目背景，根据出版行业的特殊性设计了一个基于标签关联分析的交互式推荐系统，其最终的目的是为用户推荐资源列表。根据项目的需求，对系统进行主要模块设计，如图 5-1 所示。

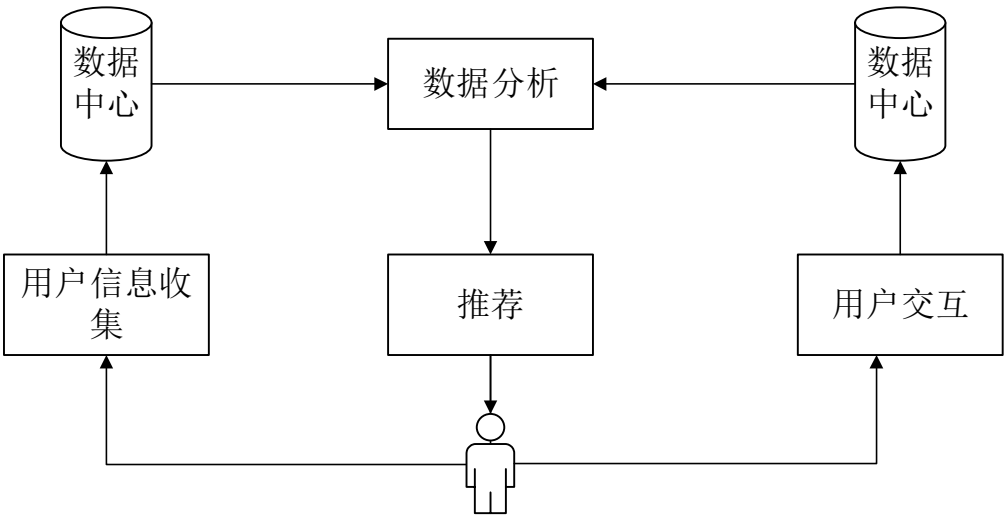


图 5-1 系统主要模块

用户信息收集功能主要用来收集用户信息，是推荐系统推荐的主要依据，它将收集用户的基本信息和有效的操作信息，通过收集到的用户反馈数据能准确的描述用户的兴趣特征。用户交互功能是用户对标签的交互操作，通过用户的交互可以快速的缩减资源备选集以提高推荐效率，同时构建用户目前的兴趣

特征。数据分析模块是通过用户的信息和交互信息进行数据的处理和加工，构建用户的兴趣特征，同时也对资源进行处理和加工。推荐功能则是推荐系统的主要功能，其目的是为用户推荐资源列表，若用户没有交互直接根据基于标签的概率矩阵进行资源推荐，若用户交互后，对资源备选集进行划分然后重排序，最后推荐资源列表给用户，达到快速响应用户的目的。

5.1.2 系统框架设计

本节将对系统进行框架设计和选择。系统的框架采用的是目前流行的分布式服务 Dubbo 框架，同时前后端分离，后端通过 REST 提供 API 接口。

Dubbo 是由阿里开源的分布式服务框架，采用 Dubbo 框架可以方便的构建分布式服务。用户根据自己实际业务应用场景，可以选择适合自己的集群模式，给广大中小型互联网公司带来了福音。Dubbo 框架的服务提供方可以进行天然的集群服务，为目前需要快速响应和高并发的场景提供很好的解决方法。该框架主要分为服务提供方、服务消费方和注册中心，它将服务进行了解耦，当一方的服务出现了故障不会引起其它服务的错误。如图 5-2 给出了整个项目的系统架构图，本文主要功能实现将在业务集群中进行代码编码。

