学校代号 10532　　学　　号 G12245165
分 类 号 TP391　　密　　级

湖南大學
HUNAN UNIVERSITY

硕士学位论文

自适应个性化图书推荐算法的研究

学位申请人姓名 刘芷茵
培 养 单 位 信息科学与工程学院
导师姓名及职称 蒋斌副教授、肖文俊高级工程师
学 科 专 业 软件工程
研 究 方 向 电气信息软件方向
论文提交日期 2016 年 5 月 12 日

学校代号：10532

学　　号：G12245165

密　　级：

湖南大学硕士学位论文

自适应个性化图书推荐算法的研究

学位申请人姓名：　刘芷茵

导师姓名及职称：　蒋斌副教授、肖文俊高级工程师

培　养　单　位：　信息科学与工程学院

专　业　名　称：　软件工程

论文提交日期：　2016 年　5 月　12 日

论文答辩日期：　2016 年　5 月　29 日

答辩委员会主席：　廖波

The Research on Adaptive Personalized Recommendation Algorithm of Book

by

LIU ZhiYin

B.E.(Guangzhou University)2003

A thesis submitted in partial satisfaction of the

Requirements for the degree of

Master of Engineering

in

Software Engineering

in the

Graduate school

of

Hunan University

Supervisor

Professor Jiang Bin,Senior Engineer Xiao Wenjun

May,2016

湖 南 大 学

学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。 除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：刘世蓬　　日期：2016 年 12 月 5 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校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。本人授权湖南大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

1．保密□ ，在___年解密后适用本授权书。

2．不保密□。

(请在以上相应方框内打"√")

作者签名：刘世蓬　　日期：2016年 12 月 5 日

导师签名：　　日期：2016年 12月 5 日

摘　要

随着信息技术的高速发展，人们积累的数据量急剧增长，如何从海量的数据中提取有用的知识成为当务之急。大数据背景下的智慧图书馆业务将面临以结构复杂、内容多样的数据为对象、以深度内容挖掘为目的的专业化要求。图书馆不仅需要通过结构化数据了解现在读者需要什么服务，更需要利用大量的非结构化、半结构化数据做多维度的分析，感知读者的真实需求，为读者提供个性化服务。

个性化推荐技术的出现为图书馆的图书推荐服务提供了一个良好的解决思路。本文以自适应个性化图书推荐算法为研究目标，以图书馆的流通业务数据为研究对象，结合协同过滤推荐、上下文感知推荐、用户行为特征分析处理、用户兴趣模型表示等相关技术，研究和分析解决图书馆个性化服务的方法，并通过实验证明了这种图书推荐算法的可行性和有效性。本文的研究工作主要包括以下几个方面：

（1）改进了读者兴趣模型表示方法。针对协同过滤算法“用户—项目”评价矩阵存在的稀疏性问题和可扩展问题，采用《中图法》的图书分类标准，设置图书类别，生成基于图书分类号的“读者—图书类型”评价矩阵，并给出了读者对某类图书偏好值的计算公式，以此构建读者兴趣模型。

（2）利用时间上下文信息，改进了基于用户的协同过滤算法。结合行为特征处理技术，挖掘读者正反馈行为中时间上下文信息与读者兴趣变化的相互关系，按读者借书行为发生时间的先后和图书借期长短对读者借阅图书的行为设置权重，预测目标读者对目标图书的感兴趣程度，提高个性化推荐的精准度。

（3）给出了基于读者负反馈行为的分析处理方法。从图书馆流通业务数据所记录的读者负反馈行为中采集负样本，计算对应的行为特征权重，进一步完善读者兴趣模型，使读者兴趣模型反映出读者最真实的需求，提高相似读者计算的准确度，提高图书推荐的效果。

（4）提出了面向公共图书馆的自适应个性化图书推荐算法，设计了五组对比实验，验证该算法的可行性和有效性。

关键词：自适应个性化推荐；图书推荐；协同过滤；上下文感知；行为特征分析

Abstract

With the rapid development of information technology, the amount of data accumulated in the rapid growth of people, how to extract useful knowledge become a top priority from the mass of data. Library Business background under the big data will be faced with complex, diverse content data as an object to dig deep content for the purpose of the specialized requirements. Libraries need not only structured data reader now understand what services, but need to use large amounts of unstructured, semi-structured data to make multi-dimensional analysis, the reader's perception of the real needs, to provide readers with personalized service.

Appear personalized recommendation technology for library books recommended service provides a good solution ideas. In this paper, an adaptive personalized book recommendation algorithm research objectives, the flow of business data to the library for the study, combined with collaborative filtering, context-aware recommendation, user behavior analysis processing characteristics and other related technologies, research and analysis solution library individuation the method of service, and experiments show the feasibility and effectiveness of the book recommendation algorithm. Research work of this paper includes the following aspects:

(1) Improve readers' interested model presenting method. "ZhongTuFa" books classification criteria for classifying books, built on words of the book "reader - Books types" access matrix can be an effective solution based library resources, "the reader - Books' access matrix due READERS' BORROWING with respect to the number of library books too few, and many readers did not produce the same behavior borrow books from each other, due to the sparsity of the matrix is too large.

(2) bonding time contextual information, the relationship between mining Library Circulation time information and business data reader interest change, according to the reader has time library behavior and by the length of the reader to borrow books acts set different weights, can make readers interested in the model to reflect the real needs of readers, book recommendations to improve the effect.

(3) Establish analytical method based on readers' negative feedback behavior. combined

with business data library circulation records Readers negative feedback acts readers behavioral characteristics process, negative feedback on negative samples collected behavior, improve the reader interested in the model, which identify the target audience of "nearest neighbors", predicting the target level of interest on the target reader of books, improving personalized recommendation accuracy.

(4) Create a self-adjusted personalized book recommended method to the public library. To test the plausibility and effectiveness of this method, five controlled experiments were designed and carried out.

Key Words： adaptive personalized recommendations; books recommended; collaborative filtering; context-aware; behavioral characteristics analysis;

目 录

插图索引

附表索引

第 1 章 绪论

1.1 课题来源与意义

20 世纪，信息技术的出现，为图书馆的发展带来了新的生机与活力。60 年代，MARC 的出现使得图书管理的电子化和网络化成为可能；基于图书馆采访、编目、验收、流通等日常业务的信息化管理系统使得传统的手工业务逐步实现了自动化；计算机应用从单机向局域网、互联网的发展，使得图书馆信息化从局部走向整体，进而走向整个行业。信息技术的发展使得图书馆的业务服务有了质的飞跃。

随着互联网技术在中国的飞速发展，1998 年，中国正式进入数字图书馆建设时代。数字图书馆的特征是信息化、网络化、数字化，把信息以数字化形式加以储存，通过互联网传输，从而做到信息资源共享。近年，随着国内“智慧城市”、“互联网+”等理念的提出，数字图书馆也逐步向智慧图书馆升级转型。智慧图书馆是传统图书馆与移动互联网、云计算、物联网、大数据等新一代信息技术融合的产物，是广泛互联和融合共享的图书馆。

大数据又称海量资料，是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的海量数据进行专业化处理，从而挖掘出其中的深层价值。大数据已经成为国家科技创新竞争的主战场。联合国“数据脉动”计划、美国“大数据”战略、英国“数据权”运动、日本“面向 2020 年的 ICT 综合战略”、韩国大数据中心战略等先后开启了大数据创新战略的大幕。2015 年 9 月 5 日，国务院印发《促进大数据发展行动纲要》，系统部署我国大数据发展工作[1]。对于很多行业而言，如何利用大数据已经成为赢得竞争的关键。对于图书馆而言，只有借助大数据的智慧，在海量增长的文献数据流中发挥长效的处理能力，搜寻新的数据运算、知识发现的新途径，探索新的服务模式，才能使图书馆在数字化阅读的新时代保有一席之地，才能确保图书馆不被高科技“边缘化”。

图书馆拥有大量藏书，随着国家对文化事业支持力度的加大，图书信息资源将进一步增长，读者要在浩瀚的书海中找到所需图书着实不易。在图书馆，文献流通存在着一种“二八现象”。 图书馆的活跃读者大概占图书馆读者比例的 20%，图书馆大量的馆藏图书中借阅率高的图书也往往集中在图书总量的 20%里面。意大利经

济学家巴莱多发现二八定律认为，在任何一组东西中，最重要的只占其中一小部分，20%，其余 80%尽管是多数，却是次要的。还有，图书馆还存在着一种“长尾现象”。如果按照每本图书的借阅量画出图书的分布曲线，可近似地得到一条递减曲线。在曲线的头部，热门图书被大量外借，随着流行程度的降低，曲线徒然下降。但有趣的是，在尾部曲线并没有迅速坠落到零，而是极其缓慢地贴近于横轴。如果按照每位读者借阅图书总量画出读者的分布曲线，也是满足长尾分布的。

Fielding Graduate University 于2006年通过对1656名用户在2005年7月到2006年6月之间下载的67060条电子文献资源进行分析，表明用户在资源使用上呈现长尾现象[2]。

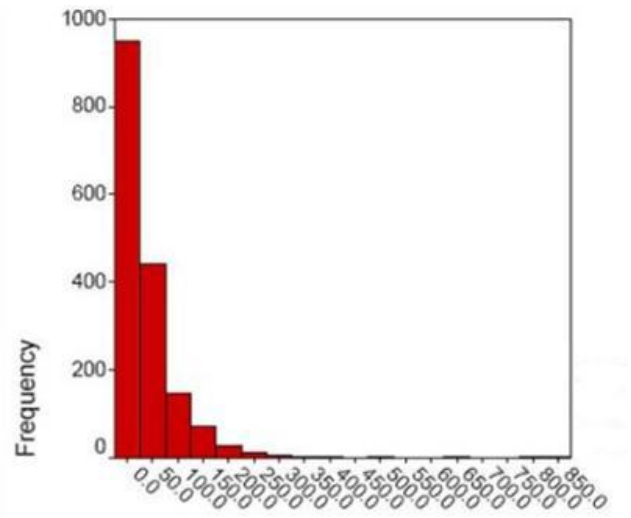

图 1.1 资源使用的长尾现象

为了缓解上述情况，智慧图书馆作为数字图书馆的更高阶段，有必要充分利用现代化信息技术手段，为读者开展智慧服务——图书个性化推荐服务。借助推荐服务，一方面可以帮助图书馆将长尾图书推荐给读者，提高馆藏资源的借阅量和利用率，另一方面也可以满足读者的借阅需求，真正实现“以人为本”。

个性化推荐是根据系统中用户的不同需求，为之提供有针对性的信息推荐服务。推荐的原理是收集用户信息和数据，分析用户的行为记录，了解用户的借阅习惯，抽取出行为特征，构建用户模型，结合推荐算法给出推荐结果。近年来，协同过滤推荐、基于内容的推荐、社会化推荐、上下文感知推荐等算法的涌现推动了个性化推荐的发展，个性化推荐的应用已渗透到各行各业。结合大数据技术，以读者为中心，开展个性化推荐服务，助力图书馆行业从传统图书馆向智慧图书馆升级转型，具有重要的现实意义。

1.2 国内外研究情况

1.2.1 图书推荐系统

在国外，图书推荐系统已经发展得比较好，资源定制模块，个人图书管理模块，最新公告模块，自选推荐（进行自选期刊最新篇目的推送）模块逐步完善，推荐系统较为可靠和高效。国外主要的图书推荐系统有：

（1）Tapestry 系统是早期的推荐系统，在 1992 年由美国施乐公司开发，开发者 Goldberg 等人在开发过程中，首次提出了“协同过滤”的概念，这对之后基于“协同过滤”思想的图书推荐系统的应用具有启发意义[3]。

（2）Fab 系统是斯坦福大学在参与数字图书馆项目时研发的成果，目的是从大量的网络信息中筛选用户感兴趣的内容。系统涉及了协同过滤推荐和基于内容的推荐两种思想，首先从互联网采集信息，将信息与用户需求匹配，推荐给用户匹配度高的信息。用户还可以对推荐信息进行评级，当其他兴趣相近的用户检索信息时，系统会自动推荐评级高的信息[4]。

（3）Citeseer 是 NEC 研究院的一个学术论文数字图书馆。系统是基于自动引文索引(Autonomous Citation Indexing, ACI)机制开发的，检索互联网上 Postscript 和 PDF 文件格式的学术论文[5]。

（4）Melvyl 系统是加州大学伯克利分校 2006 年图书馆项目的研发成果，开发者探索了基于借阅数据的推荐和基于内容的推荐两种推荐方法[6]。

（5）Amazon 系统是亚马逊网络书店采用的推荐系统，它采用了基于项目的协同过滤推荐算法，多年来对算法不断改进完善，是目前运用最成功的推荐系统之一[7]。

我国对图书推荐技术的研究起步相对较晚，笔者在 CNKI 期刊论文库检索，2000 年以后才有国内专家学者对个性化推荐研究的论文发表。

2000 年，我国第一个图书馆个性化定制系统，在北京大学顺利开展，由社会科学基金资助，可以进行个性化的图书推荐，名为“基于 web 的图书馆定制服务系统”。与此同时，中国科学院国家科学数字图书馆“我的数字图书馆”开通，深图朗恩技术有限公司也研究开发具有图书馆个性化信息服务功能的 ILAS 系统。浙江大学、清华大学和中国人民大学分别于 2001 年、2002 年和 2003 年开始研究图书馆个性化推荐系统，其中比较有代表性的是中国人民大学的“数字图书馆个性化信息服务系

统”[8]。

2000 年成立的“互动出版”是我国较早应用个性化图书推荐系统的一个网上书店。先后于 2004 年、2005 年成立的当当网、豆瓣网是我国较著名的图书推荐系统。当当网使用的个性化推荐算法主要是基于内容的推荐、协同过滤推荐、关联规则等方法。豆瓣网使用的主要是基于图书标签的推荐和协同过滤推荐算法[9]。

1.2.2 推荐算法研究进展

在图书、新闻、音乐、电影、电子商务等众多行业里面，推荐系统均具有重要的应用价值，当前已经成为数据挖掘和人工智能研究领域最热门、最具挑战性的研究课题之一。2006 年 10 月，美国 Netflix 公司宣布了一项竞赛，旨在发现更好的推荐算法解决电影评分预测问题，并准备了一百万美元的大奖。这场历时三年的竞赛吸引了来自 186 个国家的 4 万多个团队参加，推动了推荐算法的创新和发展[10]。近年，随着人工智能、数据挖掘、机器学习等技术的发展和进步，国内外研究推荐算法的专家学者越来越多。

推荐算法大都是涉及用户历史行为的，以用户行为作为推荐依据。最早的此种推荐算法是于 1992 年被提出的基于用户的协同过滤算法，当时被应用于邮件过滤系统，1994 年被应用于新闻过滤[11]。1998 年，Jhon S.Breese 等人提出通过降低用户兴趣列表中热门物品的特征权重，更准确地挖掘出相似用户或相似物品的 USER-IIF 算法[12]。另一种经典的算法是 2001 年 Sarwar 提出的基于物品的协同过滤推荐算法[13]。在国内，已有不少研究者对协同过滤算法加以改进，孙光福等提出了基于时序行为的协同过滤推荐算法[14]，张新香等提出了利用云模型改进的协同过滤推荐算法[15]，肖敏等提出了基于项目语义相似度的协同过滤推荐算法[16]。推荐算法的目的是联系用户的兴趣和物品，基于标签的推荐算法就是通过用户对物品打标签的行为建立起用户和目标物品的联系，从而发现用户的兴趣并进行推荐的一种算法[17]。用户打标签的行为由<用户，物品，标签>的三元组表示，对于每位用户，找到他最常用的标签，再从所有物品中找到具有这些标签的物品推荐给用户[18]。Delicious、CiteULike、Last.fm、Hulu、豆瓣，这些都是著名的引入了用户标签系统的网站。2008 年，ECML/PKDD 推出了基于标签的推荐系统比赛[19]，涌现了张量降维[20]、基于 LDA 的算法[21]、基于图的算法[22]等优秀的算法。

以上提到的算法都存在冷启动问题，包括用户的冷启动和物品的冷启动。解决

用户冷启动的传统方法就是采用基于人口统计学信息的推荐算法。由于推荐系统中新用户没有历史行为记录，而人口统计学特征包括年龄、性别、国籍、民族、居住地、学历、工作等信息，这些特征对预测用户兴趣能够起到很重要的作用，比如不同年龄的人兴趣不同，不同性别的人兴趣也不同。那么将用户分类，就可利用同一类中其他用户的兴趣进行推荐[23，24]。但是这种推荐算法存在两个问题：一是推荐粒度太粗[25]，二是用户基于保护私隐的考虑往往不会填写真实的信息[26]。针对物品冷启动，则可以分析物品的内容，然后通过计算用户与物品的相关程度进行推荐，即基于内容的推荐算法[27，28]。段准对内容推荐中初始模板的构建以及用户模板的更新进行分析和研究，提出了一种应用于文本推荐的基于内容的自适应推荐算法[29]。卜起荣利用计算机提取图像的视觉内容提出了基于内容的图像检索与推荐算法[30]。姜书浩等提出了一种利用协同过滤预测和模糊相似性改进的基于内容的推荐方法[31]。另一种解决冷启动的方法是社会化推荐，即利用社交网络数据的推荐。随着推特、脸书等社交网站的兴起，产生了大量的社交网络数据，凭借用户对朋友的信任关系，可利用朋友的历史行为信息进行推荐。社会化推荐主要有基于邻域的推荐、基于图的推荐、信息流推荐、基于位置的推荐等几种推荐算法[32，33]。ACM 推荐系统年会自 2009 年[34]开始举办涉及社会化推荐系统的专题讨论会，在 2011 年[35]和 2013 年[36]的专题研讨会上均指出了社会化推荐系统领域的几个研究主题和发展方向。

