

《数据管理技术与方法》

期末大作业

**题目：**基于数据挖掘的图书馆书目推荐服务的研究

学生姓名**： 王婧**

学 号**： 1831050171**

专 业**： 软件工程**

班 级**： 1802**

完成日期**： 2021年 12月 30日**

目录

[内 容 摘 要 I](#_Toc3068)

[序 言 2](#_Toc2140)

[一、 绪论 2](#_Toc20575)

[(一)论文背景 2](#_Toc28061)

[(二)论文意义 3](#_Toc28675)

[二、 图书馆信息推荐系统理论 3](#_Toc27812)

[(一) 图书馆信息推荐系统 3](#_Toc29879)

[1. 人机界面技术 4](#_Toc1614)

[2. 信息挖掘技术 4](#_Toc10501)

[3. 信息推送技术 4](#_Toc31637)

[(二) 推荐系统分类 4](#_Toc20538)

[1. 基于内容的推荐 4](#_Toc14338)

[2. 基于关联规则的推荐技术 4](#_Toc12149)

[3. 基于知识的推荐技术 4](#_Toc23788)

[4. 基于用户统计信息的推荐技术 5](#_Toc30850)

[三、 数据挖掘的理论 5](#_Toc23419)

[(一) 数据挖掘概念 5](#_Toc27731)

[(二) 数据挖掘功能 5](#_Toc15284)

[1.数据说明 5](#_Toc31300)

[2.数据分类 5](#_Toc24877)

[3.数据聚类 5](#_Toc2712)

[4.相关分析 5](#_Toc25109)

四、数据挖掘流程 5

（一）数据挖掘的流程 6

（二）决策树分析 6

（三）建立模型 6

[五、 程序设计 7](#_Toc21971)

[(一) 系统功能模块 8](#_Toc27565)

[(二) 主要流程 8](#_Toc30676)

[六、 总结 9](#_Toc18285)

[参 考 文 献 9](#_Toc9445)

基于数据挖掘的图书馆书目推荐服务的研究

内 容 摘 要

随着互联网信息的蓬勃发展,它给人们的生活方法造成了极大的负面影响。丰富快捷的资讯获得方法也带来了一次全球性的资讯革命。在此背景下,很多商业网站都构建了基于数据挖掘技术的商品个性化推荐体系,给人们带来了更直接、更高效的咨询服务。但是,到目前为止,推荐服务在商业图书馆的实际使用中还未能受到充分的关注。  
 本文将数据挖掘技术导入书库体系,并致力于提升图书馆对书目服务的推荐服务水平。本文首先对比了中国国内图书推荐系统的研发状况,提出了中国图书馆为信息推广业务所划分的若干领域及其需要哪些技术,接着阐述了最常用的推荐技术,并在对比其优缺点的基础上选取了最合适的推荐技术。然后又详尽地阐述了中国图书推广所需要的数据发掘办法:聚合数据分析、关联规则分析和决策树分析,并为各种数据挖掘方法选择最合适的算法。在引入关联规则分析方法时,通过优化Apriori算法,并引进了矩阵思想,把基于业务数据库的字符串运算转化为基于矩阵的布尔操作。这些改变不但降低了对数据库的访问次数,同时释放了大量内存空间,也大大提高了算法的执行效果。  
本文通过分析推荐书目的主要功能,在几种传统书籍推介方式的基石上,试图把各种数据发掘中的相关规律融入到图书馆书目推介工作。

本文将根据借阅实际,利用关联原理和Apriori算法对中国读者的借阅记录进行大数据分析,为中国读者提出个体化的书籍推荐服务。本文还认为,这些书籍推介方案将对我国图书馆和中国读者都有着意义。