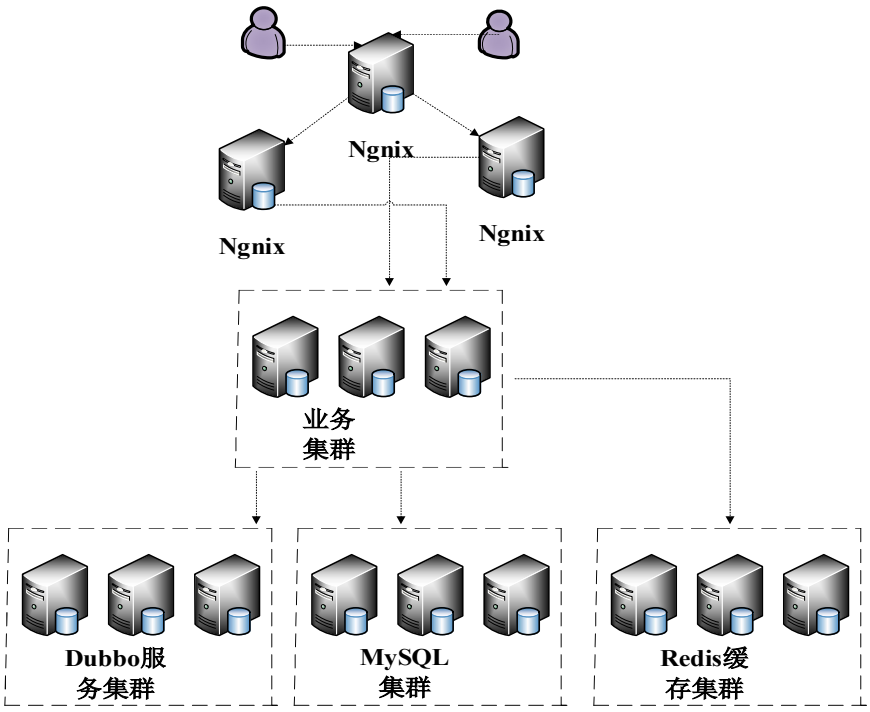


图 5-2 系统整体架构图

系统中每个单独的模块都是一个独立的服务，通过对服务进行模块化拆分可以有效的提高数据的处理速度。由于系统采用的是分布式架构风格，对每个单独的模块可以采用集群模型，当某个机器宕机后能继续保证服务的高可用。推荐系统的用户信息和用户操作反馈信息存储在持久层中，持久层采用传统的 MySQL 数据库和非关系数据库 redis，传统的数据库存储用户历史信息和日志信息，而缓存数据库 redis 主要用于存放热点数据，可以加快数据的读取，同时用户在交互式系统中通过交互可以快速的对用户感兴趣的资源进行划分和定位，可以通过对新用户的交互快速的构建出用户的兴趣特征模型。整个系统的持久层也采用集群模式，系统的瓶颈往往集中在数据库层，当用户和数据量过大时，集群模式可以有效的缓解系统压力，因此持久层也采用集群模式。模块间的通信则是通过 rpc 协议进行交互，通信速度快，也是实时交互推荐系统所必须的。