为了提高推荐系统的推荐效果，一种通过处理已感知到的上下文信息，将上下文信息融入推荐系统，为用户进行推荐的上下文感知推荐算法被提出。Adomavicius 和 Tuzhilin 等人提出了利用多维效用模型表示上下文信息的方法[37]。在目前的研究当中，主要有针对时间信息[38]、地理位置信息[39，40]和用户心情[41]等上下文信息的推荐。ACM 推荐系统年会自 2009 年[42]开始举办上下文感知推荐系统的专题讨论会，指出上下文感知推荐系统领域的几个主题，体现了当前的研究热点与难点。近几年，国内关于上下文感知推荐的研究也逐年增多，刘颖提出了一种基于二重分解的上下文预过滤推荐技术[43]，秦大路提出了基于因式分解机模型的上下文感知推荐[44]，朱煦等提出了一种利用移动设备信息的上下文感知好友推荐方法[45]。

1.2.3 智慧图书馆个性化推荐研究进展

智慧图书馆概念率先在一些欧美国家图书馆界提出，并在公共图书馆和大学图书馆中实践。2003 年前后，芬兰奥卢大学图书馆提供的一项新服务称为“Smart L

ibrary"，这一服务隶属于“Rotuaari Project”项目。此后图书馆的学者发表了题为《智慧图书馆：基于位置感知的移动图书馆服务》的会议论文，指出“Smart Library”是一个不受空间限制的、可被感知的移动图书馆服务，它可以帮助用户找到所需图书和相关资料。王世伟给出了相对完整的智慧图书馆概念：智慧图书馆是以数字化、网络化、智能化的信息技术为基础，以互联、高效、便利为主要特征，以绿色发展和数字惠民为本质追求，是现代图书馆科学发展的理念与实践[46]。

从传统图书馆到数字图书馆再到更高级的智慧图书馆，不管技术如何发展，图书馆本质理念还是“以人为本”、“以读者为中心”，所以智慧图书馆必须提供个性化的知识服务，并且要由传统的“被动知识服务”转变为“主动知识服务”。

陈卓辉介绍了从广泛使用数字和网络技术、智慧化管理和智慧化服务、服务的网络形式或移动形式等不同角度对当代图书馆的特征及发展前景的研究，指出目前国内外关于智慧图书馆个性化推荐的研究还相对较少[47]。曾子明等研究了智慧图书馆利用物联网、云计算等技术提供个性化信息服务的可能性，提出了融合情境感知的智慧图书馆个性化服务模型，分析了该服务模式的构建方法及存在问题[48]。肖理钏在充分分析高校数据孤立、业务流程低效、交互性差等实际问题的基础上，给出基于云计算和大数据等技术来构建统一、规范的图书个性化推荐与服务的统一平台[49]。马晓亭为了解决智慧图书馆个性化服务中存在的问题，提出了一种基于大数据的图书馆个性化智慧服务质量保证策略[50]。郭素君提出了一种智慧图书馆信息服务系统的解决方案，通过个性化服务推荐出适合的图书给读者[51]。

1.3 论文主要研究内容

本文研究的对象是自适应个性化图书推荐算法，以目标用户为中心，基于读者兴趣评价矩阵，采用协同过滤的思想，通过收集用户的行为信息，分析用户行为特征，感知用户需求的变化，适时调整用户兴趣模型，为目标用户提供其最感兴趣的图书信息，进行自适应个性化推荐。

具体研究内容如下：

（1）查阅和收集关于个性化推荐算法、推荐系统，以及智慧图书馆个性化推荐的相关文献资料，参考和借鉴国内外相关研究成果。

（2）分析研究图书馆的日常业务，站在图书馆和读者双方的角度分析图书馆业务和读者服务的特点。

（3）研究现有系统中各种推荐算法和技术思路，比较各种推荐算法在图书馆图书推荐应用上的适用性，选择出比较合适的基于用户的协同过滤算法作为图书推荐的基础算法。

（4）引入《中图法》的图书分类法用以改进读者兴趣模型表示方法，旨在研究如何改善用户评分矩阵的稀疏性，解决用户评分矩阵的可扩展问题。

（5）研究用户行为特征分析和处理方法，针对图书馆读者的正反馈行为、负反馈行为，研究用户行为产生的时间、用户行为持续时间，与用户兴趣偏移的关系，挖掘与分析用户行为的隐性特征，计算出用户特征权重，提出自适应个性化图书推荐算法。

（6）通过离线实验，用平均绝对偏差、准确率、召回率等评价指标证明本文研究的自适应个性化图书推荐算法的可行性和有效性。

1.4 论文组织结构

本文组织结构如下：

第一章，介绍了本文的课题背景和研究意义，并对当前国内外推荐系统现状、个性化推荐算法的相关研究和智慧图书馆个性化推荐研究进展进行了综述，最后介绍本文关于自适应个性化图书推荐算法的研究内容，及整篇论文的组织结构。

第二章，首先描述了协同过滤推荐、基于内容的推荐、基于标签的推荐、社会化推荐等推荐算法的核心思想和基本原理，并对算法的优缺点进行了说明，描述了上下文感知推荐算法的思想、定义和算法分类；然后从用户行为数据和用户行为特征两方面对用户行为特征处理技术加以说明；最后介绍了几种常用的用户兴趣模型表示方法，为后面的章节奠定了理论基础。

第三章，对图书馆的流通外借业务进行了梳理，分析图书馆日常业务和读者服务的特点。结合图书馆的实际情况，从算法的优缺点、适应度等方面比较了协同过滤推荐、基于内容的推荐、基于标签的推荐、社会化推荐等几种主要的推荐算法，最终选定基于用户的协同过滤推荐算法作为图书馆个性化推荐的基础算法。分析了该算法在图书馆的图书推荐中存在的不足，同时加以改进。首先，采用《中图法》的图书分类思想对馆藏图书进行细分，提出了基于图书分类号的“读者—图书类型”评价矩阵，给出了读者对某类图书偏好值的计算公式，改进了读者兴趣模型。还在推荐算法中融入了行为特征处理技术和时间上下文感知的方法。最后通过对图书馆

流通数据中用户的负反馈行为特征的分析进一步完善用户兴趣模型，最终提出面向公共图书馆的自适应个性化图书推荐算法。

第四章，针对第三章提出的自适应个性化图书推荐算法设计实验，选定实验数据，通过五组对比实验测量本文提出的几种改进方法的推荐效果。

最后，总结论文工作，给出了实验的结论与本文研究的创新点，并对下一步工作予以展望。

第 2 章 相关理论研究

本章主要介绍与本文研究相关的推荐算法基础理论，分别是协同过滤推荐、基于内容的推荐、基于标签的推荐、社会化推荐、上下文感知推荐这几种推荐算法，还介绍了用户行为特征处理技术和用户兴趣模型表示方式。

2.1 个性化推荐

2.1.1 推荐系统概述

在全球互联网普及率不断增长的环境下，数字化信息爆炸性增长，人们普遍面临着信息过载的问题，对于互联网用户，难以从海量的信息中发现自己真实需要的信息。对于信息发布者，则需要解决如何将信息定向发布，将信息提供给真正有需求的用户的问题。个性化推荐系统一定程度上可以解决以上问题。推荐系统不需要用户提供明确的需求，只需要收集用户的历史行为信息，建立用户兴趣模型，将用户和物品关联起来，就可以为用户提供个性化推荐。个性化推荐系统由三部分组成：收集用户历史行为信息的采集模块，分析用户兴趣的模型分析模块，推荐算法模块[27]。其中，推荐算法是最核心的部分。

2.1.2 协同过滤推荐算法

俗话说“物以类聚、人以群分”，在同一个应用系统中，总会存在相同类型的物品，和喜好相近的用户。为用户进行个性化推荐时，利用大量的用户行为数据，从中寻找相似物品或者相似用户，从而为用户推荐有效物品即为协同过滤推荐。

1. 基于用户的协同过滤

基于用户的协同过滤算法的基本思想为：首先收集用户历史行为信息，并以此来计算用户之间的相似度，并以此为依据找到与当前目标用户相似程度最高的 K 个用户集合，然后再收集这个集合中所有用户对其他项目的评分，通过对不同项目评分的高低来推测出目标用户对其他产品的喜好程度，从而实现推荐应用。算法的基本步骤为[52]：

（1）建立用户评分矩阵。设 n 为用户数，m 为项目数，用 n×m 的评分矩阵来表示第 i 个用户对第 j 个项目的评分值。

（2）生成 K 个最近邻居。通过计算所有用户之间的相似度形成最近邻居集合。

（3）推荐。通过加权目标用户的最近邻居对目标项目的评价，产生最终的推荐结果。

图 2.1 说明了基于用户的协同过滤推荐算法的基本原理。

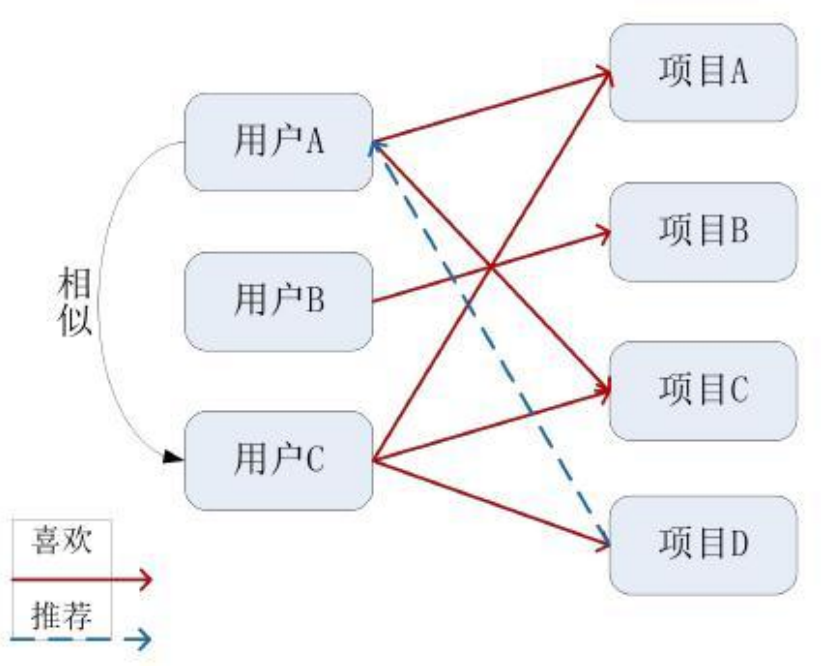

图 2.1 基于用户的协同过滤基本原理

2. 基于物品的协同过滤

基于物品的协同过滤推荐算法，就是为用户推荐那些和他们之前喜欢的物品相似的物品。首先要针对目标用户对项目的喜好程度来找到与之相似的项目集合，将与目标项目相似度高的项目作为目标项目的最近邻居，然后根据用户的历史喜好和兴趣对集合中的项目进行排序，将相似度靠前的项目推荐给目标用户。其算法的基本步骤为[52]：

（1）计算项目之间的相似性。

（2）生成 N 个最近邻居。将与目标项目相似度高的，且目标用户对其没有给出过评价的前 N 个项目作为目标项目的最近邻居集合。

（3）推荐。对项目邻居集合中所有项目的评分进行加权求和，从而得到目标用户对所有项目的预测评分。

图 2.2 说明了基于物品的协同过滤推荐算法的基本原理。

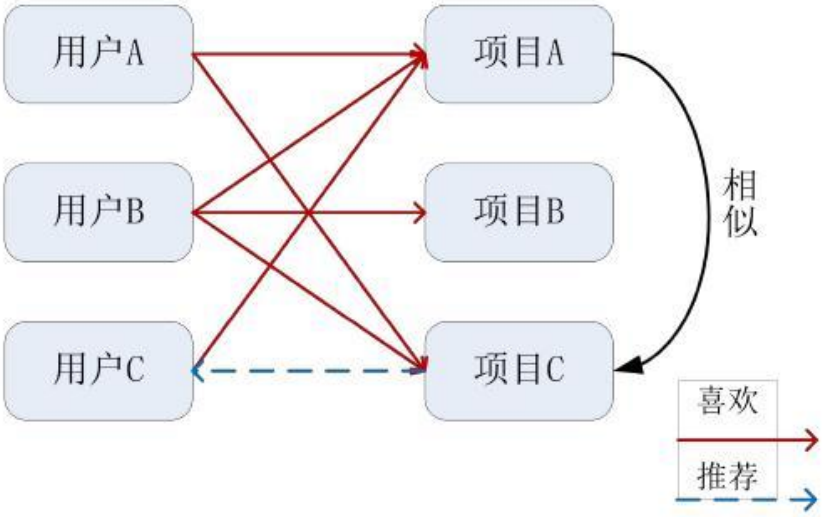

图 2.2 基于物品的协同过滤基本原理

3. 算法的优缺点

协同过滤推荐具有以下优点：

（1）能够对非结构化对象如音乐、电影、艺术品和难以表达的概念如信息品质、个人品位进行过滤。

（2）能够有效共享相似用户的反馈信息，借鉴他人的经验和意见，加快个性化学习速度。

（3）推荐的新颖性。不需要专门的领域知识，具有新异兴趣发现的能力，能够发现在内容上完全不相似的信息，能够发现用户不明显的、潜在的，而用户自己尚未发现的兴趣爱好。

同时，该算法也存在着以下的缺点：

（1）稀疏性问题。在实际应用中，用户对项目的评分信息非常少，基于用户的评分得到的用户评分矩阵，所计算出的用户相似度的准确性会受到影响。

（2）可扩展问题。即随着用户数量和项目数量的增加，在搜索目标用户的最近邻居时，计算量较大而耗时，系统性能会越来越低。

（3）冷启动问题。由于新用户在系统中没有产生过任何的行为，没有任何的评分信息，无法为其找到合适的最近邻居并做出推荐；新物品信息录入系统后，由于没有任何用户对其产生过行为，做出过评价，导致这个新物品无法被推荐。

2.1.3 基于内容的推荐算法

1. 基于内容的推荐算法概述

协同过滤推荐只考虑了用户评分数据，忽略了用户和物品本身的诸多特征，基于内容的推荐算法的根本在于内容的获取与分析，通过用户与物品的相关性进行推荐。基于内容的推荐一般包括以下三个步骤[53]：

（1）物品特征描述

物品特征描述即分析物品内容属性，为每个物品构建一个物品的属性文档。在一个应用系统里，每个物品都会有一些可以描述它特征的属性。这些属性通常分为两种：以字段形式存储在数据库中，具有明确的取值范围和意义的结构化数据；结构比较松散的，以文字、评论等形式描述物品的非结构化数据。对于结构化数据，分析处理比较容易。而非结构化数据不具备属性字段，往往含有较多自然语言，通常使用 TF-IDF 处理。

（2）用户兴趣描述

用户兴趣描述即为每个用户构建一个用户的喜好文档。在系统中，记录了大量的用户历史行为数据，包括检索、浏览、收藏、购买、评论等等，在这些数据中可以判断出用户的喜好。根据用户访问历史数据以及用户喜欢的相关物品的属性文档，可以聚合建立描述用户兴趣的用户喜好文档，即构建用户模型。

（3）物品的推荐

最后按照用户喜好文档和物品属性文档的相关程度进行推荐，可以通过计算用户没有产生过行为的物品相应的属性文档和用户的喜好文档的相似性，把相似性最大的物品推荐给用户。

2. 算法的优缺点

基于内容的推荐算法的优点是不需要其他用户的数据，也不需要物品的使用数据，不存在冷启动问题；能够为具有特殊兴趣爱好的用户提供推荐；可以根据用户的过往行为以及推荐物品的内容特征解释推荐结果，容易得到用户的信任。

该算法的缺点是对于内容特征的抽取要求较高，需要内容具有较好的结构性，对于复杂的属性难以处理。

2.1.4 上下文感知推荐算法

1. 上下文感知推荐算法概述

1991 年，施乐公司的 CTO Mark Weiser 提出了“适普计算”的概念[54]。作为其核心子领域之一的上下文感知计算理论，使系统能够自动发现和利用位置、周围环境等上下文信息，并为用户提供服务和计算资源。在普适计算的环境中，人和计算机之间不断地交互，在交互过程中，普适系统获取与用户需求相关的上下文信息来确认为用户提供什么样的服务，这就是上下文感知。所涉及的主要问题包括：获取上下文信息、上下文信息融合和处理上下文信息。

Adomavicius 等人提出了“上下文感知推荐系统（CARS，Context-aware Recommender Systems）”概念，将传统的“用户—项目”评分模型扩展为包含多种上下文信息的多个维度的评分模型[55]。

“用户—项目”评分效用模型：

$$u:U\times I\rightarrow R \tag{2.1}$$

R 为用户 U 和项目 I 映射的效用得分。

扩展为多维评分效用模型：

$$u:D1\ \times\ \cdots\times Dn\rightarrow R \tag{2.2}$$

Di 可以是用户、项目，或者多维上下文信息中的一维。

2. 上下文定义

对于上下文的定义，Dey 等人提出了：“上下文可以被认为是环境本身或者环境中可以用于描述某目标实体状态的任何信息[56]。其中，实体一般可以是人、物、地点等物理实体或者是用于与用户和应用程序之间信息交互相关的客体。”

Schilit 将上下文信息定义为：发生购买行为的地点，周围的人物，以及发生行为附近所拥有的资源[57]。利用地点、人物、资源信息可以在移动应用中带来更好的客户体验。

一般情况下，常用的上下文实例有：时间、地点、天气、社会关系、情绪、设备状态、身体状况等。

3. 上下文感知推荐算法

目前的应用当中，上下文感知推荐算法主要是跟传统的推荐算法结合使用。主要有基于协同过滤的上下文推荐、基于内容的上下文推荐、基于知识的上下文推荐三类。

（1）基于协同过滤的上下文推荐算法

基于协同过滤的上下文推荐算法是在计算用户之间的相似性的时候，考虑用户

的上下文信息，计算用户所处上下文的相似性，对用户进行推荐。该算法扩展了传统的协同过滤推荐算法，认为处于相似上下文情景下的用户拥有更高的相似性。该算法的优点是能够发现用户潜在的兴趣偏好，同时也无需对用户、物品、上下文等构建模型和分类。缺点是存在新用户问题和用户对物品评价的稀疏性问题。