关键词：数据挖掘 图书馆 书目 决策树 个性化信息服务

序 言

数据挖掘方法为人们深入分析并发现其在大量复杂信息中的存在规律,提供了高效的技术支持。目前,数据挖掘技术已经应用到商业网站中。通过数据挖掘技术网站建立商品推荐服务系统,将需要的商家的产品信息推荐给买家,从被动搜索到主动推荐,彻底颠覆传统方式,扭转被动局面。高职图书馆作为培养职业素质教育的重要机构,是教师和学生的信息来源和学习资源的重要途径。近年来,面对职教图书馆馆藏资源的不断扩大,读者受到了强烈的吸引。它需要一种方便快捷的获取信息的方式,这就要求职业教育图书馆扩大服务范围。服务空间大大提升了服务水平,将传统图书馆的被动服务模式转化为主动服务。因此,基于数据挖掘技术的个性化推荐服务系统将成为高职图书馆服务模式的主流。在比较系统分析了国内大学图书馆个性化信息服务体系发展状况的基础上,并结合了对数据挖掘的定义、任务、工作方式与流程等的系统学习,在本文中研究了图书馆分类,从而构建了个性化的大学图书馆信息管理服务体系。通过对传统图书馆信息服务模式的解析,得出我国职业技术学校图书馆所采用的数据挖掘与分析功能还不完善的结论。本文在现有图书馆体系的基础上,结合基于数据挖掘技术的个性化信息服务体系,对原有模型的组成结构、功能和工作模式进行了修改。以高职院校图书馆历史借阅记录为数据挖掘对象,采用聚类分析挖掘技术和关联规则来分析数据库中的隐藏规则。

1. 绪论
2. 论文背景

由于网络信息技术的发达,也给人类的生活方法造成了巨大的影响。丰厚便利的获取信息方法,引起了全球信息革命。在这个大历史背景下,很多商业网站都形成了基于数据挖掘技术的商品个性化服务推广体系,给人类带来了更直观、更高效的信息咨询服务。但是,到目前为止,信息推荐业务在商业图书馆的使用中,还未能受到充分的关注。二十世纪末,网络在世界各地迅速产生和发展,对人类的生活方法也产生了重大的影响。丰厚便利的信息获得方法引起了世界信息技术革命斗争。另一方面,互联网大大便利了人类的日常生活,使人类可以打破时代与地理条件的局限,更全面、更有效地掌握生活所需要的信息资料,并了解世界最新的资讯动向;但是,因为没有对网络信息资料的合理组织与分类,人民在信息检索过程中必须面临大批无序的信息垃圾,并花费着巨大的时光与精神。在这个环境下,人类迫切需要一个更直观、更智能、更人性化的资讯收集方法。

依托于数据挖掘技术在信息数据分析、分类与知识发现方面的巨大性能,人们开始运用它来进一步发现信息搜寻活动的规律性,并找到共同点,从而利用建立主动的信息传输服务为用户进行信息搜寻行为。个性化推荐系统。目前,个性化介绍模块已应用在多个商务网站中,如土豆网的"猜猜你喜欢什么"、淘宝网的"掌柜热卖"、京东商城的"同价"介绍服务、,等利用数据挖掘技术进行记录,并与数据库中记录的数据加以对比,为用户介绍同类的产品与活动。相反,大多数图书馆服务并不涉及个性化的介绍业务。图书室成为人们掌握知识、培养技术、积累知识的主要场所,给人类创造了巨大的信息资源。想要查找他们需要的图书,用户必须查阅数百万书籍,这一点非常重要。不简单的,特别是对各种书目的添加与更新,已经超越了人类的基本认知范畴。通过目前的传统搜索引擎根本无法得到所需要的书籍资料。于是,中国国内图书馆部门纷纷尝试研究书籍的推荐服务。

本文将数据挖掘技术导入书库体系,并致力于提升图书馆对书目服务的推荐服务水平。本文首先对比了中国国内图书信息推荐系统的研发状况,提出了中国图书馆对信息推荐业务所划分的若干方面及其需要哪些技术,同时推荐了常见的信息推荐技术,并在对比其优缺点的基础上选出了最合适的推荐技术。接着又详尽阐述了中国图书推荐系统所需要的数据挖掘处理方法:聚合数据分析、关联规则分析和决策树分析,并针对各种数据挖掘方法选择了最合适的算法。在介绍关联规则分析时,对Apriori算法做出了优化,并引进了矩阵的思维,把基于业务数据库的字符串计算转换为基于矩阵的布尔运算。这些改变不但降低了对数据库的访问次数,同时释放了大量内存空间,也大大提高了计算的执行效果。

