中文摘要

在当今的信息社会，日益剧增的数据量需要有效的工具来对这些数据进行建模和分析。数据挖掘就是从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的过程，它和数据仓库都是从这些海量数据中抽取知识的有效工具。近些年来，数据仓库与数据挖掘在各行业都得到了广泛的应用。

数据仓库的建立不仅需要各种建设工具，而且还需相应的数据支持。数据仓库的建设是各种先进的信息处理技术与企业管理决策结合的过程，只有将OLAP 技术、数据挖掘技术与数据仓库中的海量数据相结合，与企业管理决策方法相结合，才能真正的发挥各企业部门耗费巨资所搭建的数据集市的巨大作用。

随着数字化程度的发展，知识经济时代的到来，高校图书馆也日益成为经济社会的中心，他们肩负着信息/知识的整序、传播、开发和利用的重担，以及读者需求不断提高，要求每个馆员都是他们的知识导航员等等这些挑战都迫切要求高校图书馆必须了解并掌握读者的兴趣及需求、图书资源的使用情况等，能够主动的为读者提供他们所需要的服务，甚至能够主动的引导读者需求倾向。因此，本文选取高校图书馆系统作为数据仓库，利用数据挖掘技术针对高校图书馆的需求进行数据分析建模，试图为高校图书馆工作者、管理者、决策者提供一定的工作、管理、决策的科学参考依据。

本文介绍了我国高校图书馆的发展背景，分析了高校图书馆目前所面临的问题，提出利用数据挖掘技术来提高高校图书服务水平的思想。本文首先介绍了数据仓库和数据挖掘技术的理论知识，在此基础上总结了高校图书馆的数据特点及工作服务特点，从读者、图书、工作三个方面对数据挖掘技术在高校图书馆中的应用进行了理论分析。然后，文章介绍了高校图书馆系统数据仓库的创建，并对其中的重要数据作了探索性分析，从中得出高校图书馆利用的时间规律、图书类别规律、读者兴趣规律等。最后，利用决策树算法和聚类算法分别建立低利用率读者特性模型和读者细分模型，得出不同读者群体的特征，以期能对未来的工作进行预测作一定的参考。

关键字：数据仓库，数据挖掘，高校图书馆，知识导航员

Abstract

In this information society, huge data accumulated speedily everyday need effective tools to make models and analyze. Data mining is a process to fetch models which is efficiency, novel, latency useful and ultimately understandable. Data mining and data warehouse both are efficiency tools to take knowledge from those very large amount data. In lately years, they are both widely used in every kind trade.

Data warehouse' building needs not only all kinds of construction tools but also needs data support. Data warehouse' building is a process of combination all kinds advanced information dealing technologies with enterprise management decisions. If only to combine OLAP, data mining and the huge data in data warehouse with enterprise management decisions, can the data bazaar which costs huge play its greatness function.

With the development of the degree of digitalization and with the knowledge and economic times coming, university library is also becoming the center of economic society. It takes on the heavy load of keeping order, spread abroad, development and make use of information/knowledge. Reader's demand is keeping boost and every librarian should be their knowledge navigator, which those challenges press for that university library must find out and master readers' interests and demands, books resources' usage and so on. It should server readers for what they real need, and even can actively induct reader's tendency.

Therefore, this dissertation selects university library management system as data warehouse, using data mining to put up data analysis and model making aiming at university library demands. It tries to provide a certain scientific reference during working, management and making decision for workers, supervisor and decision maker.

This dissertation introduced the development backgrounds of university library in our country, analyzed the problems which university library being facing with and brought forward the idea of using data mining technology to improve the service level of university library. Firstly we introduced the theoretic knowledge of data warehouse and data mining, based on which we summarized the data and service

characteristics of university library. Then we theoretically analyzed the application of data mining on university library from three aspects which included reader, book and service. Secondly we introduced the data warehouse building of university library and gropingly analyzed its important data, out of which we deduced the rules of time using, book sorts and readers' interests for university library. Lastly we made two models; one is using the arithmetic of decision-making tree to set up readers' characteristic model of lower utilizing, the other is using the arithmetic of clustering to build the model of readers' subdivision, from which we educed the characters of different readers. We respect that it can act as a certain reference for future working and forecasting.