（2）基于内容的上下文推荐算法

基于内容的上下文推荐算法是将上下文信息的利用与基于内容的推荐相结合，根据用户在不同的上下文信息中的历史行为构造文档资料，与物品的属性文档资料相匹配，计算用户与物品之间的相关性，向用户推荐符合用户上下文情景的物品信息。物品特征属性的描述一般需要使用 TF-IDF 来进行处理。构建用户喜好文档需要用到分类算法，诸如决策树、最近邻、线性分类、朴素贝叶斯算法等等。基于内容的上下文推荐算法对于用户的特征描述的准确度有一定的提升，但是同样存在新用户问题。

（3）基于知识的上下文推荐算法

以上两种算法都存在着冷启动问题，而且也需要用到用户的历史行为数据，但事实上，实际应用中往往用户不一定会重复对某件物品或某类物品重复产生行为，基于知识的上下文推荐则可以解决这些问题。该算法以用户需求与物品之间的相似度，或者根据明确的推荐规则进行推荐，而不依赖于用户评分信息。基于知识的推荐通常有两种基本类型：基于实例推荐和基于约束推荐。在知识推荐系统中，可以利用上下文和一些实例，为用户推荐；也可以挖掘出在某些上下文约束下，用户对物品的兴趣，生成推荐列表。

2.1.5 基于标签的推荐算法

1. 基于标签推荐算法概述

基于标签的推荐是指通过用户给物品打标签的行为建立起用户和标签的联系，挖掘用户感兴趣的物品作出推荐。标签可以是由系统限定的，也可以是由用户自己定义的。用户使用自然语言标注物品，此类标签称之为社会化标签，社会化标签数量较多，且容易出现同义词。系统给定的标签数量相对较少，但标签的质量相对较高[58]，往往更精准。用户为物品打标签的行为可以用一个三元组（u,i,b）表示，u 代表用户，i 代表物品，b 代表标签，（u,i,b）即用户 u 为物品 i 打上标签 b[59]。

基于标签的推荐可以通过如下步骤实现：

（1）统计每个用户最常用标签；

（2）对于每个标签，统计被打过这个标签次数最多的物品；

（3）对于一个用户，找到他常用的标签，从而找到具有这些标签的物品进行推荐。

一个用户可以为一个物品打多个标签，也可以为多个物品打同一个标签，当用户使用某个标签相对频繁的时候，可以理解为用户对该标签对应的物品感兴趣。当两位用户使用同一标签标注物品的时候，可以理解为这两位用户拥有相似的兴趣。

2. 算法的优缺点

基于标签的推荐算法，优点是能够自动将用户和物品关联起来，帮助用户发现感兴趣的物品。但也存在着一些问题。一是用户的主观意识导致不同用户对标签的理解不同，往往容易出现同一物品被标注多个同义标签，甚至被标注错误标签的情况。二是标签也会有歧义，容易造成同一标签被用于标注不同的物品。这些问题会造成系统中标签使用混乱的情况，影响推荐结果的质量。

2.1.6 社会化推荐算法

1. 社会化推荐算法概述

社会化推荐是通过信息过滤技术，利用社会网络的社交行为数据，挖掘出集体智慧[60]。社会化推荐在于对人与人之间、群体之间、组织之间的关系进行描述，将个体之间的关系视为主要考虑因素[61]。这种推荐的方式不同于传统的信息推荐，它将社会网络、社交媒体视为推荐的主要平台，使用户的隐性知识在社会化推荐过程中与其他用户相互交流，在缓解传统推荐系统中数据稀疏性及冷启动问题的同时，还提高了推荐的性能。社会化推荐的形式有四种：基于社会化标引的推荐、基于社会化评价的推荐、基于用户关系的推荐、综合推荐。

2. 算法的优缺点

社会化推荐的优点主要体现在两方面[32]：

（1）缓解数据稀疏性问题。传统的推荐算法普遍存在着“用户—物品”评价矩阵稀疏问题，影响用户相似性的计算。社会化推荐使用到用户的社交网络数据，在计算用户相似性时可以把社交网络的朋友信息作为相似用户，从而缓解数据稀疏性。

（2）缓解冷启动问题。由于新用户没有历史记录信息，传统的推荐算法很难为新用户做推荐。而社会化推荐可以利用社交网络数据，为用户建立用户特征模型，

为用户推荐物品。对于新物品传统的推荐算法由于缺乏物品使用信息，同样存在冷启动问题，而社会化推荐可以依靠社交网络中的其他用户对物品的使用信息为依据进行推荐。

社会化推荐的缺点主要体现在以下两方面[32]：

（1）隐私保护和安全性问题。由于社会化推荐主要依赖于用户社交网络中的关系用户信息，涉及到用户个人信息的提取与利用，这些信息都牵涉到用户的隐私，容易引发用户信息安全问题。也由于此原因，很多社交网络系统不轻易提供用户信息，导致社会化推荐难以开展。

（2）推荐的准确性问题。社会化推荐的思想是通过用户信任的关系密切的其他朋友使用其他物品的历史行为记录为用户推荐，增加推荐结果的信任度和接受度。但是用户好友的兴趣爱好未必与用户一致，有可能推荐的物品并不是用户感兴趣的，从而影响到推荐的准确性。

2.2 用户行为特征处理

用户行为特征处理，是指在获得用户在应用系统中的所有历史操作行为的数据信息后，对有关数据进行统计、分析，从中发现用户使用系统的规律，让服务提供方更加详细、准确地了解用户的行为习惯，从而提供有针对性的服务，提高业务转化率。

2.2.1 用户行为数据

很多互联网应用系统都会用操作日志来记录用户行为数据，一条日志记录表示用户的一次行为和对应的服务。例如在图书馆的业务系统中，读者每借一本书就会生成一条借书日志，记录了读者姓名、所借图书书名、借书时间等行为数据。日志通常存储在分布式数据仓库中，可以使用 hadoop、Dremel 等进行数据分析。这些日志记录了用户的各种行为，在电子商务网站中这些行为主要包括网页浏览、检索、点击、收藏、购买、评价等。

用户行为在个性化推荐系统中，一般分为显性反馈行为和隐性反馈行为两种。显性反馈行为需要用户参与，可以明确表达用户对物品的喜好，如评论、收藏以及用户主动提供个人兴趣等。YouTube 就是一个分析用户显性行为，继而进行推荐的网站。它采用“喜欢/不喜欢”两档评分系统，收集用户兴趣。用户对视频作出评价

的同时，就向网站提交了个人对于该视频的喜好。

隐性反馈行为是相对于显性反馈行为而言的，不需要用户直接参与，但系统中往往含有大量的用户行为，虽然不能明确表达用户的兴趣爱好，但行为种类众多且数据量非常大。用户浏览网页、检索查询、点击链接、分享、购买等行为都属于隐性反馈行为。行为心理学认为人的行为可以反映人的兴趣。相关研究表明，用户在浏览网页时诸如鼠标悬停时间、移动鼠标、点击鼠标等行为虽不能直接表示用户的兴趣，但这些行为持续发生且集中于同一类对象时，可反映用户的兴趣。

表 2.1 列出了显性反馈数据和隐性反馈数据在用户兴趣、数量、存储方式等几个方面的对比[62]。

表 2.1 显性反馈数据和隐性反馈数据的比较

比较项目	显性反馈数据	隐性反馈数据
用户兴趣	明确	不明确
数量	较少	庞大
存储	数据库	分布式文件系统
实时读取	实时	有延迟

2.2.2 用户行为特征

用户的行为特征是从日志中记录的用户行为中计算出来的。一个特征向量由特征以及特征的权重组成，在处理用户行为特征时需要考虑以下因素[62]。

（1）用户行为的种类

在日志文件中，包含了多种操作类型。也就是说用户在系统应用过程中会有很多不同种类的行为，用户可以检索物品、点击物品链接、浏览物品信息、给物品打标签、将物品添加进购物车、购买物品、给物品打分等。用户的这些行为都会对特征权重产生影响，一般认为用户付出代价越大的行为权重越高。比如，购买物品需要花钱，购买行为对应的特征权重应该最高。相反，点击物品链接、浏览物品信息等行为仅是动动鼠标而已，这些行为对象对应的特征权重相比购物行为会小很多。

（2）用户行为产生的时间

一般来说，用户近期的行为反映了用户近期的需求，相比用户很久之前的行为显得更重要。因此，如果用户最近购买过某个物品，那么这个物品对应的特征权重

会相对较高。

（3）用户行为的次数

有时用户会重复产生多次行为，比如多次采购同一类型的产品，甚至多次采购同一产品，或者重复听同一首歌，多次检索同一物品。这种重复的行为都说明了用户的兴趣倾向，因此对于行为次数多的物品对应的特征权重应比行为次数少的物品高。

（4）物品的热门程度

如果用户对一个很热门的物品产生了行为，往往不能代表用户的个性化需求，有可能是因为这个物品被放在了首页显眼的位置，而被用户点击到而已。相反，如果用户对某一冷门的物品产生了行为，就很可能说明了用户的个性化需求，因为冷门物品属于长尾物品，往往不容易引起用户的关注。所以，用户产生过行为的物品当中，冷门物品对应的特征权重应该会更高。

2.3 用户兴趣模型表示方式

个性化推荐首先是要获取和感知用户的兴趣和需求，描述清楚用户的兴趣，才能通过计算比对推荐合适的内容和信息。描述用户兴趣，需要用到科学合适的方法来表示，构建出能够反映用户个体特征和兴趣偏好的兴趣模型。用户兴趣模型的表示主要采用主题表示法、关键词表示法、用户—项目评价矩阵表示法、基于向量空间模型表示法等几种方法。

2.3.1 主题表示法

主题表示法以用户感兴趣的信息的主题来表示用户模型[63]。如用户对于绘画、摄影感兴趣，则用户模型表示为{绘画，摄影}。各类网站个性化入口的用户模型通常采用该类表示方法，如 google、Microsoft、AOL 等。这种表示方法比较简单，但是主题词涵盖的范围比较广，很难精确描述用户的兴趣。

2.3.2 关键词表示法

关键词表示法是指以用户感兴趣的信息的关键词来表示用户模型[63]。它通过若干个与用户感兴趣的信息的主题相关的关键词来表示用户的兴趣模型。比如用户对数据挖掘感兴趣，则可以用{数据挖掘，大数据，hadoop，分布式数据库，算法，数

据分析}等一组关键词表示用户模型。关键词既可以由用户指定，也可以通过机器学习得到。WebWatcher 是一款网页监视工具，可以提取用户输入的个人感兴趣的关键词。TAGUS 系统、BGP-MS 系统则是通过机器学习的方式获取用户感兴趣的关键词。

2.3.3 基于向量空间模型的表示法

读者的兴趣是广泛的，对于多个兴趣往往有所侧重。基于向量空间模型是一种常用的文档描述结果，主要是通过关键词与关键词权重值组成的键值对集合{（t_1, w_1),(t_2,w_2)…(t_n, w_n) }来综合描述用户的多个兴趣，是一个 n 维特征向量，其中 t_i 为第 i 个兴趣关键词，w_i 为兴趣关键词 t_i 在文档中的权重[64]。每个向量的权重表示该兴趣向量在整个模型中的重要程度。这种方法将表示用户兴趣的文本数据转化为容易处理的结构化数据，将兴趣关键词等文本信息的处理简化为向量空间中表示兴趣项的向量的计算，是文本处理的常用方法。Fab 系统和 Web Watcher 系统都是使用一组加权关键词向量来描述用户兴趣。由于单个关键词难以全面地、完整地描述用户某方面的兴趣偏好，可能会造成推荐结果不够准确。

2.3.4 细兴趣粒度表示法

上文提到的关键词表示法、基于向量空间模型的表示法均属于粗兴趣粒度表示法。该方法主要按用户感兴趣的和不感兴趣的来分类，但表示用户兴趣的关键词无法确定相应的兴趣类别，构建的用户模型也就不能够细致地描述用户的每个兴趣主题。而细粒度的用户模型可以细分用户的兴趣种类，在各个种类中关联兴趣关键词，这样既有利于提供高质量的个性化服务，也有利于理解用户兴趣模型，方便模型修改和补充[63]。

2.3.5 用户—项目评价矩阵表示法

基于协同过滤的推荐系统多采用“用户—项目”评价矩阵表示法。“用户—项目”评价矩阵是一个 m×n 的二维矩阵，其中 m 为系统用户数，n 为项目数，矩阵中的每个元素 r_{ij} 表示了用户 i 对项目 j 的评价或偏好。这种表示方法简单易用，模型直观简洁，可根据系统中用户对项目的评价直接生成。在新闻、视频、音乐等资讯服务类网站中，通过让用户对所提供信息资源打分评级或者选择喜欢、不喜欢，来生成“用户—项目”评价矩阵。对于多维的评价空间还可以采用扩展的 N 维“用户—

项目”评价矩阵。

2.4 本章小结

本章依次介绍了协同过滤推荐、基于内容的推荐、基于标签的推荐、社会化推荐、上下文感知推荐等几种主流的推荐算法，还从用户行为数据、用户行为特征两方面介绍了用户行为特征处理技术的相关理论知识，最后介绍了主题表示法、关键词表示法、基于向量空间模型表示法、细兴趣粒度表示法、“用户—项目”评价矩阵表示法等几种常用的用户兴趣模型表示方式。

第 3 章 自适应个性化图书推荐算法

3.1 图书馆业务特点分析

图书馆的职能是保存人类文化遗产、开展社会教育、传递科学情报。图书馆的主要服务对象是读者，需要做好读者服务工作，引导读者合理利用馆藏资源。图书馆读者具有人数众多、年龄不同、专业方向不同等特点，在读者利用图书馆文献资源的过程中会留下诸如读者基本信息、借阅历史、操作记录等大量有价值的信息，这样的读者特点给图书馆提出了不同的个性化要求。如何满足读者的需求，提高读者的满意度，给读者提供更好的服务，是一个值得研究的问题。开展自适应个性化图书推荐算法研究工作，可以从大量的、不完全的、有噪声的、随机的实际应用数据中，通过仔细分析读者产生的借阅行为，提取隐含在其中的、潜在有用的信息，开发出精确的预测模型，实现以读者为中心的个性化服务。

以前，图书馆的信息管理工作是手工作业的模式，书目数据通过卡片式目录管理，图书借还通过人手进行登记。从上世纪九十年代开始，经过了二十多年的发展，我国图书馆的管理从手工操作时代发展到了自动化、信息化、智能化时代。图书馆的采访、编目、流通等业务采用自动化系统来管理，业务数据采用 Oracle、SQL 等数据库管理系统来管理，主要的业务数据有书目数据、馆藏数据、读者数据、流通数据等。随着图书馆业务的发展，读者量的不断增加，这些业务数据也不断地增多。以广州市越秀区图书馆为例，截止至 2016 年 6 月，注册读者已达 8 万多人，馆藏文献资源已达 60 多万册，读者的流通业务数据记录已超过千万条。在如此大量的数据中，要挖掘出读者的阅读习惯，潜在的阅读需求，并作出相应的图书信息推荐，需要设计算法对大规模的数据集进行运算。近年来，Hadoop、Druid、TensorFlow 等大数据技术的涌现，以及各种推荐算法的提出，为图书馆开展个性化的图书推荐服务提供了技术支持。

图书馆开展图书推荐服务需从自身实际出发，图书推荐的方式、方法与图书馆的服务对象、图书馆开展的业务有着密切的关系。例如，公共图书馆、少年儿童图书馆、科技图书馆、小学图书馆、中学图书馆、高校图书馆，这几类图书馆虽然都有着作为图书馆的共同职能，但由于服务对象不同，所开展的业务也必然不尽相同，有着各自的服务特色，所收藏的图书也有着各自的馆藏特点。不同的目标群体，有

着不同的知识水平，自然也有着不同的阅读需求，图书馆开展的图书推荐服务也要因应各自的实际情况制定相应的推荐策略。本文以广州市越秀区图书馆为例，对其日常流通业务和读者管理业务进行分析，并设计合适的自适应个性化图书推荐算法。

3.1.1 图书馆流通业务特点分析

图书馆最基本最主要的业务是图书借阅，通过借阅行为，实现图书资源共享以及知识的传播。传统的借阅行为由图书馆工作人员、图书、读者三个要素构成。随着科技的快速发展，近年来很多图书馆都将高新技术与传统的图书馆服务融合在一起，为读者提供多元化、人性化、智能化的服务。引入 RFID 技术开展自助服务是当前图书馆服务的一大特色。RFID 技术是图书馆应用得比较多的技术之一，RFID (Radio Frequency Identification)又称无线射频技术，是一种自动识别技术，能通过射频信号和空间耦合传输特性来实现对图书的自动识别。RFID 技术结合图书馆特有的工作流程，从而形成图书馆 RFID 系统。20 世纪 90 年代后期 RFID 技术开始应用于图书馆，2006 年国内图书馆开始应用并得到迅速的扩大。图书馆引进 RFID 技术后，图书馆工作人员的角色由配备 RFID 系统的自助借还设备来替代，读者的借阅行为演变成读者的自助借阅行为。自助借还服务模式对于读者来说，具有操作便捷、自由度高，私密性强等特点。对于图书馆来说，既提高了流通外借的服务效率，又提高了馆藏图书的利用率。对于图书馆的图书推荐业务，由于自助借还设备介入借还书操作流程，在分析读者借阅行为时，需要考虑自助借还操作的特性，例如读者借还书操作频率加快，图书借阅次数增多等。读者在自助借还设备上的借阅行为与人工借还相比，可能会有细微的变化。图书馆为读者推荐图书，要考虑自助借还这一因素。

3.1.2 图书馆读者管理业务特点分析

图书馆的服务对象是读者，对于读者的标识是为每位读者分配一个读者证号。图书推荐是以读者过往的借阅记录作为依据进行推荐的。在实际系统中，就是在数据库中抽取一个读者证号的所有借阅记录进行数据挖掘与分析。由于业务的需要，在公共图书馆除了开通个人读者证外，还开通了集体借阅证，即一个读者证号对应多位读者。以广州市越秀区图书馆为例，在系统中就存在校园证和家庭证这两种集体借阅证。校园证，是学校团体使用的，是图书馆与辖区范围内的中小学合作共同推广课外阅读而开通的一种集体借阅证。学校用校园证统一借书，供本校学生阅读。