(二)论文意义

1.有利于打通本馆和图书查阅者间的搜索障碍,发挥本馆的重要功能。

书库中保存了大量的信息数据,是信息资料的大集合。它的功能也不容忽视。但是,因为信息资料利用得不够,大量的图书全年都陈列在书架上。很多人因为无法进行学术研究和资料采集,就无法正确地查找自身所需要的图书。而信息检索制度的滞后直接造成了图书馆资源的闲置,也削弱了图书馆在人民日常生活中的服务功能。对图书馆书目所提供服务的深入研究,则有利于图书馆更好地为民众生活服务,发挥图书馆的公共服务能力,从而提升了图书馆的社会影响力。  
 2.改变了图书馆目前的被动落后的业务模式,使其更加适应了信息时代的发展。  
 原来的"我给,你要"的卖方市场将完全转变为"你需要,我提供"的买方市场,将使用体验变成了行业的领导者。目前,图书馆的图书搜索方式都较为简单。因为使用者只有对自身需求的图书有了清晰的了解,才能实现为自身定位书籍的目的,这就对使用者提供了很高的需求。7.设计和完善图书馆书目检索服务系统,是信息时代发展的大趋势。

1. 反映了该馆的人文关怀,并遵循该馆"以客为本"的基本管理原则与服务宗旨。

图书馆用户到图书馆工作并非只是为了寻书娱乐,也并非为了寻找书本学习。他的真正需要是个性化和深入的服务。图书推介业务的开展有助于增强用户的社会主体地位,让图书馆管理人员在满足读者借阅目的的同时,也体会到图书馆对读者的人文关爱。对图书推介业务的深入研究,将影响并引导未来图书馆的整体经营发展趋势。

1. 图书馆信息推荐系统理论
2. 图书馆信息推荐系统
3. 人机界面技术

人机界面,即用户界面,是用户查询和借阅图书馆图书的媒介。他的探索必须遵守如下准则:第一,使用户在使用时不产生不适,并满足要求,及时征求用户意见。最好符合所有用户需要的页面,让用户挑选自己感兴趣的个人页面;第二,各板块的功能与作用都应清楚,并能实现相应的操作;第三,注意界面之间的顺序关系,使用户在切换界面时不会出错;四、监控接口;五、界面的颜色、字体大小等应符合惯例,色调也应与图书馆的氛围一致,界面的配色应合理。  
 2. 信息挖掘技术  
 信息资源挖掘技能和传统信息挖掘技能也不一致。此处的信息挖掘,是指通过一定的关键字从信息系统中查找并提取资料的流程。图书馆信息挖掘技术主要是根据应用系统的运行记录,通过发掘用户的使用习惯,完善信息系统的模板设定以及页面组合样式,可以丰富使用者在注册后的个人页面。  
 3. 信息推送技术  
 信息推荐是图书馆推荐业务的核心内容,而信息推荐技术又是推荐业务中至关重要的科技基础。目前,信息推荐主要应用于商业网站,而其中最受欢迎的就是广告推荐。它响应及时,对象的覆盖范围很广阔。不论是分类组的新书介绍或是对个人用户感兴趣的图书介绍,都是至关重要的。