Key words: data warehouse, data mining, university library, knowledge navigator

学位论文独创性声明

本人郑重声明：

1、坚持以“求实、创新”的科学精神从事研究工作。

2、本论文是我个人在导师指导下进行的研究工作和取得的研究成果。

3、本论文中除引文外，所有实验、数据和有关材料均是真实的。

4、本论文中除引文和致谢的内容外，不包含其他人或其它机构已经发表或撰写过的研究成果。

5、其他同志对本研究所做的贡献均已在论文中作了声明并表示了谢意。

作者签名：张海燕

日　　期：2008.6.1

学位论文使用授权声明

本人完全了解南京信息工程大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版；有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅；有权将学位论文的内容编入有关数据库进行检索；有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

作者签名：张海燕

日　　期：2008.6.1

关于学位论文使用授权的说明

本人完全了解南京信息工程大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。

（保密的论文在解密后应遵循此规定）

作者签名：张海燕　　导师签名：

日　　期：2008.6.1　　日　　期：2008.6.1

第一章 引言

近些年来，尽管可以直接感受到各个领域中的变化，但还是很难发现哪个领域的变化能超越信息数量的增加。“信息爆炸”为各行各业提供了新的机遇，同时也引致了一些新的问题，从制造业到医药行业再到市场营销无不如此[1]。面对浩如烟海的信息知识，绝大多数人都会感到无所适从。因为知识的巨量，人们很容易被吓倒，这就是所谓的“人们被知识淹没却又饥饿于知识”。然而，“道高一尺，魔高一丈”，智慧的人类总是能够走出窘境，大量的数据分析与数据挖掘（Data Mining）的方法和策略都在被各行各业广泛的应用。

本文就是利用当今最受关注且最具广泛应用前景的数据挖掘技术来对高校图书馆这一行业进行理论分析和应用研究。

1．1 课题研究背景和意义

1．1．1 高校图书馆的发展

作为近年来对图书馆事业发展影响最大、同时也是最为活跃的领域——计算机技术和数字通讯网络的发展前景如何，成为图书馆界关心的热点。

我国图书馆界从开始使用计算机到今天有二十多年的时间，其间至少经历了两次大的变迁。第一次是微型计算机的出现，使图书馆计算机的应用得到普及；第二次是国际互连网的推广，它使图书馆由封闭服务走向开放。[2]

知识经济时代图书馆的地位和作用问题，也是当前图书馆界普遍关心的一个热点问题。农业经济时代，土地和简单劳动力是社会主要资源，图书馆游离于经济社会之外，其发挥的不过是“藏书楼”的作用；工业经济时代，资本和能源构成了主要资源，图书馆处于经济社会的边缘，发挥着“知识守护者”的作用，而随着知识经济时代的到来，信息和知识构成了主要资源，图书馆日益成为经济社会的中心[3]，发挥着对首要资源——信息、知识的整序、传播、开发和利用的作用，图书馆员应当充当着“知识领航员”的角色。

如今，图书馆的建筑设施从简单建筑楼到现代化设施齐全的大厦；图书馆的借阅方式从闭架的目录检索服务到开架借阅服务；再有数量倍增的图书馆藏及可供无限量借阅使用的电子资源等这些变化都预示着图书馆行业的一次次革新。伴随着这些革新的进程，图书馆的核心价值体系及价值观念也迫切的需要有新的定义。

大学图书馆业务是目前图书馆这一行最包罗万象的领域。大学图书馆主要服务该机构所属的学术社群，包括其中的大学生、研究生、教师、行政人员，以及职员。综合性大学图书馆应该要为研究生和教员提供资料来辅助他们的研究。现在的国内大学图书馆目前已经在不断尝试提升自己的个性化服务和导航作用，不断的缩小新时代读者需求与图书馆服务之间的差距。

目前国内图书馆提供的服务主要包括：纸制文献的查找借阅；电子资源的检索、利用；另外还有在不断发展完善的信息知识服务，主要包括馆际互借、科技查新、代检代查、学科服务等。这些都是等待读者上门的被动服务。另外有些图书馆提供的特色服务包括图书服务和读者服务两方面[4]。清华大学、南京大学等诸多大学图书馆都可以使用其管理系统通过 e-mail 为读者发送流通通知（包括过期通知、催还通知、预约取书通知）；中国人民大学提供图书导读、报刊导读、热点追踪、专题探微和图书外借预约排行榜服务；台湾交通大学图书馆定期举办图书阅读系列活动、专题音乐节等，同时设立阅读 365 知识点数累计，可以换取纪念品或者提升图书馆借阅权限来鼓励读者参与图书馆活动及利用图书馆馆藏。