家庭证，是供家庭使用的，是图书馆为了方便那些每位家庭成员都办了读者证的家庭到图书馆借书不用携带多张证而开通的一种集体借阅证。家庭证的借阅量是家庭证下关联的多张读者证借阅量的总和。对于这两种集体借阅证，一个证里面包含了多位读者的借阅记录，在数据分析时，如何区分不同读者的借阅记录，预测读者的阅读偏好，为读者做好图书推荐工作，是个性化图书推荐算法要解决的问题。

综上所述，图书馆必须根据图书馆自身的业务特点，选择科学高效的、最佳的推荐算法和推荐策略。所设计的图书推荐算法需兼顾考虑图书馆和读者双方的切身利益，既要充分地调动与利用图书馆的馆藏文献资源，尽量让每本图书都找到它的潜在读者，实现图书馆的资源效用最大化，也要根据读者在参与图书馆各项业务时留下的行为记录信息，充分研究了解读者，洞悉读者的潜在需求，让每位读者都找到所需的图书，从而实现图书馆和读者两者的双赢。

3.2 图书馆个性化推荐算法的适用性分析

当前，推荐算法有很多种，究竟哪种算法适合图书馆，需要结合图书馆的实际情况来考虑。下面，笔者结合图书馆的具体情况，根据各类推荐算法的特点，对各类推荐算法进行分析比对，选择较为适合图书馆图书推荐的算法。

推荐算法按用户数据类型来分类，主要可分为以下几种：

（1）协同过滤推荐算法

协同过滤推荐算法是基于系统中众多用户的历史行为数据计算目标对象之间的相似度，再根据相似度和用户的历史行为为用户推荐那些他可能会感兴趣而又没听说过的物品。该算法分为两类：给用户推荐那些和他们之前喜欢的物品相似的物品。这种推荐方法叫做基于物品的协同过滤推荐算法（简称 ItemCF）；建立 m×n 的用户评分模型，找到和目标用户相似度较高的邻居用户，然后将邻居用户喜欢但是当前用户未听说过的物品推荐给当前用户，这种推荐方法叫做基于用户的协同过滤推荐算法（简称 UserCF）。

（2）基于内容的推荐

为每个物品构建一个物品的属性文档，为每个用户构建一个用户的喜好文档，然后比较用户配置文件与产品配置文件之间的相似度，并直接向用户推荐与其配置文件最相似的产品[52]。其关键任务是对物品内容信息的提取，通常采取提取关键词的方法，如使用人工标注的标签抽取关键词，借助自然语义分析和情景感知提取等

[65-67]。内容特征的提取多以文本处理为主，对于图像、音频、视频等多媒体要提取内容特征目前还具有一定的技术障碍。

（3）基于标签的推荐

用户给物品定义一个或多个标签，根据用户最常用的标签，然后找到具有这些标签的热门物品推荐给用户。这种推荐方式需要系统先设置标签，用户亲自为物品选择标签。或者用户自定义标签，再为物品标识标签。

（4）社会化推荐

用户由共同的兴趣结识于不同的社交网络，形成了相互之间的信任。社会化推荐就是根据用户在某一社交网络中的好友感兴趣的内容为用户作出推荐。现在互联网上存在着各种各样带有社交性质的应用系统，如 Twitter、Facebook、微博、微信等，它们的每位用户都有好友列表，这里包含了用户之间的联系。

以上几种算法都需要用到数据信息，协同过滤推荐，需要用到读者的历史借阅记录，图书馆信息化系统中记录了大量的诸如图书借出时间、还回时间、借阅次数、操作人员、读者证号等借阅信息。基于内容的推荐，需要用到物品和用户的属性。图书的书名、作者、摘要等属性信息可在图书馆信息化系统的书目库获取到，但由于目前大部分图书馆都开通了自助办证服务,读者凭第二代身份证即可办理读者证，系统中保存的读者信息仅有姓名、年龄、性别、家庭住址等简单的基本信息，导致无法对读者做进一步的分析。基于标签的推荐，需要读者对图书定义标签，但是当前图书馆业务系统里一般没有这一功能，缺乏标签数据。社会化推荐，通过读者最信任的朋友作推荐，需要获取读者的好友列表，以及好友过往的行为数据。但是现在很多的社交网络，如微博、微信等开发商基于各种原因并没有向图书馆开放数据接口，没有读者的好友信息，此种推荐也很难开展。表 3.1 列出了四种推荐算法需要用到的数据的比较。表 3.2 根据第二章的个性化推荐算法理论知识给出了四种推荐算法的优缺点比较情况[68]。

表 3.1 四种推荐算法数据使用的比较

算法	需要使用的数据	数据在图书馆系统中是否具备
协同过滤推荐	用户历史行为	是
基于内容的推荐	物品属性、用户属性	是（用户属性数据较少）
基于标签的推荐	标签	否
社会化推荐	社交网络数据	否

表 3.2 四种推荐算法的优缺点比较

推荐技术	优点	缺点
协同过滤推荐	发掘用户新异兴趣；	冷启动问题；
	推荐个性化、自动化程度高；	稀疏性问题；
	能够处理复杂的非结构化数据；	推荐质量取决于历史数据集；
基于内容的推荐	推荐结果直观，容易使用户接受；	稀疏性问题；
	不需要领域知识；	无法处理复杂属性；
基于标签的推荐	连接用户与物品；	标签的随意性、歧义性；
	帮助用户发现喜欢的物品；	未区分场景；
		标签信息源单一；
		缺乏个性化；
社会化推荐	推荐结果容易使用户接受；	好友之间兴趣不一致导致推
	可以解决冷启动问题；	荐精度不高；

在 3.1 节提到图书馆有校园证和家庭证两种集体借阅证，不同的群体，虽然包含的个体不一样，但相互之间是存在着共性的。小学或中学办理的校园证所借图书能够体现出小学生或中学生的性格特征，同一年龄层、同一地区的学生会有近似的兴趣爱好。每个家庭成员构成、收入水平、受教育程度的不同，阅读兴趣偏好也会有所不同。反过来，如果能够在图书馆的借阅记录中找到与某一所学校相似的同类学校，或者找到与某个家庭相似的同一类家庭，就可以将相似群体借过的图书推荐给目标对象。

综上所述，面向图书馆读者的图书推荐使用基于用户的协同过滤算法是比较适合的。此算法的关键是计算目标用户与其他用户之间的相似度，主要有皮尔逊相关系数、余弦相似性或修正的余弦相似性等三种计算公式。

（1）余弦相似度

每个用户可对所有物品进行评分，两个物品 i ，j 视作为两个 m 维用户空间向量，通过计算两个向量的余弦夹角计算相似度。那么，对于 m×n 的评分矩阵，i ，j 的相似度 sim(i , j) 计算公式为：

$$sim(i,j) = cos(I,J) = \frac{I \bullet J}{\|I\| \times \|J\|} \tag{3.1}$$

（2）修正余弦相似度

余弦相似度计算仅考虑向量维度方向上的相似而没考虑到各个维度的量纲的差异性，也就是说没有考虑到不同的用户的评分标准的差异性，有的用户评分更宽容普遍打分较高，有的用户评分更严格，普遍打分较低。修正余弦相似度为了克服这一缺点，需要去中心化，对每位为物品 i 评过分的用户 u，计算其评分的均值，调整评分向量为评分偏差向量，再进行求解余弦相似度。

$$sim(i,j)=\frac{\sum_{u\in U}(R_{u,i}-\overline{R_u})(R_{u,j}-\overline{R_u})}{\sqrt{\sum_{u\in U}(R_{u,i}-\overline{R_u})^2}\sqrt{\sum_{u\in U}(R_{u,j}-\overline{R_u})^2}} \tag{3.2}$$

（3）皮尔逊相关系数

皮尔逊相关系数是一种度量两个变量间相关程度的计算方法。考虑到读者评分的差异性，也采用了中心化方法，但与修正余弦相似度不同的是，先计算每个物品 i 被评分的均值，再计算相关度。

$$sim(i,j)=\frac{\sum_{u\in U}(R_{u,i}-\overline{R_i})(R_{u,j}-\overline{R_j})}{\sqrt{\sum_{u\in U}(R_{u,i}-\overline{R_i})^2}\sqrt{\sum_{u\in U}(R_{u,j}-\overline{R_j})^2}} \tag{3.3}$$

在图书馆开展的业务里，读者不用为图书评分，想看的书直接借走，因此，可将借书操作视为评分操作，读者每借一本书即视为对此书评 1 分。由于每本书分值相同，不存在评分的差异性问题，且每位读者所借图书数量对于图书总量来说是非常少的，评分矩阵比较稀疏，所以，采用余弦相似度计算用户之间相似性会比较合适。

得到读者之间的相似度后，基于用户的协同过滤算法会给目标读者推荐和他兴趣最相似的 K 个读者借过而目标读者未借过的图书。以下公式计算了该算法中目标读者 u 对图书 i 的感兴趣程度：

$$p(u,i)=sim(u,v)\,r_{vi} \tag{3.4}$$

r_{vi} 代表了读者 v 对图书 i 的兴趣。最后，按照感兴趣程度由高到低排序，取 Top-N 得到最终的推荐列表。

3.3 读者兴趣模型的改进

基于用户的协同过滤推荐算法，关键的步骤是读者相似度的计算，主要利用读者行为的相似度计算兴趣的相似度。在传统的推荐系统中，基于协同过滤的推荐系统多采用二维“用户—项目”评价矩阵 (Recommendation Space, RS)表示读者兴趣

模型。模型通过系统中的用户行为生成，计算用户之间的相似度，将相似用户喜欢的物品推荐给目标用户。

在图书馆业务系统里，读者的行为隐含在系统的日志数据中，每位读者对图书馆馆藏图书的借阅情况可用 0 或 1 表示，0 为未借阅过，1 为借阅过，用 $\{b_1,b_2,b_3,\cdots\cdots,b_n\}$（n 为馆藏图书数量）即可表示系统中读者与图书的借阅关系。可用表 3.3 的“读者—图书”矩阵表示所有读者的行为。

表 3.3 协同过滤推荐算法“读者—图书”矩阵表示

读者	$Book_1$	$Book_2$	……	$Book_n$
$User_1$	R_{11}	R_{12}	……	R_{1n}
$User_2$	R_{21}	R_{22}	……	R_{2n}
……	……	……	……	……
$User_m$	R_{m1}	R_{m2}	……	R_{mn}

m 表示数据库中读者数量，n 表示数据库中图书数量，R 表示第 m 个读者是否借过第 n 本图书。

这种矩阵表示法是读者兴趣模型的一种表示方式，用 m×n 的二维“用户—项目”评价矩阵描述读者的兴趣模型。其中，m 为读者数量，n 为图书数量，矩阵中的每个元素 R_{ij} 表示了读者 i 对图书 j 的评价或偏好。该读者兴趣模型可以用以下公式表示：

$$\{(B_i, R_i)| i=1,2,\cdots,n\} \quad (3.5)$$

B_i 是图书，R_i 是读者对图书 i 的评价，R_i 的取值为 1 或 0，n 是图书的总数。

通过计算两位读者兴趣模型的相似度，就可以找到目标读者的最近邻居集，并根据最近邻居的借阅行为为目标读者做出合适的推荐。

由于在我国大多数的图书馆都普遍存在着被借阅的图书占图书馆总馆藏量比例过低的现象，如广州市越秀区图书馆，其注册读者数量约 8 万人，馆藏中外文图书文献 60 万余册，导致在整个图书馆数据库中，n 比 m 大很多。而在海量的馆藏图书中，被读者借阅过的图书占图书馆总馆藏的比例过少，而且不同读者的借阅记录之间，图书借阅的重叠项也不多。这些现实情况容易造成图书借阅信息存在极大稀疏性，使用传统的协同过滤算法就面临着目标用户最近邻居集难找而导致推荐效果降低及相似性计算耗费大等问题。而且，随着图书馆读者数量和馆藏图书数量的逐年递增，在搜索目标读者的最近邻居时，计算量大而耗时，系统性能会越来越低。

举个例子，假如读者 A 的借阅书单为：社交日语、日语会话、三体、时间移民、IOS 程序设计、IOS 9 应用开发实战；读者 B 的借阅书单为：凡尔纳科幻小说、星际穿越、寻找外星人、摄影的艺术、Android 编程实战。

用“读者—图书”矩阵来表示两位读者的借阅行为，见表 3.4。

表 3.4 读者 A 和读者 B 的“读者—图书”矩阵

读者	社交日语	日语会话	三体	时间移民	IOS 程序设计	IOS 9 应用开发实战	凡尔纳科幻小说	星际穿越	寻找外星人	摄影的艺术	Android 编程实战	Bi	…	Bn
A	1	1	1	1	1	1	0	0	0	0	0	0	…	0
B	0	0	0	0	0	0	1	1	1	1	1	0	…	0

用余弦相似度公式计算表 3.4 中读者 A 和读者 B 的相似度，$sim(A, B)=0$。

由于两位读者的借阅记录当中没有相同的图书，两者兴趣相似度为零。倘若仔细观察两位读者所借图书类别，其实两者所借图书是存在共同点的，读者 A 借阅的三体、时间移民和读者 B 的凡尔纳科幻小说、星际穿越、寻找外星人都属于科幻小说，虽然两者所借图书书名不一致，但是图书所属类别是一致的，这说明两位读者都对科幻图书感兴趣。用表 3.5 把两位读者所借图书按类别分组。

表 3.5 读者 A 和读者 B 借阅清单

读者 A：

常用外国语类	科幻小说类	编程类
社交日语 日语会话	三体 时间移民	IOS 程序设计 IOS 9 应用开发实战

读者 B：

科幻小说类	摄影类	编程类
凡尔纳科幻小说 星际穿越 寻找外星人	摄影的艺术	Android 编程实战

从表 3.5 可以看出，读者 A 和读者 B 都借阅过科幻小说类图书和编程类图书，可以认为两者对此类图书有共同的兴趣，他们之间的兴趣相似度不应该为零。

图书馆的藏书虽然数量庞大，但图书是具有分类属性的，相同类别的图书之间或多或少存在一定的相关性。读者所借图书代表了读者的阅读兴趣，那么图书所属类别也能够反映出读者感兴趣的图书类型。读者的借阅行为除了可以用“读者—图书”矩阵表示外，也可以用“读者—图书类型”矩阵来表示。这样，读者评分的项

目数就由图书馆的所有馆藏图书数量变为图书类型数量，矩阵中 n 的数量可大为降低。而且，只要读者对同类型图书产生过借阅行为，都可以视为相近邻居。

在我国图书馆，普遍采用《中国图书馆分类法》（简称《中图法》）对图书分类。《中图法》使用字母与数字相结合的混合号码，基本采用层累制编号法，根据类目的不同等级，配以相应不同位数号码。《中图法》包括“马列主义、毛泽东思想，哲学，社会科学，自然科学，综合性图书五大部类，22 个基本大类，214 个二级类目，1482 个三级类目，51861 个主表，1969 个复分类目数。在图书馆，为了方便读者查找图书，每种图书均有唯一的索书号。通过索书号可以准确地确定馆藏图书在书架上的排列位置。索书号由分类号和书次号两部分组成。分类号依据《中图法》取号。书次号即具有相同分类号的图书的流水次序号，由 1-3 位阿拉伯数字组成。根据每种图书的索书号中分类号的一级类目、二级类目和三级类目，就可以对图书准确归类。

参照《中图法》，可以构建一个根据图书学科种类划定图书分类的“读者—图书类型”矩阵，见表 3.6：

表 3.6 协同过滤推荐算法“读者—图书类型”矩阵表示

读者	$Type_1$	$Type_2$	……	$Type_n$
$User_1$	P_{11}	P_{12}	……	P_{1n}
$User_2$	P_{21}	P_{22}	……	P_{2n}
……	……	……	……	……
$User_m$	P_{m1}	P_{m2}	……	P_{mn}

m 表示读者数量，n 表示图书类型数量，P 表示第 m 个读者对第 n 类图书的偏好值。P 值用读者借阅某类图书的数量与其借过的所有图书总数的比值量度。采用图书类型表示读者对图书的访问矩阵，矩阵中的项目数就由上万种图书数降低为上千个图书类目，有效解决了传统的协同过滤算法存在的数据稀疏性问题。

用“读者—图书分类”矩阵为读者兴趣建模，还可以提高用户相似度计算的精准度。

用读者借阅某类图书的数量与其借过的所有图书总数的比值表示读者对该类图书的偏好，结果如表 3.7：

表 3.7 读者 A 和读者 B 的“读者—图书类型”矩阵

读者	常用外国语类	科幻小说类	编程类	摄影类
A	0.33	0.33	0.33	0
B	0	0.6	0.2	0.2

$$sim(A,B)=\frac{0.33\times 0+0.33\times 0.6+0.33\times 0.2+0\times 0.2}{\sqrt{0.33^2+0.33^2+0.33^2+0^2}\times\sqrt{0^2+0.6^2+0.2^2+0.2^2}}$$

$$=0.696$$

读者间的相似度由零变为 0.697，有显著的提高。图书所属类别反映了图书的主题、内容等方面的属性，即使两位读者所借的不是同一本书，但如果所借图书属于同一类型，可以认为他们的兴趣是相同的或者相似的。