1. 推荐系统分类
2. 基于内容的推荐

协作筛选科技对使用者自己的打分重视程度不足,使用者也对自身特征的打分重视程度不足。所以,当使用者对自身特征的影响因素很多,甚至出现了影响使用者客观打分的原因时,协作筛选科技的选择结论偏差很大。为克服这一实际提问,在协同过滤的科学技术基石上进一步开发了根据信息内容的推介科技。根据信息内容的推介科技不再取决于对使用者的评价,而是依赖于使用者的历史行为等信息来建立使用者偏好文件。使用机器学习方法,通过对文件记录信息具体内容和产品之间的匹配程度进行匹配,从而向使用者提供相似程度最大的商品。  
 2.基于关联规则的推荐技术  
 关联规则推荐技术的关键是找出与用户行为的相关模式。它一般用来寻找在批量成交数据与购买记录之间的规律。而基于关联规则的技术推广是了解规律(购买的产品)并向用户推荐最后一项规则(可能感兴趣的产品)的步骤。该技术中最关键的部分就是关联规则发掘过程。最初的关联规则发生于调查超市客户交易记录的项目集中。此后,大量研究人员对关联规则发掘理论做出了补充与完善,进一步优化了原有算法,引进了并行性、随机性等新数学思想,大大提高了数据挖掘效果,并扩大了算法的实际应用。  
 3.基于知识的推荐技术  
 系统利用自身的知识信息,并通过所采集到的用户信息,自动形成与应用和产品之间的关联,进而提供服务给使用者,满足用户的需求。它是一个推理技术,并不要求使用人特征、喜好或者其他信息上的帮助,也没有产生用户分数。所以,在用户的偏好发生变化时,系统并不要求修改推荐标准。而只要向使用者提供需求的含义以及信息描述等基本信息,系统就可以向用户推荐。  
 4.基于用户统计信息的推荐技术  
 类似于协同过滤推荐技术,按照用户的喜好、性别、年龄段以及个人特点对用户进行了划分,然后根据分类结果推荐相似用户。不过,根据使用者统计的推介科技和协同过滤的区别在于,协同过滤科技主要是根据记录收集使用者的个性特点数据,而根据使用者统计的推介科技,顾名思义,也就是根据使用者的个性特点数据收集的。实时用户统计与调查,通常使用人机对话的方法。

1. 数据挖掘的理论
2. 数据挖掘概念

数据挖掘的概念从大部分统计不完整、有噪音、模糊和随意改变的资料中获得隐蔽的、意义不详,但却可能有益的消息和经验的过程。从概念上可以发现,数据挖掘通常具备如下特征:1、数据挖掘的主要对象是大量的原始资料,由于这种对原始数据存在着干扰性,因此必须在数据挖掘之前加以整理;数据挖掘的主要功能是发现,而不是证明。因为在数据挖掘以前,人类并不了解数据挖掘将会产生哪些成果;3.数据挖掘的结论实用,容易掌握和接受。这些信息都隐匿在原始数据中,且从未被发现。

1. 数据挖掘功能
2. 数据说明

数据挖掘的数据描述功能等同于传统的数据分析方法。它主要描述数据特征,如总和、均值、方差和标准差。此外,它还具有生成点图、直方图、直方图、饼图等功能。  
 2.数据分类

通过数据挖掘进行数据分类主要是编写一个分类函数,并在数据库中设置各种标准。每个标准对应一个类型,该分类函数也称为分类器。它不能直接作用于数据库中的记录,而是作用于数据库中的属性。  
 3.数据聚类

数据聚类和数据分类的功能相似。他们都对数据进行了分类,但他们的方法和想法不同。在数据聚类过程中,不同类别之间的分类标准不是预先确定的。按"类间差异明显,类内相似"的原则进行分类。它是动态的和最优的。在数据聚类之前,我不知道数据库中的记录将被划分为多少个类别。