5.1.3 系统功能详细设计

系统的主要框架设计和选择后，将对系统的功能模块进行设计。如图 5-3 是推荐系统的详细功能框架，包含需求分析要求的四个主要模块。

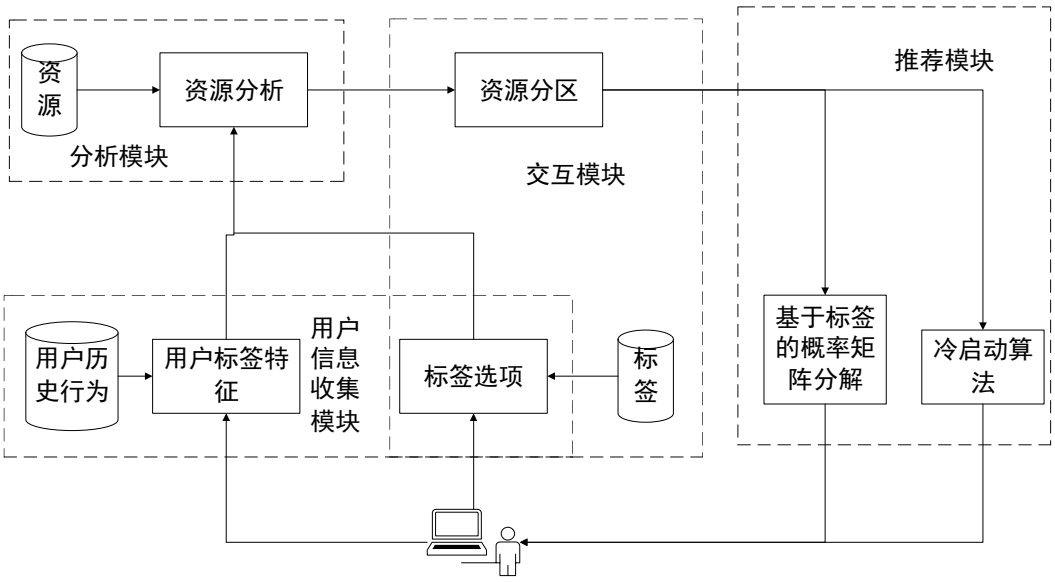


图 5-3 系统详细功能框架

用户信息收集模块：用户信息收集模块将会收集用户的历史行为信息、用户交互式选择信息，用户基本信息等。用户的历史行为信息主要包括对资源的

反馈，如阅读、收藏、评论、购买等。收集的用户信息将被收集存放于数据库持久层中，由统一的分析模块进行分析，构建用户的兴趣特征向量。

分析模块：本文采用规范化的标签，通过标签能描述资源的属性，本文将资源通过一维的标签特征向量进行描述。用户在系统中浏览资源时会有不同的操作行为，通过用户的行为可以被动的贴上资源的标签，同时根据用户行为权重、标签标注的时间、用户的交互行为等构建用户的兴趣模型，最后由用户的兴趣特征寻找相似用户。

交互模块：本文通过基于标签关联分析的交互式方法进行资源推荐，交互模块主要适用于用户的交互过程。用户如果对推荐的资源不满意可以通过交互快速定位感兴趣的资源，从而提高推荐的质量。系统提供规范化标签引导用户进行选择，用户对标签的选择能表示用户目前的兴趣特征，在用户交互过程中也进一步完善了用户的特征模型。

推荐模块：基于本文的算法研究，对于非新用户采用基于标签的概率矩阵分解算法进行推荐，本文在传统的概率矩阵分解算法的基础上对其进行改进与优化，首先通过规范化的标签获取用户和资源的特征向量，然后寻找用户和资源的近邻关系，将用户和资源的近邻关系融入到评分矩阵中进行概率矩阵分解，重构评分矩阵，最后为用户推荐感兴趣的资源。新用户则通过决策树分类推荐然后进行交互。

5.2 交互式系统推荐流程设计

通过对本文推荐算法的分析，设计了交互式系统的推荐流程。图 5-4 对交互式系统推荐流程进行详细的说明。当用户进入系统后，会对用户进行区分，如果是新用户采用决策树分类算法，计算用户的兴趣特征主题，推荐相关主题的资源列表。如果是非新用户会进行数据处理，分析用户的标签特征、行为特征、时间权重构建用户的兴趣特征向量，同时根据资源-标签矩阵构建资源的特征向量，然后计算用户和资源的相似邻居，接着采用本文改进的基于标签的概率矩阵分解算法为用户推荐资源。用户的交互操作不区分新旧用户，都提供标签供用户交互，若用户对推荐的资源满意则停止交互，否则直到用户对推荐的资源满意为止。第二章对交互式框架进行了设计，其中用户信息和标签信息等非实时性的信息会放于线下进行分析处理，同时用户和资源的特征模型在线下进行计算，线上只进行标签交互选择和推荐过程，从而加快整个系统的推荐速度，