引入了图书分类来描述读者的兴趣偏好后，读者的兴趣模型可以由公式（3.5）改进为：

$$\{（C_{i,}\ P_i）\mid i=1,2,\cdots,n\} \qquad （3.6）$$

上式中，C 是图书类型 i，P 是读者对图书类型 i 的偏好值，n 是图书类型的数量。设定读者借阅过的每类图书的数量用 y 表示，那么对于某一类图书 C_i（i=1,2…n），P 可按以下公式计算：

$$P_i = y_i/(y_1+y_2+\cdots+y_n) \qquad （3.7）$$

本文将本节提出的读者兴趣模型改进方法记为 CUserCF 算法。

3.4 算法的改进

3.4.1 算法改进的思路

协同过滤算法是基于读者的行为数据分析设计的，读者行为数据在图书馆的业务系统里是以日志方式存在的。在系统运行过程中，会产生大量的原始日志文件，存储在文件系统中。这些日志记录了读者的各种行为，与图书借阅相关的行为主要包括借阅图书、还回图书、续借图书等。每条日志都详细记录了操作类型、用户（读者）、对象（图书）、操作人员、时间等要素。一般来说，在个性化推荐中用户的行为按照反馈的明确性来区分的话，可分为两种：显性反馈行为和隐性反馈行为。显性反馈行为包括用户明确表示对物品喜好的行为，比如读者对图书的评分。但是，有时侯读者对于自身兴趣也不十分清晰，或者由于隐私的原因而无法提供准确的信息。隐形反馈行为是指那些不能明确反映用户喜好的行为。比如在图书馆的日志数据中所记录的图书借阅日志就属于隐性反馈行为,这些日志的数据量是非常庞大的，其中包含了许多的用户行为特征。另外，用户行为还可以按照反馈方向分为正反馈和负反馈。正反馈指用户的行为倾向于指用户喜欢该物品，负反馈指用户的行为倾

向于指用户不喜欢该物品。比如在图书馆的日志数据中读者借阅图书的记录就属于正反馈行为。正反馈的隐性行为中包含了许多的用户行为特征。

用户行为特征是从用户的行为中计算出来的特征向量。一个特征向量由特征和特征的权重组成。比如，读者阅览图书的行为和借阅图书的行为，这两种行为都会对图书特征的权重产生影响，但不同的行为影响不同。阅览图书的行为不代表读者喜欢该书，可能只是想大致了解图书的内容再决定是否外借。而借阅图书这一行为则代表读者确实喜欢该书，想把书借回去认真看。因此，借阅图书的行为应具有较高的权重。再比如，图书的借阅时间先后、图书的借期长短等行为特征都反映了读者对图书的感兴趣程度。在得到读者的所有特征向量后，可以综合考虑各种特征向量，并根据特征找到推荐的图书。图 3.1 是基于特征的图书推荐示意图。

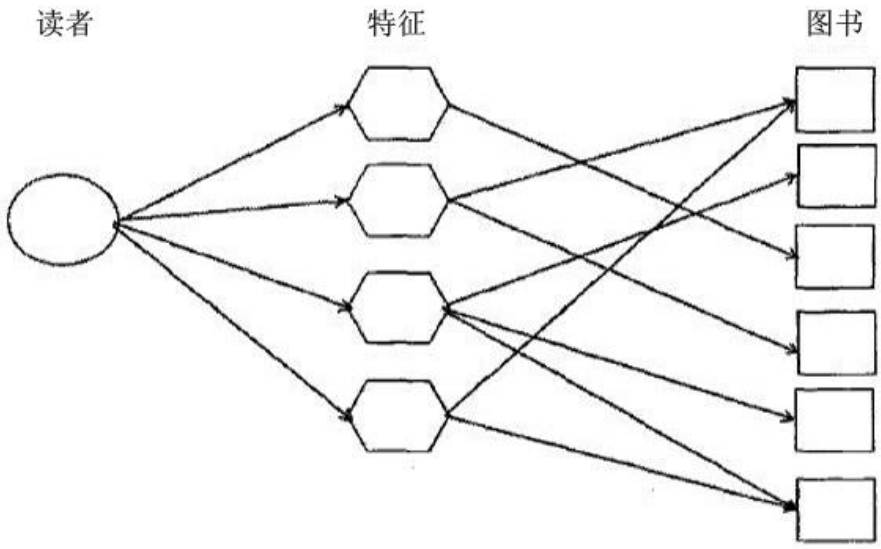

图 3.1 基于特征的图书推荐示意图

个性化图书推荐就是要分析所有读者行为，抽取出行为特征，揭示读者行为与其兴趣爱好之间的关联关系，从而为读者推荐图书。

在图书馆业务系统的流通日志里面，记录了某一时刻系统发生的行为事件。每条记录用操作人员、操作对象、操作类型、操作时间来表示。操作人员即读者，操作对象即图书，操作类型即借书、还书、续借等操作，操作时间即该次操作发生的时间。这些日志隐含了读者借书、图书借出时间、图书借阅时长、图书借阅次数等行为特征。基于用户的协同过滤算法就用到了读者借书这一行为特征，根据这一行为构建出“读者—图书”评价矩阵，但对于其他行为特征则并未考虑到。第 3.3 节提出的“读者—图书类型”评价矩阵，也是仅考虑到读者借书这一行为特征。其实，

时间是很重要的行为特征，近期的行为应比远期行为具有更重要的参考意义。还有，图书的借阅次数反映了图书的受欢迎程度，在图书推荐的时候同样是一项重要的参考依据。针对图书馆的个性化推荐，在采用协同过滤推荐思想的同时，也应着眼于读者行为特征的综合分析处理。

世界的万事万物都在发展变化当中，读者的兴趣与阅读需求也并非一成不变，而是不断地动态变化着的。本文研究的对象是自适应个性化图书推荐算法，也就是说所推荐的图书既要满足读者的个性化需求，又要随着读者需求的变化而随之改变推荐策略。“自适应”就是指不需要读者事先明确提出，就能及时地捕捉到读者阅读兴趣的变化，随之而动态调整推荐的图书列表，使最终的推荐结果始终符合读者的真实需求。读者的阅读兴趣其实都隐含在个体的各种行为里面，某种借阅行为发生改变，并不会马上对全局的分析产生影响，但是当这种行为持续发生的时候，该行为特征对应的权重就会发生变化，如果能够自动跟踪读者行为特征的变化，就能够不断修正该行为特征的权重值。从图 3.1 可以知道行为特征是用于联系读者和推荐图书的，感知到读者某种行为特征的权重改变，就可以调整图书的推荐列表。

本文 3.2 节提到，由公式（3.4）可得到图书的推荐列表。公式中 p 的大小主要取决于两个参数——sim(u,v)和 r_{vi}。其中，r_{vi} 是读者 v 对自己所借图书 i 的兴趣度，在分析读者行为生成特征向量时计算，sim(u,v)是根据“读者—图书类别”矩阵计算的读者之间的相似度。自适应个性化图书推荐算法要反映读者当前的阅读偏好，实现自适应推荐，在计算 sim(u,v)和 r_{vi} 时，要充分考虑读者借阅行为特征变化对 sim(u,v)和 r_{vi} 的影响。定义读者行为特征向量的权重为 W，那么读者 u 和读者 v 的相似度可用以下公式计算：

$$sim(u,v) = cos(U,V) \times W \tag{3.8}$$

r_{vi} 可用以下公式计算：

$$r_{vi} \times = W \tag{3.9}$$

读者行为包含正反馈行为和负反馈行为，将读者借阅某本图书所发生的所有相关行为特征向量的当前权重值综合计算，通过公式（3.4）、（3.8）、（3.9）可计算出读者对该本图书的兴趣度，再取 TOP-N 就能得到图书推荐列表。

3.4.2 基于用户行为特征处理技术的算法改进

基于用户的协同过滤算法通过计算读者之间借阅图书的行为的相似性进行推

荐，其中读者的借阅行为表明读者对所借图书感兴趣，但读者借书这一行为本身并不包含读者对图书感兴趣程度的描述。虽然这种对借阅过的每本图书同等对待的推荐方式也能起到一定的推荐效果，但如果能够考虑感兴趣程度这一因素，推荐结果必定更加丰富、更加灵活，甚至乎，如果能够把不感兴趣的图书也考虑到，那么推荐算法将更加完美。读者的正反馈行为和负反馈行为，分别倾向于指其感兴趣和不感兴趣的图书，基于用户行为特征处理技术，充分挖掘读者行为信息，可以获取这些数据。

3.4.2.1 基于正反馈行为的改进

读者的正反馈行为中，时间是度量读者对图书感兴趣程度的重要因素。读者阅读图书的行为是持续的行为，随着时间的流逝，读者的成长，读者的阅读兴趣也会改变，读者在不同时点阅读的图书反映了读者在当前时刻的阅读兴趣。上下文感知推荐算法将上下文信息融入到推荐算法中，计算处于相似时间上下文用户的相似性，让推荐系统能够准确预测用户在某种特定上下文情景下的兴趣。对于图书推荐来说，在使用协同过滤算法推荐时，搜集带有时间上下文信息的读者个性化信息，能够掌握读者兴趣偏移，图书感兴趣程度的改变，提高读者对图书兴趣度计算的准确度，提高自适应推荐的整体性能。

1. 图书借阅行为发生的时间

传统的协同过滤算法，只注重读者之间的相似性，将读者对每本图书的感兴趣程度均等化，而忽略了用户兴趣会随时间动态变化发展的情况。在现实生活中，时间对于读者兴趣的影响主要表现在以下方面：

（1）读者随着年龄的增长，周遭环境的改变，自身知识的积累，兴趣爱好是会发生变化的。比如说，某位读者小时候可能喜欢看漫画和动画片，长大之后就不怎么喜欢了。再比如某位读者参加工作后，因为工作的原因兴趣发生了改变。如果要准确预测读者当前的兴趣，就要关注读者近期的行为，因为近期的行为最能体现本人当前的兴趣。

（2）部分图书是有生命周期的。比如教辅类图书，随着国家教育理念的更新，教学课程的调整，考试大纲的更新，旧的教辅图书可能已经不再适合当前的教学要求了，那么这些书就慢慢地淡出读者的视线了。再比如计算机应用类图书，随着应用软件的更新，新技术的涌现，低版本的产品已经逐渐退出市场，相应的图书也逐渐地被淘汰了。因此，在为读者推荐某本图书时，需要考虑该图书在该时刻是否已

经过时了。

读者兴趣会不断改变这一现实情况对自适应的个性化图书推荐算法提出了以下要求：

（1）读者兴趣的变化体现在读者不断增加的新行为中，一个具备自适应功能的推荐算法需要能够自动跟踪读者新的行为，及时更新读者的兴趣模型，让推荐的图书列表及时调整，始终适应读者兴趣的变化。

（2）推荐算法需要平衡考虑读者近期行为和长期行为，即推荐图书列表既要反映出读者近期行为所体现的兴趣变化，又要兼顾考虑读者过往兴趣的延续性。

下面，提出 UserCF 算法的改进算法 TUserCF。UserCF 算法，两个读者兴趣相似，是因为他们对同一本书产生过借阅行为，但是，如果读者 A 和读者 B 同时借阅了图书 i，而读者 C 在更早的时间借阅图书 i，那么，可以认为读者 A 和读者 B 的兴趣相似度更高。

在系统日志中，读者借过的每本书均记录了借出时间，可以基于读者借书时间设置数据权重，以提高最近所借图书在推荐生成过程中的重要性。定义 WT(u,i)为基于读者 u 对图书 i 外借行为产生时间所反映的读者兴趣度的权重，ΔT_{ui} 为读者 u 借图书 i 的时间与当前时间的时间差。WT(u,i)是与ΔT_{ui} 相关的函数，ΔT_{ui} 越大，WT(u,i)越小，即对于$\Delta T_{ui}>\Delta T_{uj}$，有 WT(u,i)< WT(u,j)。WT(u,i)定义为：

$$WT(u,i)=1/(1+e^{\Delta Tui})$$

$$\Delta T_{ui}=T_{ui}-t \qquad (3.10)$$

这里，T_{ui} 为读者 u 借阅图书 i 的时间，t 为当前时间。WT(u,i)权重值在（0，1）区间内。

两读者相似度的计算公式为：

$$sim(u,v) = cos(U,V) \times WT \qquad (3.11)$$

在找到和读者 u 兴趣相似的一组读者后，这组读者近期的兴趣显然相比这组读者以前的兴趣更接近读者 u 当前的兴趣。那么，应该为读者推荐和他兴趣相似的读者近期喜欢的图书。p 的计算公式如下：

$$p(u,i) = sim(u,v) \times r_{vi} \times WT \qquad (3.12)$$

2. 图书借阅行为持续时间的长短

图书馆是提供图书借阅服务的场所，读者可以在图书馆阅读图书，也可以将想看的书借走，借期限定为一个月，可以续借一次。根据图书馆的日志记录，不同的

人借不同的书，借期各有不同。这与书的厚度、属性以及个人的阅读速度有关。笔者以为，不论借期长短，只要借期在合理范围内，都代表了读者对该书的阅读兴趣和需求。在给读者作图书推荐时，应该以读者曾经借过的书作为参考依据。但对于借期不在合理范围内的书则应区别对待。

随着高新科技的发展，RFID 技术也引入到图书馆行业，图书馆通过 RFID 实现了图书自助借还服务。这为读者借还图书带来了极大的便利性、快捷性、自主性和私密性。对读者借阅行为进行分析，改进推荐算法时，也要结合自助借还这一因素充分考虑。为了让广大读者都能够享受到图书馆的文献资源，实现服务均等化，图书外借是有数量限制的。通常，读者在书架上找到一堆想借的书后，到自助机借书时如果发现想借的书超过最大可借数，读者会选择将已借图书中感兴趣程度较低的图书还回，再借更感兴趣的图书。或者，当读者在某个阅览区借完书后，在另一阅览区找到更想看的图书，而此时如果所借图书已达最大可借数上限，读者也会选择还回部分之前所借书中感兴趣程度较低的图书。这就会造成在图书外借记录中有部分图书的借期非常短。如果没有自助借还机，读者借还书需要图书馆工作人员代劳，可能还需要排队等候，而自助机上借还操作的便捷性和自主性让读者借书更随心所欲更自由。笔者对广州市越秀区图书馆 2010 年至 2015 年 8 万多位读者的 300 多万条借阅记录进行统计分析，其中图书外借借期在一天以内的达 6 万多人次，约占总体借阅量的 2%。这些图书的借期不长，从几分钟到几小时不等。这些图书外借之后又在短时间内被还回，极有可能是读者看过书后觉得不喜欢还回，又或者是发现有更想看的书而还回的。那么，如果在计算读者相似度以及计算读者对图书的兴趣度时，将这类书与借期在合理范围内的图书都视为读者感兴趣度相同的图书，显然是不合理的。因此，在生成用户特征的时候，应该降低这一类图书所对应的特征权重，削弱这类图书对图书推荐的影响。

本文提出 UserCF 算法的改进算法 DUserCF。定义 WD(u,i)为基于读者 u 借阅图书 i 这一行为的持续时间所反映的读者兴趣度的权重，T 为图书馆日常开放时间，为一常量，计算公式如下：

$$\begin{cases} WD(u,i) = (r_{ui} - l_{ui})/T & (r_{ui} - l_{ui} \le T) \\ WD(u,i) = 1 & (r_{ui} - l_{ui} > T) \end{cases} \quad (3.13)$$

其中，r_{ui} 是读者 u 借出图书 i 后的还回时间，l_{ui} 是读者 u 外借图书 i 的借出时间， WD 权重值在（0，1）区间内。

读者之间相似度的计算公式为：

$$sim(u,v) = cos(U,V) \times WD \quad (3.14)$$

在找到和读者 u 兴趣相似的一组读者后，这组读者曾经借过的而借期又不在合理范围内的图书需要降权。p 的计算公式如下：

$$p(u,i) = sim(u,v) \times r_{vi} \times WD \quad (3.15)$$

基于以上两种正反馈行为可综合计算读者 u 对所借图书 i 的感兴趣度的特证权重 W。W 计算公式如下：

$$W = WT \times WD \quad (3.16)$$

那么，基于以上两种正反馈行为生成的读者兴趣模型计算读者相似度的公式为：

$$sim(u,v) = cos(U,V) \times W \quad (3.17)$$

找到目标读者的最近邻居集后，目标读者 u 对图书 i 的兴趣可用以下公式计算：

$$p(u,i) = sim(u,v) \times r_{vi} \times W \quad (3.18)$$

3.4.2.2 基于负反馈行为的改进

个性化推荐系统的核心是用户兴趣模型，通过对模型相似度的计算进行推荐。用户兴趣模型是能够管理和记录用户行为，表述与存储用户兴趣偏好的一种数据结构。基于用户的协同过滤算法采用 m×n 的二维“用户—项目”评价矩阵表示用户兴趣模型，m 是用户数量，n 是项目数量。在图书推荐系统中采用这种模型表示方法，就是将读者行为历史中借阅图书的行为看作评分行为，将读者所借阅的图书看作读者感兴趣的图书，以此描述读者的兴趣偏好。读者作为一个富有情感、具有社会属性的自然人，兴趣是广泛的，同时也是具有两面性的，对于图书馆的书，必定有其喜欢的，也有其不喜欢的。如果用户兴趣模型只反映读者感兴趣的一面，而忽略其不感兴趣的一面，那么这个模型是片面的，是不完整的，并不能很好地表示用户的整体兴趣偏好。通过这样的模型做出的推荐准确度可能不是很高。换言之，在对读者行为进行跟踪与管理的时候，全方位地分析读者的兴趣特征，综合考虑读者阅读兴趣的正反两面，全面了解读者，感知读者感兴趣与不感兴趣的图书，生成“用户—项目”评分矩阵，有助于提高读者相似度计算的准确性。同时，读者感兴趣与不感兴趣的事物也是在不断变化的，为了实现自适应的个性化推荐，用户兴趣模型也必须跟踪到这些变化。用户兴趣模型中的数据不能一成不变，必须随时根据获取到的最新的读者信息动态更新。这就要求推荐算法必须持续不断地观察和记录读者行为，将从读者行为中得到的信息与用户兴趣模型中现存的信息融合，并且消除信