1. 相关分析

为发现与数据库中记录的关联,人们一般采用关联规则和序列模型。前者用来寻找在同一个事件中各种记录之间的关联,而后者用来寻找在各个事件之间和时间有关的行。

1. 数据挖掘流程
2. 数据挖掘的流程

图书馆推荐系统有三个部分：数据库、数据挖掘和用户界面层；

1.数据库层

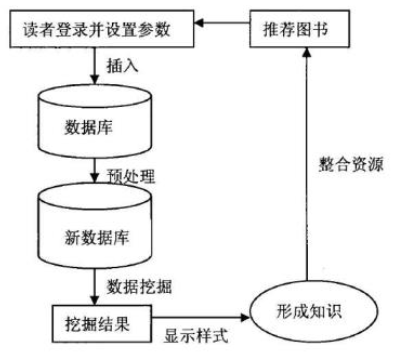
通过图书馆数据库中的借阅信息、用户信息等数据，将其转换成系统可以收入的形式，收纳到数据库中，传递给数据挖掘层。

2.数据挖掘层

在数据挖掘层通过代码可以转换成实际可以运行的程序，用从数据库传出来的数据，实施数据挖掘的过程，产生一定的规律，来进行图书的推荐，并将其最终结果推荐给用户界面层。

3.用户界面层

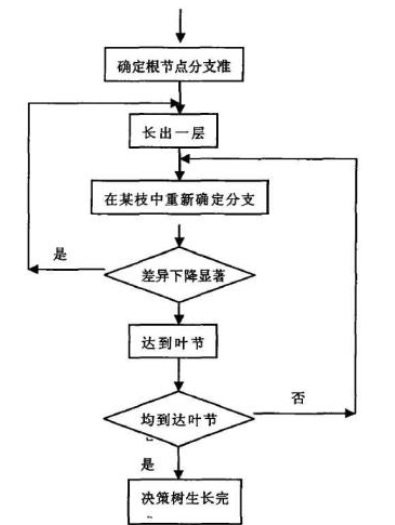
用户可以通过这一层对系统进行操作。



1. 决策树分析

1.决策树生长

将样本分组，按照主要算法确定决策树上的分支，直到细化到一定的程度，对决策树上的一个分支在分组没有意义的时候，这一段分支就不会再生长，直到决策树的所有分支都停止生长，决策树就结束生长，需要进行剪枝。



2.决策树修剪

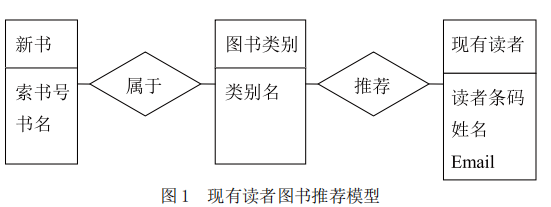
决策树在分组过程中，越细化包括的数据独特性就越强，总体的代表性就越差，这样决策树就失去了代表性，也就失去了对数据的预测能力，因此需要修剪决策树。