国外一些发达国家的图书馆事业在很多方面都要领先于国内的发展水平。马塞诸塞大学杜博图书馆，能够在用户需要的时候，随时提供服务；能够针对自己的新馆藏提供符合读者时间表的全方位服务。罗格斯大学图书馆致力于寻找采购电子资源与传统资源如书籍等之间最适当的平衡。哈佛大学图书馆在其网站首页简明地列出其推荐的学术论文及其内容介绍。

国内很多图书馆也纷纷学习先进的服务理念和方式，但是由于专业人才的缺乏，以及其他种种条件的限制，大部分图书馆的服务水平和方式仍需要相当大的改进。所有措施的制定和实施都需要对图书馆自身利用情况以及服务的读者的特性进行科学的分析，才能制定出真正能够促进图书馆工作效率和提高其服务质量的科学措施，来主动拓展读者视野，培养读者综合素质，引起“懒惰”读者的阅读兴趣，促进读者的阅读行为以及扩大读者的阅读范围，真正地改变目前大部分图书馆的“坐等读者上门[5]”的借、阅、还的简单服务方式。

1．1．2　我国高校图书馆面临的问题

随着我国高校的发展和层次的提升以及办学类型的转变，高校图书馆也面临着非常大的挑战。学科的不断整合与发展、读者需求不断的变化与提高、信息技术的发展、科研能力的加强、教学教育的改革、和谐社会与和谐校园的建设等等，这一切都要求图书馆要不

断的改变，不断的提高，才能保证跟得上高校快速发展的步伐，甚至在某些方面能够引领高校的发展方向。

如何有效地开发利用信息资源，将成为高校图书馆提升自身作用和地位的关键；如何以现有的资源提供最优的服务，如何根据读者的需求对图书进行合理的布局和搭配，如何根据目前的图书馆利用情况来引导读者的未来需求等一系列问题都是高校图书馆应该思考并努力给出科学解答的行业问题。

面临着如此巨大的挑战以及伴随着高校规模的不断扩大，学科的整合所带来的一系列问题，高校图书馆应利用已有的、有价值的信息和知识积极面对和迎接挑战，实现经营理念创新、制度创新、技术创新[6]，通过对图书馆信息资源的充分利用，不断提高工作效率，提升服务质量，找准在教学科研中的定位等，以增强高校图书馆的功能性。

国内高校图书馆应用现代化信息技术刚处于起步阶段，对读者行为数据的收集和利用都还不够，对读者数据、图书数据中所包含的大量的潜在的、有用的信息利用还不够充分。加强高校图书馆工作的信息化水平，增加其工作、决策、管理的科技含量，是图书馆在信息时代发展的必由之路。

1．1．3 数据挖掘技术对改进高校图书馆服务水平的作用

当前，无论在学术界还是产业界，Data Mining 都是一个相当时髦和红火的专题。Data Mining 的中文翻译有：数据挖掘、数据淘金和数据采矿[7]。它被美国麻省理工学院（MIT）评价为未来影响最大、发展最有前途的十大技术之一，它名列第三。

数据挖掘广泛地应用于零售/营销、银行、保险、交通、电信、医疗及故障诊断等许多领域，在市场预测、股票分析、客户行为分析及决策支持等许多方面都取得了可喜的成果。它在互联网资源、信息检索、搜索引擎等方面也都起着非常重要的作用。数据挖掘技术可以用来支持广泛的商务智能应用，如顾客分析、定向营销、工作流管理、商店分布和欺诈检测等。数据挖掘还能帮助零售商回答一些重要的商务问题，如“谁是最有价值的顾客？”“什么产品可以交叉销售或提升销售？”“公司明年的收入前景如何？”等。

如今，在医学、科学与工程技术界、金融、保险等领域，数据的积累增长速度也是惊人的。NASA 轨道卫星上的地球观测系统 EOS 每小时会向地面发回 50GB 的图像数据[8]；美国零售商系统 Wal-Mart 每天会产生 2 亿左右的交易数据；我校图书馆每个月的借阅记录也高达数万条。面对海量数据库和大量繁杂信息，如何才能从中提取有价值的知识，进一步提高信息的利用率，如何从庞大的用户数据库中吸取有用的用户行为模式和各种潜在的

信息，为用户提供更好的服务，为决策者提供更好的决策支持，就是数据挖掘研究的意义所在。数据挖掘技术为我们从大规模的数据库中提取有用信息提供了强有力的解决工具。

目前，图书馆的系统数据主要包括：图书馆馆藏信息，读者信息，读者借阅历史记录等。我馆到目前为止，馆藏纸质文献 118 万册，每天的数量还在不断的增加，注册读者 2 万多人，每天读者的借还平均达 4000 次，阅览达 1500 余次，这样的数据每天都在增加，图书馆系统内已经藏有海量的数据，它们都需要昂贵的存储器来存储，但是它们除了说明一些历史记录，是否隐藏有大量的、具有更大价值的隐性信息？