息的不一致，从而动态建立与更新模型，实现模型对读者兴趣的自适应。

读者行为既包括正反馈行为，也包括负反馈行为。读者借阅图书的行为倾向于表明读者喜欢该图书，属于正反馈行为；读者从未借阅过某类图书的行为倾向于表明读者不喜欢该图书，属于负反馈行为。正反馈行为和负反馈行为反映了读者的喜恶，可以用来分析读者对图书的感兴趣程度，并生成相应的兴趣模型。公式（3.6）用{(Ci，Pi)| i=1,2,…,n}表示读者兴趣模型，用 P 衡量读者对某类图书的感兴趣程度，就是基于读者正反馈行为建模。除此之外，还可以基于读者的负反馈行为建模，Pi 为正值表示读者对 i 类图书感兴趣，Pi 为负值表示读者对 i 类图书不感兴趣。这样的模型能够充分反映读者的兴趣面，基于读者模型计算读者相似度，能够提高图书推荐的准确度。而且，在向读者进行图书推荐时，如果事先知道了读者对哪些图书不感兴趣，还可以在最终的推荐列表中去除这部分图书，进一步提升自适应个性化图书推荐的精准度，使推荐结果更符合读者需求。

一般来说，读者借书这一行为属于正反馈隐性行为，所借的书基本上都代表了读者的阅读偏好。但反过来，读者没有产生过借阅行为的图书，并不能完全说明读者对这些书不感兴趣。对于读者没有产生过借阅行为的图书，如果是冷门图书，即使是属于读者感兴趣的图书范畴的，也极有可能因为读者不知道这本书的存在而没有在图书馆的检索系统中发现到它，或者由于书库太多书而没办法让读者注意到它。而如果是热门图书，则不存在这一种情况。读者可以通过借阅排行榜、畅销书架、媒体宣传或者亲朋好友介绍等多种途径知道哪些是近期的热门图书，只要图书馆有这些书，读者想借的话总有办法能借到。因此，可以认为没有出现在读者的借阅记录中而又很热门的图书更能说明读者对这本书不感兴趣，甚至很有可能读者对与该书相似的同一类书都不感兴趣。每个图书馆都会公布热门图书排行榜，向读者揭示近期受欢迎的图书。可以将图书排行榜里面读者没有借过的，同时又不属于读者兴趣模型{(C_i, P_i)| i=1,2,…,n}里面读者曾经产生过借阅行为的图书类型 C_i 的书，视为读者不感兴趣的书，把这些书组成图书列表 L：

$$Li=\{l_1,l_2,l_3,\cdots\cdots,ln\} \quad (3.19)$$

其中 l_i 代表读者不感兴趣的图书 i，n 是读者不感兴趣图书的数量。

然后将图书列表 L 中的图书按类别补充到读者兴趣模型{(C_i, P_i)| i=1,2,…,n}里面。设定读者不感兴趣的图书列表 L 中每类图书的数量用 z 表示，那么模型里每一种读者不感兴趣图书类型 C_i 的 P 值可按照以下公式计算：

$$P_i = z_i/(z_1+z_2+\cdots+z_n) \quad (3.20)$$

按以上方法将读者不感兴趣的图书类型反映到读者的兴趣模型中，一方面可以降低模型中 P 值为零的图书类型数量，改善矩阵模型的稀疏性，另一方面可以使模型涵盖读者感兴趣的和不感兴趣的图书类型，让读者的兴趣模型更符合实际，进而全面地揭示读者的兴趣特征。

图书列表 L 中，每本图书的热门程度与读者对该图书的不感兴趣程度有关，越热门的图书越能代表读者对该书或与该书相同类别的图书不感兴趣。在寻找读者的最近邻居时，如果能够考虑这一关系，就可以突出读者不感兴趣的图书，以便于找到最相似的读者。根据这一思路，可以通过以下公式计算读者对图书列表 L 中每本图书的不感兴趣度。

定义 WN(u,i)为基于读者 u 没借过的热门图书 i 所反映的读者不感兴趣度的权重，权重取值范围在（0，-1）区间。

$$WN(u,i) = 1/\ln(p_{ui}+e) - 1 \quad (3.21)$$

其中，p_{ui} 是读者 u 没有产生借阅行为的热门图书 i 的热门程度。

基于包含读者不感兴趣图书类型的读者兴趣模型$\{(C_i, P_i)| i=1,2,\cdots,n\}$，计算读者之间的兴趣相似度：

$$sim(u,v) = cos(U,V)\times WN \quad (3.22)$$

推荐算法的性能指标之一是提高推荐的准确率，与读者相似度最高的 K 个读者中，可能某位读者借过目标读者不感兴趣的图书，这些书即使放在推荐列表中推介给目标读者，最终转化为读者借阅行为的可能性也不大，为了提高推荐的准确率，可以在推荐列表中删除这部分图书。这种基于读者负反馈行为进行图书推荐的算法笔者记为 NUserCF。例如，《盗墓笔记》是当前流行图书，属于灵异类图书，读者 u 过往的借阅记录中并没有借阅过这套书，也未借阅过此类图书，那么，根据 NUserCF 推荐算法，可推算出读者 u 对灵异类图书不感兴趣，将这一分析结果补充到读者 u 的兴趣模型中，计算读者 u 的 K 个最近邻居，在根据借阅历史得出的 K 个最近邻居借阅过，而读者 u 没有借阅过的所有图书列表中剔除读者 u 不感兴趣类型的图书，得到的就是面向读者 u 的推荐列表。

社会学领域有一个著名的马太效应，即强者愈强，弱者愈弱。如果一种推荐算法会加大热门物品和非热门物品的流行度差距，让热门的物品更加热门，让不热门的物品更加不热门，那么这种算法就具有马太效应[69]。比如，图书馆的图书借阅排

行榜，就是对借阅量大的热门图书的一种推荐，增加了热门图书的曝光机会，无形中让上榜的图书更加热门。而没上榜的图书由于得不到推广，则变得更加不热门。再如，协同过滤算法通过计算用户的相似度进行推荐，会使得借阅了相同热门图书的读者相似度较高，使推荐结果偏向于热门图书，加大了马太效应。NUserCF 推荐算法关注到读者对图书种类的好恶，可以推断出读者不感兴趣的图书，同时避免读者不感兴趣的热门图书在推荐列表中出现，有助于缓解马太效应。

3.5 自适应个性化图书推荐算法

本文采用协同过滤算法的思想，以基于用户的协同过滤推荐算法为基础，设计了面向图书馆的自适应个性化图书推荐算法。通过对协同过滤推荐算法的优缺点分析，主要从三方面对该算法的不足之处进行了算法改良：

（1）针对协同过滤算法“用户—项目”评价矩阵存在的稀疏性问题和可扩展性问题，采用《中图法》的图书分类标准，设置图书类别，生成基于图书分类号的“读者—图书类型”评价矩阵，并给出了读者兴趣模型表示公式，读者对某类图书偏好值的计算公式，以此构建读者兴趣模型。

（2）针对协同过滤算法将读者借阅图书的行为同等对待，未有考虑读者对图书感兴趣程度会随时间发生改变这一情况，将协同过滤算法与上下文感知推荐算法混合使用，借助时间上下文信息，感知读者对所借图书的感兴趣程度的变化状况，给出了算法中度量读者 u 对图书 i 感兴趣程度的公式。

（3）针对协同过滤算法构建的读者兴趣模型中未能涵盖读者不感兴趣图书的不足之处，运用用户行为特征处理技术，提出基于读者负反馈行为的分析处理方法，找出读者不感兴趣的图书类型，将其纳入进读者兴趣模型中，使模型更贴合实际。

自适应个性化图书推荐算法的步骤如下：

输入：读者 u。

输出：读者 u 的 top-N 推荐集。

（1）对每位读者 $u \in U$，遍历借阅记录，找到该读者产生过借阅行为的图书，根据公式（3.10）按照图书借阅时间先后计算对应的特证权重 WT。

（2）对每位读者 $u \in U$，遍历借阅记录，找到该读者产生过借阅行为的图书，根据公式（3.13）按照图书借期长短计算对应的特证权重 WD。

（3）根据（1）、（2）两步计算出的特证权重，按照公式（3.16）计算每位读者

u 借阅每本图书的行为对应的特证权重 W。

（4）按照 3.3 节提出的读者兴趣模型改进方法，根据公式（3.6）构建读者兴趣模型，同时引入基于借书时间和图书借期的特征权重，利用第（3）步计算出的特证权重 W 调整模型中读者 u 对图书类型 i 的偏好值 P_i。

（5）对每位读者 $u \in U$，按照 3.4.2 节提出的研究读者负反馈行为特征感知读者不感兴趣图书类型的方法，找出读者 u 不感兴趣的图书，生成图书列表 L，根据公式（3.21）计算读者 u 对图书列表 L 中每本图书的不感兴趣度的权重 WN，按公式（3.6）和（3.20）补充完善读者兴趣模型，使读者模型涵盖读者感兴趣和不感兴趣的项。

（6）根据第（5）步生成的读者兴趣模型，由公式（3.1）计算目标读者与其他读者之间的相似度，生成目标读者的最近邻居集。

（7)从目标读者最近邻居集的图书借阅历史记录中删除目标读者 u 已经借阅过的图书，再删除属于目标读者 u 兴趣模型中不感兴趣类型的图书，得到候选推荐集。

（8）根据公式（3.9）用第（3）步得到的特证权重 W，计算最近邻居集中读者 v 对候选推荐集图书 i 的感兴趣度 r_{vi}。

（9）基于第（8）步计算结果，根据公式（3.4）计算目标读者 u 对候选推荐集图书 i 的感兴趣度 p_{ui}。

（10）将候选推荐集中的图书按目标读者 u 对候选推荐集图书 i 的感兴趣度 p_{ui} 从大到小排列，取 TOP-N 得到目标读者 u 的推荐集。

3.6 本章小结

本章围绕本文的研究主题，分析了公共图书馆的业务特点，分析对比了各类推荐算法用于公共图书馆开展图书推荐服务的适用性，结果显示采用余弦相似度公式计算用户兴趣相似度的协同过滤推荐算法相对于其他算法适用性更高。经过对基于用户的协同过滤推荐算法优缺点的研究分析，针对该算法存在的“用户—项目”评价矩阵稀疏性较大、可扩展性差，未考虑时间效应与读者对图书感兴趣程度的关系，读者兴趣模型未能从正反两面全面揭示读者兴趣偏好等问题，分别提出了 CUserCF、TUserCF、DUserCF、NUserCF 四种改进的推荐算法，通过将这几种算法组合运用提出一种面向公共图书馆的自适应个性化图书推荐算法。

第 4 章 实验与分析

4.1 实验目的

本文结合《中图法》的图书分类方法，对传统基于用户的协同过滤算法存在的矩阵稀疏性问题加以改善，并通过用户行为特征处理的技术手段从读者阅读行为的变化中找出隐含的阅读倾向的变化，对原有算法作进一步的改进。为了测量改进后算法的推荐效果，本文设计了一系列实验，与基于用户的协同过滤算法进行比较。

4.2 实验数据

实验使用了广州市越秀区图书馆自动化管理系统中的真实数据，包括书目数据库、馆藏数据库、读者数据库、流通数据库。图书馆业务系统中的原始数据，存在数据不完整、冗余、分散等问题，如果不加处理直接使用，会影响用户行为分析，导致推荐结果有偏差。在实验之前要对数据进行预处理：1、数据清理。系统在应用过程中不可避免地会发生一些故障，导致数据出现错误，这些错误的异常数据将影响数据分析的准确性，必须去除。2、数据简化。在数据库里会设置很多的字段记录各种信息，而对于用户行为分析，其实只需要诸如读者证号、图书索书号、图书题名、图书条码号、借阅行为类型、操作时间等字段，其余多余的数据可以去除。3、数据合并。理清各个数据库之间的字段关联关系。比如书目数据库中的书名和索书号，跟馆藏数据库中的图书条码号是一对多的关系。流通数据库中记录的读者借还的图书信息是条码号，而不同的条码号有可能实际上是对应同一个书名、同一个索书号的。那么就需要创建视图，把条码号转换为索书号，以降低数据分析过程中的时间复杂度。4、数据筛选。由于数据量过于庞大，涉及到图书馆 60 多万册图书以及 8 万多名读者，而且还有部分读者只有少量的图书借阅记录，有的甚至连一条记录也没有，导致数据集过于稀疏。为了提高实验的效率，同时使实验数据满足实验的要求，需要在流通数据库中筛选出使用图书馆服务时间跨度较长的，借阅量达到一定数量的，存在部分借期较短的借阅记录的活跃读者。

通过对数据的预处理，最终选定 2011 年 1 月 1 日至 2012 年 12 月 31 日期间，图书借阅量大于 30 册的借阅记录，作为实验数据集。数据集当中共包括 17702 位读

者对全馆 278371 种图书的 442223 条借阅记录，还包括了书目数据库、馆藏数据库和读者数据库的所有数据记录。

4.3 评测指标

在一般的提供推荐服务的系统中，是通过向用户提供个性化推荐列表的形式实现推荐的，这种推荐方法叫做 Top-N 推荐。在 Top-N 推荐中测评指标主要用于评价推荐算法的推荐效果。在统计精度量度方法上，通常采用平均绝对偏差 MAE 来测量算法预测的精度。在决策支持精度量度方法上，常采用准确率（Precision）、召回率（Recall）这两个评测指标来测量算法的推荐质量。

平均绝对偏差 MAE 是所有单个测量值与算术平均值的偏差的绝对值的平均，在推荐质量评价中，用来度量算法预测出的用户评分与实际用户评分之间的偏差程度，偏差值越小，说明算法预测准确率越高。设测试集中读者 u 对图书 i 的评分为 p，而实际的评分为 q，N 为图书数量，则 MAE 定义为：

$$MAE = \frac{\sum_{i=1}^{N} | p - q |}{N} \tag{4.1}$$

准确率是正确推荐的物品数量与推荐列表中所有物品数量的比率，用于考察为用户产生的推荐列表中有多少是用户真正感兴趣的物品。

推荐准确率：

$$Precision = \frac{\sum_{u \in U} | R(u) \cap T(u) |}{\sum_{u \in U} | R(u) |} \tag{4.2}$$

召回率是推荐列表中正确推荐的物品数量与测试集中用户全部感兴趣物品数量的比率。

推荐召回率：

$$Recall = \frac{\sum_{u \in U} | R(u) \cap T(u) |}{\sum_{u \in U} | T(u) |} \tag{4.3}$$

在以上两条公式中，R(u)是根据训练集数据计算得到的推荐列表，T(u)是用户在测试集上的访问数据列表。有时为了全面评测 Top-N 推荐的推荐效果，会选取多个推荐列表长度，计算全部推荐准确率和召回率，进行比较。

4.4 实验设计与结果分析

将实验数据集里面，借还操作时间在 2011 年 1 月 1 日至 2012 年 9 月 30 日期间的借阅记录作为训练集，在 2012 年 10 月 1 日至 2012 年 12 月 31 日期间的借阅记录作为测试集。从中随机挑选出 30 名在训练集和测试集里都有借书记录的读者作为目标对象，在训练集上训练，在测试集上对目标对象的借阅行为进行预测，最后评测改进的推荐算法在测试集上的预测结果。

为了全方位地测量本文提出的算法，共设计了五组实验。在所有实验中，实验对象均为随机挑选的同一批在训练集和测试集中都有借书记录的 30 名读者，读者最近邻居数 K 设为 80，观察 Top-N 推荐中 N 值从 10—50 以 10 为单位递增的情况下，推荐算法三个性能评测指标的比较。三个评测指标的实验结果是对 30 位读者推荐结果的平均值。

1. 第一组实验

将引入《中图法》的图书分类思想对“读者—图书”矩阵改进过的推荐算法 CUserCF，推荐的结果与 UserCF 算法对比。结果如下：

表 4.1 CUserCF 算法和 UserCF 算法实验结果对比

N 值	平均绝对误差		准确率		召回率	
	CUserCF	UserCF	CUserCF	UserCF	CUserCF	UserCF
10	0.815	0.851	48.49	40.29	4.89	3.26
20	0.803	0.835	46.25	38.12	7.83	5.51
30	0.801	0.826	43.31	35.38	8.94	6.51
40	0.792	0.819	40.27	33.56	9.71	7.6
50	0.787	0.813	38.62	32.57	10.39	8.58

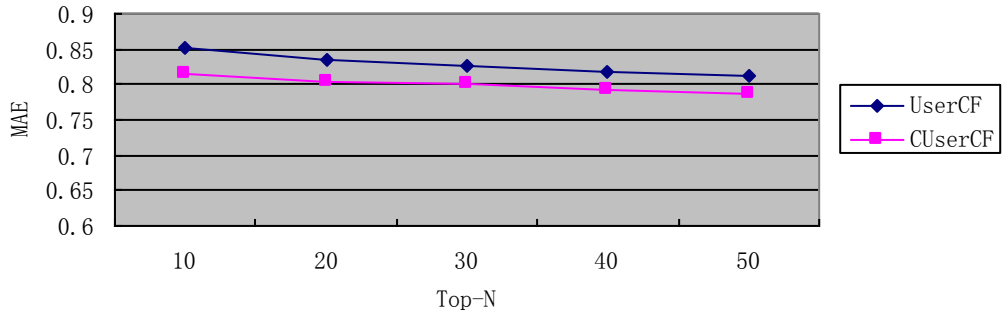

图 4.1 CUserCF 算法和 UserCF 算法实验 MAE 对比

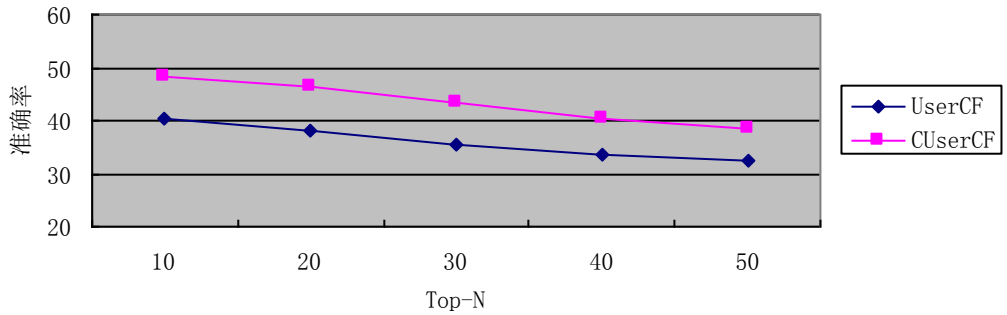

图 4.2 CUserCF 算法和 UserCF 算法实验准确率对比

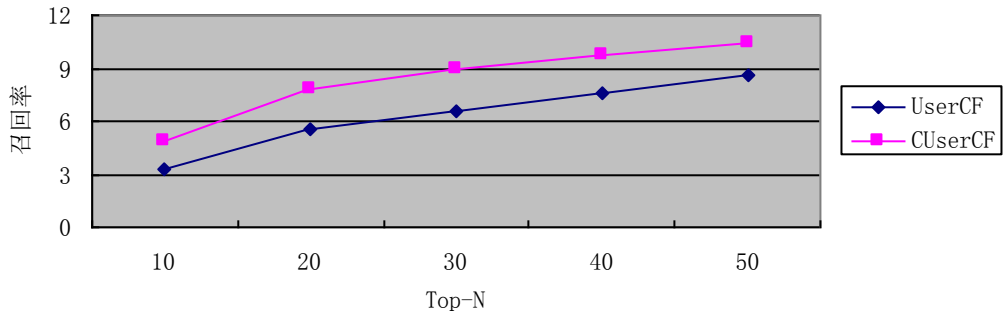

图 4.3 CUserCF 算法和 UserCF 算法实验召回率对比

从上表可看出，改进后的 CUserCF 算法与 UserCF 相比，平均绝对误差 MAE 有比较明显的提高，无论准确率还是召回率都有一定程度的提升。本实验结果表明结合《中图法》的分类方法构建的“读者—图书类别”矩阵，有效改善了原有评价矩阵的稀疏性问题，使原本毫无关联的图书因为分类属性而聚集在一起，将访问矩阵转化为读者对各类图书阅读倾向的描述，使对目标读者邻居集的计算更准确，从而在为目标读者进行自适应个性化图书推荐时能给出更准确、更全面的推荐结果。