满足用户快速获取知识服务的需求。

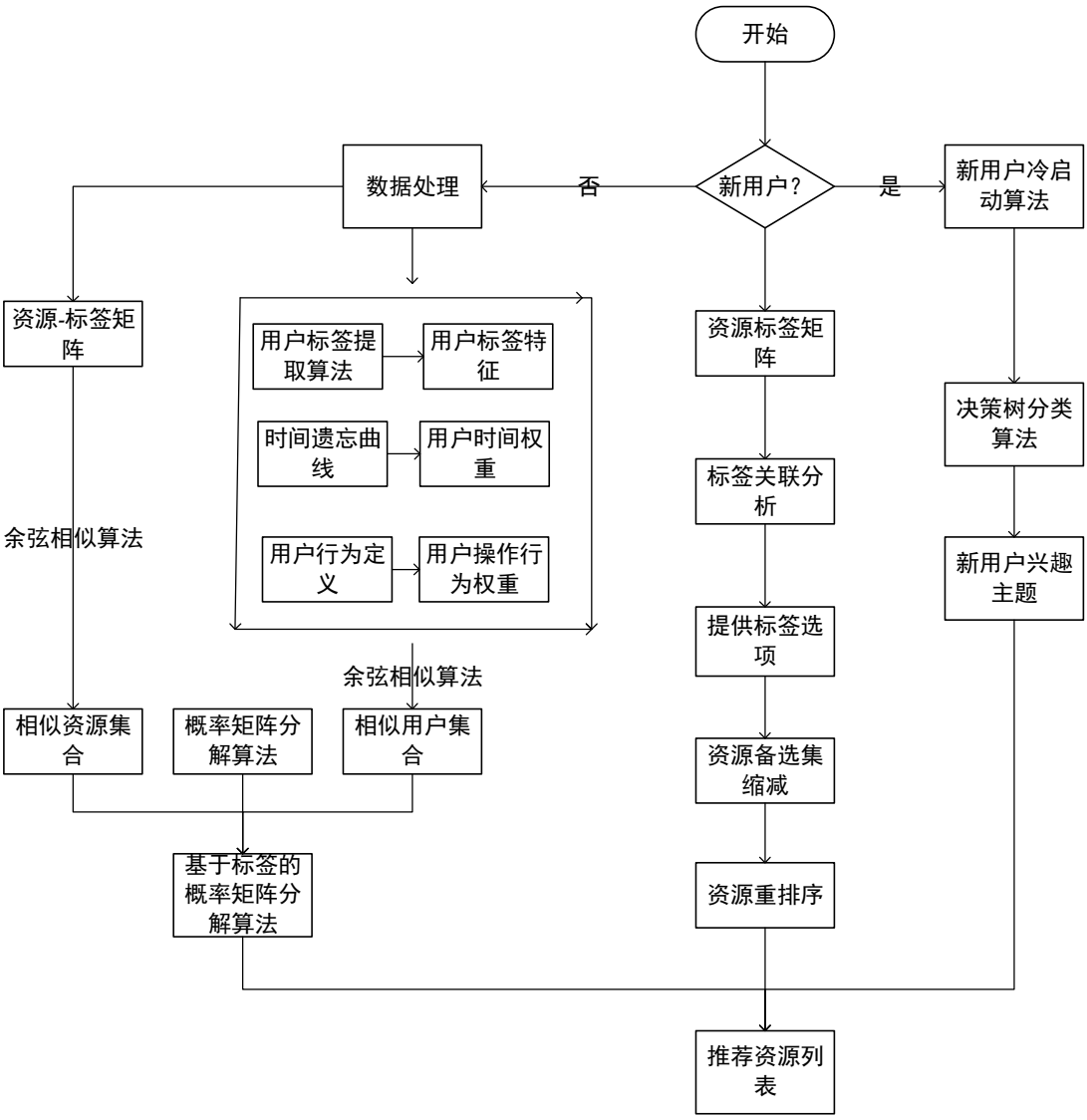


图 5-4 交互式系统推荐流程图

5.3 系统主要验证界面

基于本文的项目背景设计了基于标签关联分析的出版资源交互式推荐方法，对算法进行了研究证明了本文算法的可行性，然后进行了交互式系统的设计，其中包括系统框架的选择和详细功能的设计，并对整个推荐流程进行了分析，通过推荐流程和功能的分析，最后将对系统实现展示和检验，根据不同的需求将系统分为管理端和客户端。管理端为管理人员分析和管理用户，客户端则为

用户提供出版资源。

5.3.1 用户信息分析验证

首先对管理端进行功能实现和验证，图 5-5 展示的是系统的用户。其中管理端主要目的是对用户的信息进行分析，了解用户的特征。

图 5-5 系统读者展示界面

系统的信息收集模块会收集用户的基本信息，并将所有用户进行简要的信息展示，读者查询界面可以发现读者关注的出版社，了解读者的类型，也是本文算法中构建的用户兴趣主题。

如果要分析用户的详细信息，可以进入用户的详情分析页面，该页面会记录用户的行为操作，本文对资源的隐式评分也就是来源于用户的操作行为。本文对用户“大白(lily)”进行分析，当需要获取用户的详细信息，管理端的编辑或其它出版行业工作者只需进入用户的详情界面，如图 5-6 对用户的详细信息进行展示。