1. 建立模型

1.现有读者阅读兴趣

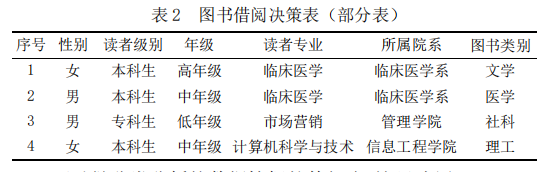


首先可以从图书借阅表中统计每种图书的读者数，当图书编写好索引号后就可以放入某一类图书中，该书本上架后就可以向喜欢这类图书的用户推荐该书。

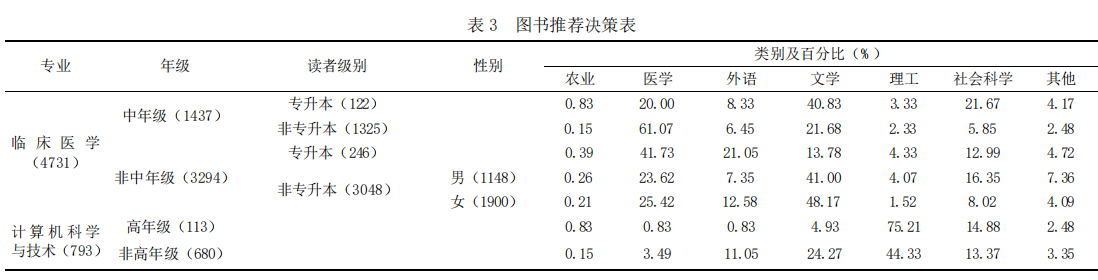


2.新读者阅读兴趣模型建立

对表中的图书借阅数据进行挖掘，得出分类规则，图书类别为决策类别属性，其他未条件属性。



对决策树分类算法进行分析



得出以下规则：

条件：专业 = 临床医学 并且 年级 = 中年级 并且 读者 级别 = 专升本

结论：借文学类书的占 40.83%，借医学类书的占 20%，借社 科类书的占 21.67%。

1. 程序设计

（一）系统功能模块

1.用户登录:读卡器可按卡号输入。初始密码是卡号的最后四位数字。成功登录后,个人信息将得到改善

2.新书推荐:近30天的新书推荐,方便读者更快地借阅新书;  
3.藏书楼:最具个性化的部分,读者可以在此收集自己喜爱的书籍和信息;此外,还保存了读者的历史借阅记录;  
4.大众推荐:向用户推荐借阅量大的书籍。

在该模块下,可分为学校流行推荐和专业流行推荐。学校最受欢迎的推荐方式是系统地向读者推荐高借阅率的书籍,按借阅量降序排列;大众专业推荐是根据读者的专业特点向读者推荐更高层次的书籍。这种推荐方法更智能,更贴近读者的需求。使用此模块可以减轻新读者的盲目性,不会导致新读者拒绝系统;

5.图书检索:是图书馆图书检索的检索部分。

使用书名、作者、书号、出版社、ISBN、图书内容等关键词进行搜索,使读者能够及时获取相关信息,查看图书相关信息。

（二）主要流程

在浏览器中输入用户名、密码和验证码后,成功登录图书推荐系统。成功登录后,如果用户是新用户,系统将根据书籍内容自动跳转到推荐。系统根据用户的基本信息、专业、性别等信息创建用户图书库,然后对读者的专业、年龄等基本信息进行聚类分析,找到与用户相似的用户,获取用户的浏览信息并进行推荐。用户生成借阅记录时,可以使用与借阅相关的推荐算法,也就是匹配与用户相关的图书进行推荐。

1. 总结

学校图书馆的发展反映了学校的文学遗产。在人生阶段的求知道路上,大学生对图书馆的需求

和依赖日益增加。图书馆已成为大学生获取知识的重要渠道。

在信息日益多样化的今天,传统的书本知识已经远远落后于时代,知识需求水平与现代高职学

生的差距越来越大。职业技术教育的最大特点是因材施教。通过掌握不同学生的特征,努力提高了学生的职业技能。作为校方给学生提供的教学平台,学生可以根据自己的兴趣在图书馆获得所需的知识。图书馆藏书种类繁多,积累了丰富的资料。然而,现有的图书馆管理系统只具备图书基本信息的收集、统计和查询功能,不能为读者提供高效的服务。结果,图书馆里的一些书籍几乎成了摆设,造成书籍的闲置和无形损耗。此外,随着知识时代的标志越来越明显,人们对信息的数量、质量、针对性和专业性的要求也越来越高。如何帮助读者从图书馆的海量图书中快速、高效地找到自己需要的图书,是图书馆不断改进服务的重点,也成为图书馆发展的方向。

参 考 文 献

[1]张泳涛.关联规则算法研究.硕士学位论文.南京:南京大学，2010.

[2]朱明.数据挖掘.第2版.合肥:中国科学技术大学出版社:2008:258-265.

[3]廖琴郝志峰陈志宏.数据挖掘与数学建模.北京:国防工业出版社，2010:186-199.

[4]方洪鹰.数据挖掘中数据预处理的方法研究.硕士学位论文.重庆:西南大学，2009.

[5]林郎碟,王灿辉.Apriori算法在图书推荐服务中的应用与研究[J].计算机技术与发展,2011,(21),22-24.

[6]路永和,曹利朝.基于关联规则总和评价的图书推荐模型[J],应用实践(2),81-86.