2. 第二组实验

在 CUserCF 算法基础上，引入基于借书时间的特征权重，对读者进行图书推荐。将此推荐算法（TUserCF）的推荐结果与 CUserCF 推荐算法对比。结果如下：

表 4.2 TUserCF 算法和 CUserCF 算法实验结果对比

N 值	平均绝对误差		准确率		召回率	
	TUserCF	CUserCF	TUserCF	CUserCF	TUserCF	CUserCF
10	0.797	0.815	52.16	48.49	5.81	4.89

20	0.785	0.803	50.02	46.25	8.94	7.83
30	0.781	0.801	47.28	43.31	9.98	8.94
40	0.770	0.792	44.19	40.27	11.01	9.71
50	0.762	0.787	42.26	38.62	11.51	10.39

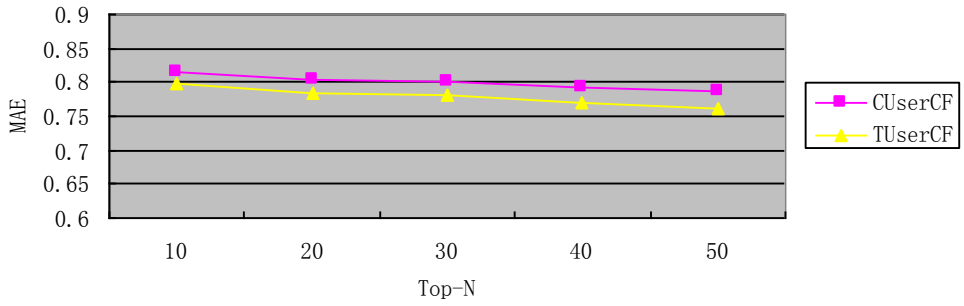

图 4.4 TUserCF 算法和 CUserCF 算法实验 MAE 对比

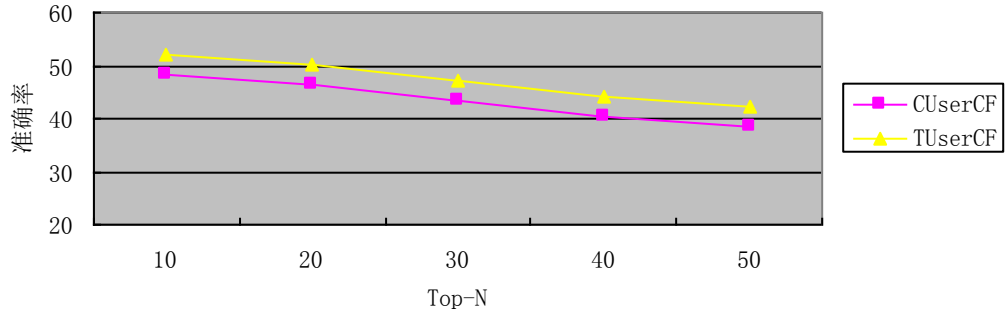

图 4.5 TUserCF 算法和 CUserCF 算法实验准确率对比

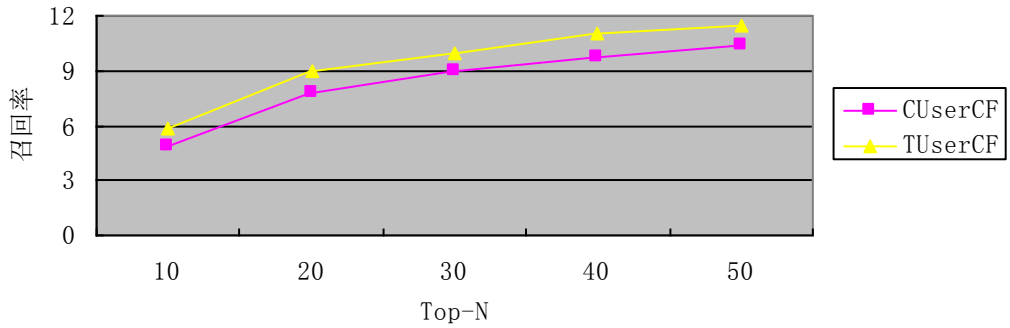

图 4.6 TUserCF 算法和 CUserCF 算法实验召回率对比

从上表可看出，在 N 不同取值情况下，TUserCF 算法的平均绝对误差 MAE 更小，无论准确率还是召回率，TUserCF 算法都明显优于对照的算法。尤其是 N 取值较少时，准确率有较明显的提高。这说明，引入基于借书时间的特征权重，有效增

加了每位读者近期借阅图书对读者当前兴趣偏好计算的贡献度，同时也突出了兴趣相似的一组读者当中，这组读者近期借阅的图书对目标读者兴趣预测的重要性。此种改进的算法，适应读者兴趣动态变化的特点，具有较好的自适应能力。

3. 第三组实验

在 CUserCF 算法基础上，引入基于图书借期的特征权重，对读者进行图书推荐。将此推荐算法（DUserCF）的推荐结果与 CUserCF 推荐算法对比。结果如下：

表 4.3 DUSERCF 算法和 CUSERCF 算法实验结果对比

N 值	平均绝对误差		准确率		召回率	
	DUserCF	CUserCF	DUserCF	CUserCF	DUserCF	CUserCF
10	0.807	0.815	50.16	48.49	5.63	4.89
20	0.795	0.803	47.83	46.25	8.65	7.83
30	0.790	0.801	44.39	43.31	9.59	8.94
40	0.779	0.792	41.2	40.27	10.92	9.71
50	0.773	0.787	39.87	38.62	11.27	10.39

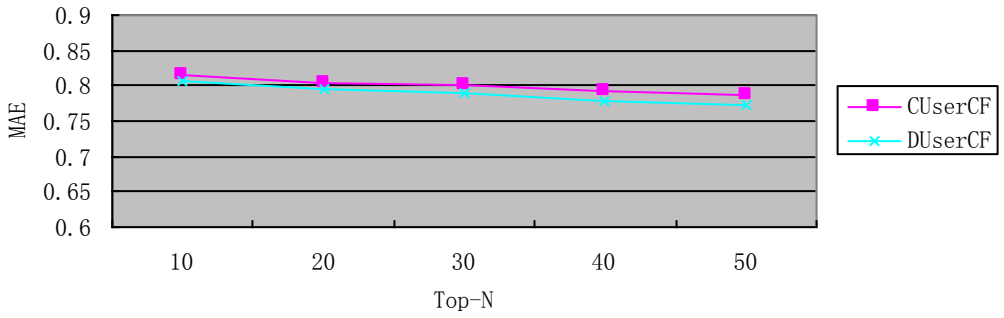

图 4.7 DUserCF 算法和 CUserCF 算法实验 MAE 对比

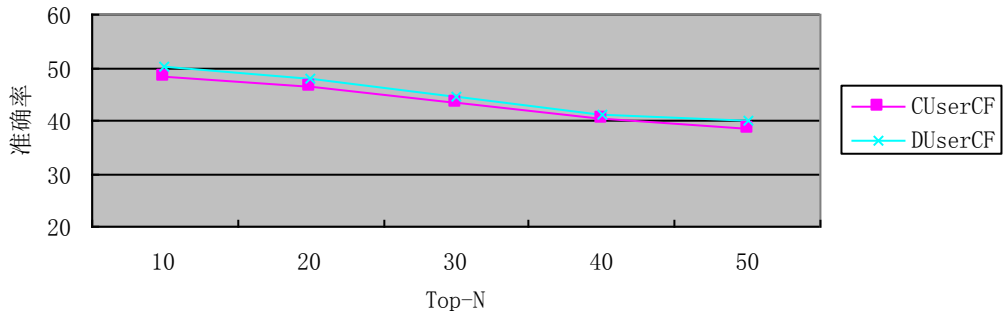

图 4.8 DUserCF 算法和 CUserCF 算法实验准确率对比

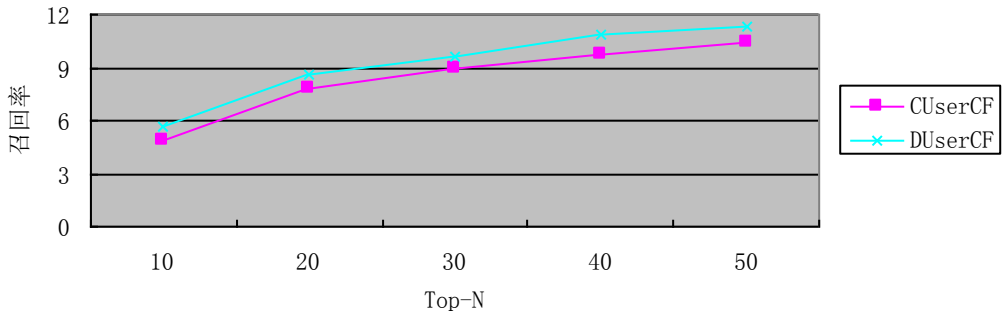

图 4.9 DUserCF 算法和 CUserCF 算法实验召回率对比

从上表可看出，在 N 不同取值情况下，DUSERCF 算法的平均绝对误差 MAE、准确率和召回率都略优于对照的算法。这说明，在计算读者兴趣相似度和读者对图书感兴趣度时，考虑图书借期这一因素，降低借期非常短，可认为读者对其兴趣度低的图书的权重，是有必要的。

4. 第四组实验

在 CUserCF 算法基础上，基于负反馈行为推测读者不感兴趣的图书类型，对算法加以改进，将此推荐算法（NUserCF）的推荐结果与 CUserCF 推荐算法对比。结果如下：

表 4.4 NUserCF 算法和 CUserCF 算法实验结果对比

N 值	平均绝对误差		准确率		召回率	
	NUserCF	CUserCF	NUserCF	CUserCF	NUserCF	CUserCF
10	0.799	0.815	50.98	48.49	5.57	4.89
20	0.786	0.803	48.66	46.25	8.59	7.83
30	0.782	0.801	45.39	43.31	9.17	8.94
40	0.775	0.792	42.67	40.27	10.23	9.71
50	0.766	0.787	40.25	38.62	10.86	10.39

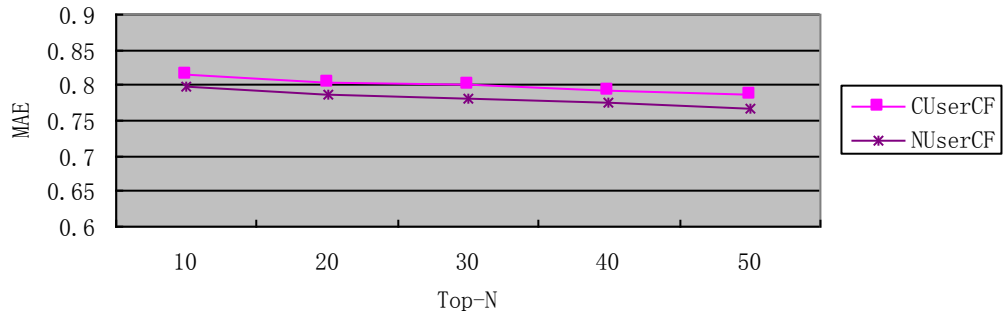

图 4.10 NUserCF 算法和 CUserCF 算法实验 MAE 对比

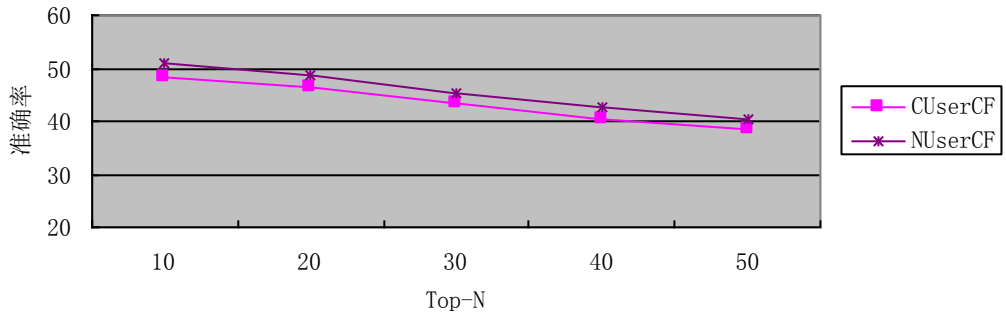

图 4.11 NUserCF 算法和 CUserCF 算法实验准确率对比

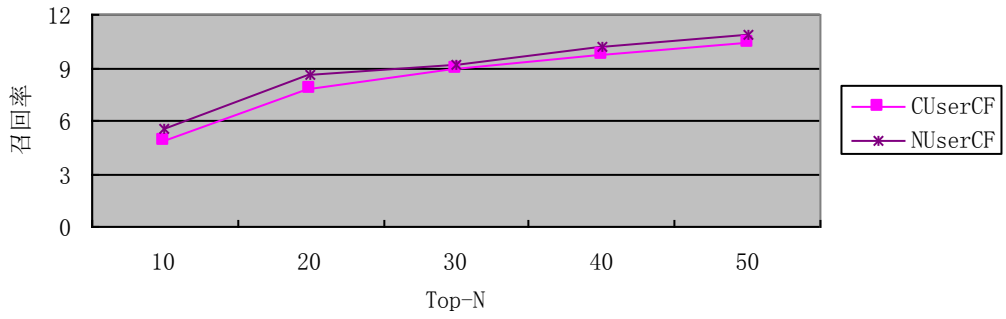

图 4.12 NUserCF 算法和 CUserCF 算法实验召回率对比

从上表可看出，NUserCF 算法的各项性能指标都有所提高。这说明基于读者负反馈行为进行读者行为特征处理，能够充分挖掘读者历史行为中隐含的内容和规律，将读者不感兴趣的图书补充到兴趣模型中，有助于提高图书推荐的准确度。而且，在推荐结果中去除读者不感兴趣的热门图书，能够避免推荐的图书趋于热门书，有助于发掘长尾图书。

5. 第五组实验

综合本文在改进 UserCF 算法时使用到的所有方法，与 UserCF 算法对比，从准确率、召回率等几方面评价算法。

表 4.5 CTDN-UserCF 算法和 UserCF 算法实验结果对比

N 值	平均绝对误差		准确率		召回率	
	CTDN -UserCF	UserCF	CTDN -UserCF	UserCF	CTDN -UserCF	UserCF
10	0.798	0.851	56.34	40.29	6.02	3.26
20	0.779	0.835	53.26	38.12	8.83	5.51
30	0.765	0.826	49.86	35.38	9.96	6.51
40	0.754	0.819	46.02	33.56	10.98	7.6
50	0.749	0.813	43.78	32.57	12.07	8.58