图 5-6 用户详细信息展示页

图 5-6 记录了用户“大白(lily)”2017/12/31 到 2018/1/3 的行为记录。通过用户的行为，数据分析模块进行分析，得到用户的标签特征，如图 5-7 所示，通过对用户的标签进行分析得到用户主要标签特征，然后通过本文的基于标签的概率矩阵分解算法为用户推荐资源。

图 5-7 用户标签特征界面

5.3.2 用户交互推荐验证

当用户在系统中进行浏览、购买、评论、收藏等操作后，系统的信息收集模块会收集用户的历史信息，然后数据分析模块分析在后台分析用户行为对用户进行建模，进而为用户推荐其感兴趣的资源。系统后台对用户进行数据分析，然后通过客户端将资源推送给读者。如图 5-8 所示。

图 5-8 客户端交互推荐界面

用户“大白(lily)”进入系统后首先进入书城页面，上方是交互式按钮，然后是系统的基本功能，包含文化墙、资源的分类，最后是系统为用户推荐的资源列表。通过推荐的资源列表，用户可以选择资源进行阅读、购买等操作。如用户对推荐的资源不满意，用户可以点击交互式按钮，进入交互页面。交互页面为用户展示的是交互标签和其他读者喜欢的资源，用户点击标签可以进行交互式操作，选择标签进行交互后，系统对资源备选集进行划分与重排序，用户每次将会降低资源备选集进行缩减，因此交互后推荐的效率越来越快。

例如：用户“大白(lily)”对此次的推荐结果不满意，通过标签交互后，了解用户此次想获取游戏类的资源，推荐系统在缩减后的资源备选集中给用户推

荐。如图 5-9 所示，展示的是用户“大白(lily)”交互后得到的推荐结果。

图 5-9 客户端资源详情界面

假设用户对《互动圈加入游戏互动圈，看本书游戏的不同玩法》感兴趣，用户点击进入详细页面，该页面将会有作品的详细介绍，在该页面用户可以进行购买，然后就可以阅读购买后的资源，若是免费作品，用户可以直接进行阅读。通过管理端和客户端的实际系统展示，分析了用户的兴趣特征，然后为用户进行资源推荐。

5.4 本章小结

本章主要是对基于标签关联分析的出版资源交互式推荐算法的具体系统实现。首先基于项目的背景进行需求分析，然后对整个系统进行了框架和流程设计，分析了各模块的详细功能。最后给出了系统的主要验证界面，将系统分为管理端和客户端，管理端主要分析用户信息，客户端为用户进行资源推荐，方便用户获取自己所需的资源。

第 6 章 总结与展望

6.1 工作总结

随着互联网的飞速发展导致信息爆炸，推荐系统对解决信息服务行业个性化内容的生产和传播大有助益。特别是对于类别多样化、特征难以表示的出版资源而言，进行资源个性化推荐研究有很好的现实意义与理论意义。因此本文重点对出版资源个性化推荐方法进行了深入研究，本论文的主要工作总结如下：

（1）用户特征提取算法改进。针对出版行业资源的特点，构建了资源-标签矩阵，通过用户的隐式反馈构建用户-资源评分矩阵以及用户偏好的标签矩阵。针对传统推荐算法中用户特征提取不准确的问题，本文通过用户行为操作权重、用户的标签特征、时间权重综合考虑，构建完善的用户兴趣模型，同时介绍了基于 TF-IDF 的用户标签特征提取方法。

（2）概率矩阵分解算法改进。首先介绍了概率矩阵分解算法的应用和优缺点，然后对其进行了改进。将本文规范化的标签应用于用户的特征提取中，通过改进的特征提取算法获取用户和资源邻居集合，并融入到概率矩阵分解模型中，充分考虑用户和资源的影响关系，通过梯度下降算法得到每个用户和资源的特征向量，重构用户-资源评分矩阵，提高推荐算法的准确率。