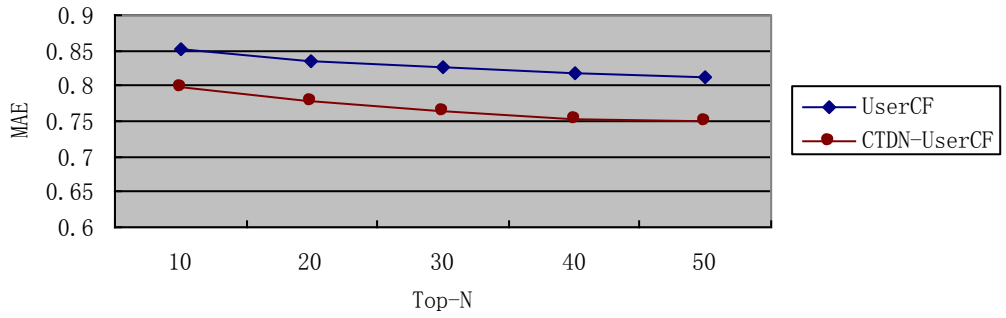

图 4.13 CTDN-UserCF 算法和 UserCF 算法实验 MAE 对比

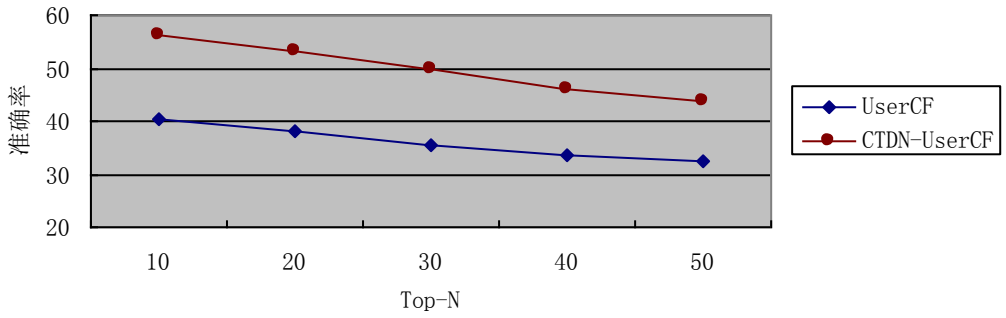

图 4.14 CTDN-UserCF 算法和 UserCF 算法实验准确率对比

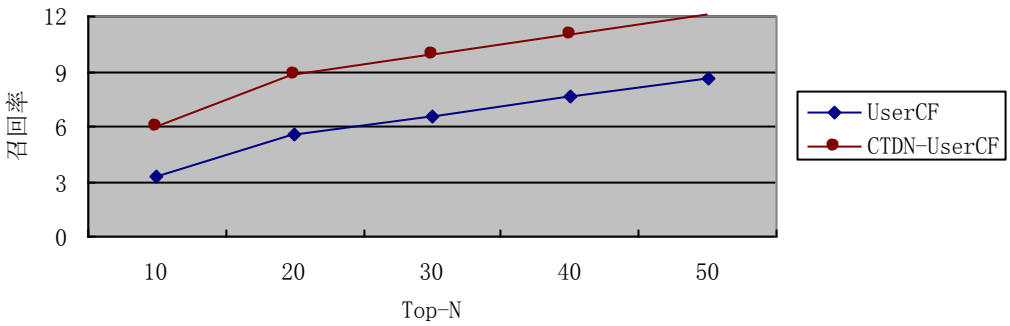

图 4.15 CTDN-UserCF 算法和 UserCF 算法实验召回率对比

从上表可看出，由于综合考虑了降低矩阵稀疏性、图书借出时间先后、图书借期长短，读者负反馈行为等方面的因素，推荐的效果比起单一考虑其中某个因素有显著的提升。

4.5 本章小结

本章针对第三章提出的自适应个性化图书推荐算法设计了五组实验，从平均绝对偏差、准确率和召回率三个方面对算法进行了验证，并对实验结果进行分析，结果表明该算法取得了较好的推荐效果。

结论

图书馆信息系统中有着丰富的图书，如何将这些大量的图书有效地、贴切地推荐及传达给读者，进而吸引读者到馆借阅及利用，以提升图书馆的利用率及效益，是图书馆管理者必须思考的问题。读者依其需求及兴趣到图书馆借阅图书，从这些被借阅过的图书中，可反映出读者对图书的偏好倾向及图书之间的关联性，若能从读者的借阅数据中找出图书彼此间的关联性，对图书馆管理者在拟订图书最佳的推荐策略时，必可提供相当有用的参考信息。

本文将基于用户的协同过滤推荐技术与用户行为特征处理技术相结合，应用于公共图书馆的图书推荐，通过对图书馆读者图书借阅行为的分析研究，在不同角度不同层面上，提取隐含信息的价值，构建精确的预测模型，自动追踪发现某位读者的需求改变，即使那位读者从未明确提及，以此为依据实现自适应个性化图书推荐。具体结论如下：

（1）采用《中图法》的图书分类标准，用图书分类号关联馆藏图书，描述读者兴趣，构建基于图书分类号的“读者—图书类型”访问矩阵，可有效解决基于图书资源的“读者—图书”访问矩阵由于读者借阅图书数量相对于图书馆馆藏数量过少，且很多读者相互之间没有对相同的图书产生借阅行为，导致的矩阵稀疏性过大的问题，以及馆藏图书数量和读者数量逐年递增造成的可扩展性问题。

（2）借鉴时间上下文感知推荐技术，结合行为特征处理技术，挖掘图书馆自动化管理系统数据库记录所隐藏的读者正反馈行为中时间信息与读者兴趣变化的相互关系，按读者借书行为发生时间的先后和借期长短对读者借阅图书的行为设置不同权重，对读者兴趣变化敏感，能够在读者兴趣模型中突出读者最近喜欢的图书，提高图书推荐的效果。解决了在协同过滤算法中将读者曾经借阅过的图书平等对待，没有考虑时间效应的问题。

（3）针对图书馆管理系统中记录的读者正反馈行为和负反馈行为进行读者行为特征处理，从读者正反馈行为中采集正样本，从读者负反馈行为中采集很热门但读者没有产生过借阅行为的图书作为负样本，补充完善读者兴趣模型，能够进一步解决矩阵稀疏性问题，还能够使读者兴趣模型反映出读者最真实的阅读倾向，有助于发现与目标读者相似度更高的读者群，提高个性化推荐的精准度。

本文的创新点是：结合上下文感知技术，将时间效应对图书推荐的影响融合到

个性化图书推荐算法当中，在读者借阅行为日志记录中，挖掘出图书借阅行为持续性长短与读者阅读兴趣预测的关联关系，设置图书借期的合理范围，只要借期在合理范围内即可视为有效借阅，而对于借期不在合理范围内的，降低这一类图书所对应的特征权重，削弱这类图书对图书推荐的影响；在进行读者行为分析和处理时，综合分析研究读者正反馈行为和负反馈行为，使读者兴趣模型能够全面综合地反映读者的兴趣偏好，而且在推荐结果中剔除了读者不感兴趣的图书，当中包括了读者不感兴趣的热门图书，能够在一定程度上缓解马太效应，以及进一步提高推荐的准确率和新颖度。

由于个人能力有限，本文的研究工作尚存在不足之处，日后还需从以下几方面加以改进：

（1）本文对于图书分类仅考虑《中图法》三级类目，对图书分类不够细致，粒度较粗，且仅从一个维度对图书分类，未能多维度描述图书属性，影响了挖掘结果的精确程度，下一阶段将针对此问题优化图书分类方法。

（2）本文在用户行为分析和处理方面，虽然有综合考虑到读者的隐性行为和负反馈行为，但挖掘的深度和广度还不够，下一阶段将围绕隐语义分析技术对读者行为特征作深层次挖掘。

（3）本文虽采用了时间上下文信息对协同过滤算法进行改进，但未考虑到特定时点与读者阅读偏好的关系。例如每年 5、6 月备考时期，或者每年固定的若干职业资格考试日之前一段时间，相应的图书会比较受欢迎。每年寒暑假，公共图书馆的青少年读者会增加，青少年读物借阅量也会随之而增加。以后可以沿着这一思路深度挖掘时间效应对自适应个性化图书推荐算法的影响。

（4）文本的工作主要体现在对基于用户的协同过滤算法的改进方面，但对于该算法存在的冷启动问题，尚且缺乏有效的解决方法。在今后的研究当中，可着重从如何给新注册读者做个性化推荐、如何将新采购的图书推荐给可能对它感兴趣的读者，这两方面入手探讨解决方案。

参考文献

[1] 国务院关于印发促进大数据发展行动纲要的通知.中华人民共和国国务院公报,2015(26)：26-35

[2] 图书馆中的长尾现象.http://blog.sina.com.cn/s/blog_492afa5001007rk5.html, 2007-12-20

[3] Resnick.P,VarianH.R. Recommender systems Communications of The ACM,1997,40(03):56-58

[4] Marko Balabanovic,Yoav Shoham. Fab:content-based collaborativere commendation Communications of the ACM,1997,40(03):66-72

[5] Chandrasekaran.K,Gauch.S,Lakkaraju.P,et al. Concept-based Document recommendations for citeseer authors. Lecture Notesin Computer Science,2008

[6] 王艳翠.Melvyl 推荐项目——发展中的图书馆推荐服务.图书馆杂志,2007(10)：55-57

[7] 张宁昳.Amazon 个性化推荐系统的文本组织结构研究.图书与情报. 2013(05)：103-106

[8] 李卫双.数字图书馆的个性化信息服务.教师,2010(02)：112-113

[9] 贾君枝,李婷.图书标签与书目记录结合方式.图书情报工作,2013(03)：96-99

[10] 杨博,赵鹏飞.推荐算法综述.山西大学学报（自然科学版）. 2011.34(3)：337-350

[11] Resnick, P., N. Iacovou, M. Suchak, et al. GroupLens: an open architecture for collaborative filtering of netnews. in Proceedings of the 1994 ACM conference on Computer supported cooperative work.1994. ACM. 175-186

[12] John S Breese,David Heckerman,Carl Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering.1998.43-52

[13] Badrul M. Sarwar, George Karypis, Joseph A. Konstan, et al. Item-based collaborative filtering recommendation algorithms. Uncertainty in Artificial Intelligence • UAI，1998,43-52

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

[15] 张新香,刘腾红.利用云模型改进基于项目的协同过滤推荐算法.图书情报工作.

2009(01)：117-120

[16] 肖敏,熊前兴.基于项目语义相似度的协同过滤推荐算法. 武汉理工大学学报. 2009(03)：21-32

[17] 王金燕,刘亚军.基于标签的个性化推荐技术研究.计算机科学期刊，2012 (02)：25-30

[18] 吴幸良,涂风华.采用图模型的个性化标签推荐方法.计算机工程与应用，2015 (09)：142-146

[19] Marek Lipczak.Tag Recommendation for Folksonomies Oriented towards Individual Users.In Proceedings of ECML PKDD Discovery Challenge(RSDC08).84-95

[20] Panagiotis Symeonidis,Alexandros Nanopoulos,Yannis Manolopoulos.Tag recommendations based on tensor dimensionality reduction. Proceedings of the 2008 ACM conference onRecommender systems:43-50

[21] Ralf Krestel,Peter Fankhauser,Wolfgang Nejdl.Latent dirichlet allocation for tag recommendation.Proceedings of the 2008 ACM conference onRecommender systems:259-266

[22] Andreas Hotho,Robert Jaschke,Christoph Schmitz,Gerd Stumme.Folkrank:A ranking algorithm for folksonomies.(Proc.FGIR 2006,2006)

[23] Krulwich, B. Lifestyle finder: Intelligent user profiling using large-scale demographic data. AI magazine, 1997. 18(2): 37

[24] Das, M., S. Amer-Yahia, G. Das, et al. Mri: Meaningful interpretations of collaborative ratings.Proceedings of the VLDB Endowment, 2011.4(11)

[25] 项亮.动态推荐系统关键技术研究,2011,中国科学院自动化研究所

[26] Montaner, M., B. Lopez, and J.L. De La Rosa. A taxonomy of recommender agents on the Internet.Artificial intelligence review, 2003. 19(4): 285-330

[27] 刘建国,周涛,汪秉宏.个性化推荐系统的研究进展.自然科学进展，2009.19(1)：1-15

[28] Cleger-Tamayo, S., J.M. Fern}ndez-Luna, and J.F. Huete. Top- N news recommendations in digital newspapers. Knowledge-Based Systems, 2012. 27: 180-189

[29] 段准. 基于内容的自适应推荐系统研究：[上海交通大学硕士].上海：上海交

通大学，2015

[30] 卜起荣. 基于内容的图像检索与推荐技术研究：[西北大学博士].西安：西北大学，2015

[31] 姜书浩,薛福亮. 一种利用协同过滤预测和模糊相似性改进的基于内容的推荐方法. 现代图书情报技术，2014 (02)：41-47

[32] 孟祥武,刘树栋,张玉洁,等.社会化推荐系统研究.软件学报，2015 (06)：1356-1372

[33] 刘树栋,孟祥武.基于位置的社会化网络推荐系统.计算机学报，2015 (02)：322-336

[34] Jannach D, Geyer W, Dugan C, et al. RecSys 09: Workshop on recommender systems and the social Web. In: Proc. of the ACM RecSys 2009. New York: ACM Press,2009.421-422

[35] Freyne J, Anand SS. Recommender systems and the social Web. In: Proc. of the ACM RecSys 2011. New York: ACM Press, 2011.383-384

[36] Mobasher B, Jannach D, Geyer W, et al. Recommender systems and the social Web. In: Proc. of the ACM RecSys 2013, ACM Press, 2013. 477-478

[37] Adomavicius G,Tuzhilin A. Context-Aware Recommender Systems.Recsys'
08，2008，335-336

[38] Ding Y, Li X. Time weight collaborative filtering. Acrn International Conference on Information&Knowledge Management,2005,1:485-492

[39] Lee H, Smeaton A F, O'connor N E, et al. User Evaluation of FischlarNews: An Automatic Broadcast News Delivery System.//ACM Transactions on Information Systems (TOIS.2008:M 102-M 107(6)

[40] Park M, Hong J, Cho S. Ubiquitous Intelligence and Computing. Springer Berlin Heidelberg, 2007:1130-1139

[41] G Adomavicius,A Tuzhilin. Context-Aware Recommender Systems.ACM Conference on Recommender Systems, 2008, 16(3):2175-2178

[42] Adomavicius G, Ricci F. RecSys'09 Workshop3: Workshop on context-aware recommender systems (CARS 2009). In: Proc. of the RecSys 2009. New York: ACM Press, 2009. 423-424

[43] 刘颖.基于二重分解的上下文预过滤推荐技术研究：[杭州电子科技大学硕士].

杭州：杭州电子科技大学，2015

[44] 秦大路. 基于因式分解机模型的上下文感知推荐系统研究：[郑州大学硕士]. 郑州：郑州大学，2015

[45] 朱煦,陈志奎,凌若川.一种移动上下文感知的好友推荐方法.小型微型计算机系统，2015 (04)：744-748

[46] 王世伟.未来图书馆的新模式——智慧图书馆.图书馆建设，2011 (12)：1-5

[47] 陈卓辉.当代图书馆的个性化信息推荐.科教文汇，2015(01)：195-197

[48] 曾子明,陈贝贝.融合情境的智慧图书馆个性化服务研究. 图书馆论坛，2016 (02)：57-63

[49] 肖理钏.基于云计算模式的图书文献个性化推荐技术研究.科技广场，2015 (08)：22-27

[50] 马晓亭.大数据时代图书馆个性化智慧服务 QOS 保障研究.现代情报，2014 (12)：69-73

[51] 郭素君.高校智慧图书馆信息服务系统设计与实现：[河北农业大学硕士]. 保定：河北农业大学，2015

[52] 罗文.协同过滤推荐算法综述.科技传播，2015 (07)：115-196

[53] Pazzani, M.J. and D. Billsus. Content-based recommendation systems, in The adaptive web2007 Springer. 325-341

[54] Weiser. Mark. The computer for the 21st Century. Scientific American. 1991. 265(3): pp. 66-75.September 1991

[55] Adomavicius G, Tuzhilin A. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge&Data Engineering, 2005, 17(6):734-749

[56] Dey AK. Understanding and using context. Personal and Ubiquitous Computing 2001，5(1):4-7

[57] Schilit B, Adams N, Want R. Context-aware computing applications. //IEEE Workshop on Mobile Computing Systems&Applications. IEEE, 1994:85-90

[58] 李慧, 马小平, 胡云,等. 融合主题与语言模型的个性化标签推荐方法研究. 计算机科学，2015 (08)：70-74

[59] 琚春华,鲍福光,刘中军.基于社会化评分和标签的个性化推荐方法.情报学报，

2014 (12)

[60] Ido Guy , David Carmel. Introduction to SocialRecommendation, in Social recommender systems, Proceedings of the 20thinternational conference companion on World wide web, March 28-April 01, 2011, Hyderabad, India:1355-1356

[61] 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.计算机学报，2014 (01)：219-228

[62] 项亮.推荐系统实践.北京:人民邮电出版社,2012

[63] 王巧容,赵海燕,曹健.个性化服务中的用户建模技术.小型微型计算机系统，2011 (01)：39-46

[64] 陈永光.基于 OPAC 的高校图书馆个性化图书推荐算法研究：[南京理工大学硕士].南京：南京理工大学，2013

[65] Song, Y., Z. Zhuang, H. Li, et al. Real-time automatic tag recommendation. in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. 2008. ACM. 515-522

[66] Sen, S., J. Vig, and J. Riedl. Tagommenders: connecting users to items through tags. in Proceedings of the 18th international conference on World wide web. 2009. ACM. 671-680

[67] Zhang, Z.-K., C. Liu, Y.-C. Zhang, et al. Solving the cold-start problem in recommender systems with social tags. EPL (Europhysics Letters), 2010. 92(2): 28002

[68] 曾子明.电子商务推荐系统与智能谈判技术.武汉:武汉大学出版社,2008

[69] 王伟,王洪伟,孟园.协同过滤推荐算法研究:考虑在线评论情感倾向.系统工程理论与实践，2014 (12)：3239-3249

致　谢

在论文即将完成之际，回顾过往，这些年确实不易，我的内心充满了感激，我由衷地感谢一直以来帮助我、支持我的老师、同学和家人。

首先，我要感谢我的导师蒋斌教授，是他将我带进了个性化推荐算法这一研究领域，让我了解到推荐算法在当今时代应用之广泛，领略到推荐算法的独特魅力。蒋教授学识渊博、年轻随和，在繁忙的公务之余还为我论文的研究与写作予以指导和帮助。在此期间，从论文的选题、研究思路的确定，到论文的组织、攒写和修改，每一个环节，都离不开蒋教授的悉心指导。在蒋教授的教导下，我学到的所有知识，将受益终身，对我的工作也会有极大的帮助。

我还要感谢我的另一位导师肖文俊，他长期以来在学习上为我提供了大量的帮助，在论文研究上也提供了大量宝贵的建议，使我的论文得以顺利完成。

其次，我要感谢和我一起攻读硕士学位的同学们。在求学路上，他们给予了我各种帮助。我们在学习中共同成长，不断进步，愿同窗友谊之树长青。

最后，我要感谢我的家人。在论文写作期间，他们给予了我莫大的支持和鼓励，为我分担了很多。他们的关心、陪伴和帮助，让我克服了学习、工作和生活中遇到的种种困难。他们是我困难中前行的坚强后盾，也是我心中最大的港湾。我还要感谢我的外公和外婆，感谢他们对我从小无微不至的关怀和爱护，以及悉心的教导，他们的音容笑貌永远留在我的心中。