（3）基于标签关联分析的交互式方法分析与框架设计。针对传统推荐算法随资源数量增加导致的推荐效率下降问题，用户对推荐资源不满意问题，提出了基于标签关联分析的交互式方法，设计了算法了整体推荐框架。通过标签的关联分析对资源备选集进行划分，在用户交互的过程中缩减资源备选集，完善用户信息，提高推荐的效率。

（4）冷启动问题解决策略。本文主要针对新用户冷启动问题提出了解决方案。首先对非新用户进行决策树分类，当新用户进入系统后对其进行决策树分类分析，推荐新用户感兴趣的主题资源，同时新用户可以通过交互快速获取感兴趣的资源，有效的解决了新用户冷启动问题。

（5）实验对比分析。在出版行业真实数据集下进行算法的实验验证，其中推荐算法部分包含参数 λ 的影响实验、邻居数量 D 的影响实验、不同特征向量维度 K 的对比实验、推荐长度 L 对比实验、算法耗时对比实验。交互式部分包含新用户冷启动实验、交互式实验。通过大量实验验证了本文提出的基于标签

关联分析的出版资源交互式推荐算法的正确性，基于标签概率矩阵分解算法的有效性，同时证明了本文冷启动解决策略的可行性。

（6）交互式推荐系统实现。针对本文提出的基于标签关联关系的出版资源交互式推荐方法的系统性验证，分析了推荐系统的需求关系，并设计了系统的整体框架和流程，最后对推荐系统的各模块进行了系统展示。

6.2 工作展望

本文针对出版行业提出了相关个性化推荐方法，并对现有的推荐技术存在的缺陷进行了研究和改进，但是本文方法在实际应用中仍存在缺陷，主要表现在以下几个方面。

（1）本文提出的规范化标签不依赖用户的反馈标签，并不具有大众标签的自由标注的特性，在一定程度上杜绝了无效标签和冗余标签的问题，但对资源创建者有着要求，增加了资源提供方的负担，后期研究可以考虑在保证标签准确率的前提下，减少资源创建者的工作。

（2）本文介绍的算法虽然分析了标签的关联关系，并通过标签的关联关系对资源进行分区，但是随着资源和标签信息的增长，对资源的划分和标签关联分析需要进一步优化。

（3）在本文改进的概率矩阵分解算法中，对于用户的相似度和邻居的选择是无向的，在实际应用中用户间的相互影响因素是不同的，在下一步的研究中可以考虑用户的社交关系和信任关系。

致谢

三年的研究生生涯，转眼即逝。论文工作已进入最后的阶段，马上就要离开校园，内心不禁感慨良多。在过去的三年中，遇到了各种困难，是老师们的谆谆教诲，同学和朋友的热心关怀以及家人的支持和爱护使得我克服一个个的困难，不断的成长。在此，我将对他们表示最诚挚的谢意。

首先，衷心感谢我敬爱的恩师刘永坚老师，在三年的研究生阶段，刘老师悉心指导，耐心授教，给我提供了很好的学习成长机会，让我在思维方法、学习能力及科研能力等方面得到了全面的锻炼和提高。刘老师严谨务实的科研态度、广阔的视野和高度的责任心，不仅让我的学生生涯受益匪浅，对于我今后的人生道路也会有很大的帮助，鞭策我努力学习。接着要感谢解庆老师，解老师在论文的选题、研究和撰写的过程中给予我诸多帮助，给我提出了许多宝贵的建议和修改意见。然后感谢中心的白立华老师，白老师在三年的研究生生活和学习中给予了非常大的帮助，提供了良好的环境供我们学习，提高了我们的实践能力。

感谢在读研期间给予我帮助的实验室同学们、中心同事们，感谢你们在学习，工作和生活中对我的帮助和教导，感谢你们这三年来的陪伴和关心。

最后，衷心感谢各位评审专家在百忙之中能够对我论文的细心审阅，向各位参加论文评审和答辩的老师专家表达诚挚的谢意！

参考文献

[1] 戴俊潭. 传播逻辑与受众诉求:数字化阅读时代的出版创新[J]. 出版发行研究, 2015,1(6):24-27.

[2] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003, 7(1):76-80.

[3] Meo P D, Quattrone G, Ursino D. A decision support system for designing new services tailored to citizen profiles in a complex and distributed e-government scenario[J]. Data & Knowledge Engineering, 2008, 67(1):161-184.

[4] Capuano N, Gaeta M, Ritrovato P, et al. Elicitation of latent learning needs through learning goals recommendation[J]. Computers in Human Behavior, 2014, 30(1):663-673.

[5] Nguyen T T S, Lu H Y, Lu J. Web-Page Recommendation Based on Web Usage and Domain Knowledge[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(10):2574-2587.

[6] 张佳乐, 梁吉业, 庞继芳, 等. 基于行为和评分相似性的关联规则群推荐算法[J]. 计算机科学, 2014, 41(3): 36-40.

[7] 薛福亮, 马莉. 利用动态产品分类树改进的关联规则推荐方法[J]. 计算机工程与应用, 2016, 52(4): 135-141.

[8] 江周峰, 杨俊, 鄂海红. 结合社会化标签的基于内容的推荐算法[J]. 软件, 2015,18(1): 1-5.

[9] Koohi H, Kiani K. User Based Collaborative Filtering using Fuzzy C-Means[J]. Measurement, 2016, 91(1): 134-139.

[10] Salehi M, Kmalabadi I N. A Hybrid Attribute-based Recommender System for E-learning Material Recommendation[J]. Ieri Procedia, 2012, 2(4): 565-570.

[11] Imielienskin T, Swami A, Agrawal R. Mining association rules between set of items in largedatabases[J]. ACM Sigmod Record, 1993, 22(2):1-14.

[12] 文俊浩, 何波, 胡远鹏. 基于社交网络用户信任度的混合推荐算法研究[J]. 计算机科学, 2016, 43(1):255-258.

[13] 黄璐, 林川杰, 何军,等. 融合主题模型和协同过滤的多样化移动应用推荐[J]. 软件学报, 2017, 28(3):708-720.

[14] Zhang J, Lin Y, Lin M, et al. An effective collaborative filtering algorithm based on user preference clustering[J]. Applied Intelligence, 2016, 45(2):230-240.

[15] Zhang Z, Kudo Y, Murai T. Neighbor selection for user-based collaborative filtering using covering-based rough sets[J]. Annals of Operations Research, 2017, 256(2):1-16.

[16] 印鉴, 王智圣, 李琪, 等. 基于大规模隐式反馈的个性化推荐[J]. 软件学报, 2014(9): 1953-1966.

[17] 黄震华, 张佳雯, 田春岐, 等. 基于排序学习的推荐算法研究综述[J]. 软件学报, 2016, 27(3): 691-713.

[18] 于洪, 李俊华. 一种解决新项目冷启动问题的推荐算法[J]. 软件学报, 2015, 26(6): 1395-1408.

[19] Zhou K, Yang S H, Zha H. Functional matrix factorizations for cold-start recommendation[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011:315-324.

[20] Shi L, Zhao W X, Shen Y D. Local Representative-Based Matrix Factorization for Cold-Start Recommendation[J]. ACM Transactions on Information Systems, 2017, 36(2):1-28.

[21] 杨秀梅, 孙咏, 王美吉,等. 新闻推荐系统中用户冷启动问题的研究[J]. 小型微型计算机系统, 2016, 37(3):479-482.

[22] 郭晓波, 赵书良, 牛东攀,等. 一种解决稀疏数据和冷启动问题的组合推荐方法[J]. 中国科学技术大学学报, 2015, 45(10):804-812.

[23] Ma H, King I, Lyu M R. Learning to recommend with social trust ensemble[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2009:203-210.

[24] Zhou T C, Ma H, Lyu M R, et al. UserRec: A User Recommendation Framework in Social Tagging Systems[C]// Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, Usa, July. DBLP, 2010:1486--1491.

[25] 郭磊, 胡燕. 基于人工蜂群的项聚类推荐算法[J]. 微电子学与计算机, 2017, 34(6):31-35.

[26] 朱昆磊, 黄佳进. 基于信念网络的协同过滤图模型的推荐算法[J]. 模式识别与人工智能, 2016, 29(2):171-176.

[27] 刘付勇, 高贤强, 张著. 基于改进贝叶斯概率模型的推荐算法[J]. 计算机科学, 2017, 44(5):285-289.

[28] 海本斋, 解瑞云. 基于贝叶斯网络的上下文推荐算法[J]. 计算机科学, 2014, 41(7):275-278.

[29] Salakhutdinov R, Mnih A. Probabilistic Matrix Factorization[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2007:1257-1264.

[30] 孙光福, 吴乐, 刘淇,等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013(11):2721-2733.

[31] 杨阳, 向阳, 熊磊. 基于矩阵分解与用户近邻模型的协同过滤推荐算法[J]. 计算机应用, 2012, 32(2):395-398.

[32] Song Y, Zhang L, Giles C L. Automatic tag recommendation algorithms for social recommender systems[J]. ACM Transactions on the Web, 2011, 5(1):1-31.

[33] Hotho A, Jäschke R, Schmitz C, et al. Information retrieval in folksonomies: Search and ranking[C]//European Semantic Web conference. Springer Berlin Heidelberg, 2006: 411-426.

[34] 宋伟伟, 杨德刚, 郑敏. 基于时间加权标签的协同过滤推荐算法研究[J]. 重庆师范大学学报(自然科学版), 2016,3(5):113-120.

[35] 张艳梅, 王璐. 适应用户兴趣变化的社会化标签推荐算法研究[J]. 计算机工程, 2014, 40(11):318-321.

[36] 孔欣欣, 苏本昌, 王宏志,等. 基于标签权重评分的推荐模型及算法研究[J]. 计算机学报, 2017, 40(6):1440-1452.

[37] Hao P, Zhang G, Martinez L, et al. Regularizing Knowledge Transfer in Recommendation With Tag-Inferred Correlation[J]. IEEE Transactions on Cybernetics, 2017, 90(8):1-14.

[38] Ali S M,Ghani I,Latiff M S A.Interaction-based Collaborative Recommendation[J].Ksii Transactions on Internet & Information Systems,2015,9(10):20-35.

[39] Nepal S,Paris C,Pour P A,et al.Interaction Based Content Recommendation in Online Communities[M]//User Modeling,Adaptation,and Personalization. Springer Berlin Heidelberg,2013:14-24.

[40] Kuramoto I,Yasuda A,Minakuchi M,et al.Recommendation System Based on Interaction with Multiple Agents for Users with Vague Intention[M]//Human-Computer Interaction. Interaction Techniques and Environments.Springer Berlin Heidelberg,2011:351-357.

[41] Liu N,Jiang Q,Chen H S,et al.Personalized recommendation using implicit interaction information[C]//International Conference on Computer Science & Education. IEEE, 2011:1340-1345.

[42] Cheng Y, Qiu G, Bu J, et al. Model bloggers' interests based on forgetting mechanism[C]. International Conference on World Wide Web. ACM, 2008: 1129-1130.

[43] 罗军, 朱文奇. 考虑物品相似权重的用户相似度计算方法[J]. 计算机工程与应用, 2015, 51(8):123-127.

[44] 李斌, 张博, 刘学军,等. 基于 Jaccard 相似度和位置行为的协同过滤推荐算法[J]. 计算机科学, 2016, 43(12):200-205.

[45] Liu C, Cao J, He J. Leveraging Kernel Incorporated Matrix Factorization for Smartphone Application Recommendation[C]// International Conference on Database Systems for Advanced Applications. Springer, Cham, 2017:459-474.

[46] Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// ACM Conference on Recommender Systems. ACM, 2010:135-142.

[47] 郭磊, 马军, 陈竹敏,等. 一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报, 2014, 37(1):219-228.

[48] Sobhanam H, Mariappan A K. Addressing cold start problem in recommender systems using association rules and clustering technique[C]// International Conference on Computer Communication and Informatics. IEEE, 2013:1-5.

[49] Le H S. Dealing with the new user cold-start problem in recommender systems: A comparative review[J]. Information Systems, 2016, 58(1): 87-104.

攻读学位期间发表的学术论文

[1] Xie Q, Xiong F, Han T, et al. Interactive resource recommendation algorithm based on tag information[J]. World Wide Web-internet & Web Information Systems, 2018(1):1-19.

[2] Feng Xiong, YongJian Liu, Qing Xie. Recommendations Based on Collaborative Filtering By Tag Weights [C]// International Conference on Semantics, Knowledge and Grids. IEEE, 2017:62-68.