我们知道，数据挖掘不但能够学习已有的知识，而且能够发现未知的知识；Bigus 将数据挖掘系统定义为“从大量数据集合中发现有价值但不明显的信息[9]”。数据挖掘得到的知识是显式的，既能为人所理解，又便于存储和应用，因此一出现就得到广泛的重视。

然而，因为数据挖掘在以上所说的各个领域的应用都能带来可观的经济效益，所以它能够迅速的得到认可和充分的发展，在高校图书馆的工作中，由于其自身不单纯追求经济利益的服务特性导致目前还没有成熟的应用，但是如果我们能够清楚的了解图书馆资源的利用规律和读者的使用特性，很大程度上就可以了解我校师生的需求以及学习科研等状况，也可以为管理部门和决策者提供制定政策和方法的科学依据。

综合来看，实现高校图书馆的信息化之路，大致要求国内高校图书馆重点要实现下列转变[10]：收藏由纸质转向复合图书、工作重点由文献收藏向文献保障转变、定位由文献中心向文献和信息中心转变、服务模式由被动服务向主动服务转变、服务内涵由文献服务向知识服务转变、工作中心由资源建设向学科服务转变、服务方式由面向面服务向网络化服务转变、服务教学方式由第二课堂向第一课堂转变、服务范围由服务校园向服务社会拓展、性质由服务机构向学术机构、培训教育机构转变、向信息产业部门转变。

本文希望通过数据挖掘技术，对我校图书馆系统的数据进行研究，能够改进图书馆的管理和服务理念，增强图书馆的特色及个性化服务，在图书馆的信息化转变的过程中，发挥基础作用；能够有效地缓解图书馆工作与师生的实际需求脱节的现象，使高校图书馆能够在今天的信息社会，随着社会文化、文明的进步，满足人们对图书馆提出的更高的要求，真正的成为科学文化和精神文明的殿堂。利用数据挖掘的有关技术，在高校图书馆系统中可以展开如下分析：

（1）读者社群与借阅行为模式分析

叙述统计：以叙述统计的方式对数据仓库的数据进行分析，找出借阅率高的图书和读者，并分析其代表的意义。

分类统计：利用借阅记录和读者信息库，分析读者的不同群体间借阅行为的差异，以

了解读者的行为模式。

孤立点分析：利用读者的借阅信息，找出特殊需要的读者群体，可以为其提供特殊服务；利用图书的借阅历史表，找出最受欢迎图书系列与最不受关注的图书系列。

（2）个性化服务工作

关联规则分析：利用读者借阅记录库，找出各类图书的相关性，可以向读者提供相关图书推荐服务。

时间序列分析：利用读者借阅记录库，找出读者借阅图书的顺序特性，可以向读者适时进行图书推荐，引导读者的进一步借阅；同时找出读者借阅的时间特性，可以为图书馆的工作安排提供科学依据。

1．2 目前国内外相关领域研究状况

1．2．1 目前数据挖掘国内外研究状况

数据挖掘的工作虽然是近年来数据库应用领域中相当热门和时髦的技术，但 Data Mining 使用的分析方法，如预测模型、数据库分割、连接分析、偏差侦测等，美国政府从第二次世界大战以前，就在人口普查以及军事方面使用过。

到了 20 世纪 80 年代末，最早是以从数据库中发现知识（Knowledge Discovery in Database, KDD）研究起步，KDD 一词首先出现在 1989 年人工智能国际会议上，1995 年，召开了第一届知识发现与数据挖掘国际会议，以后每年召开一届，规模由原来的专题讨论会发展到国际学术大会，研究重点也逐渐从发现方法转向系统应用，并且注重多种发现策略和技术的集成以及多种学科之间的相互渗透。

目前，除了 KDD 是数据挖掘国际会议，SIGMOD（Special Interest Group On Management Of Data），VLDB（Very Large Database），ICDE(International Conference on Data Engineering) 是数据库领域的三大顶级会议，研究涉及的范围较广，主要偏向应用，但是，数据库研究者会倾向于把数据挖掘看作一个数据库的子领域来研究。事实上，对于很多其他背景的人而言，数据挖掘是相对独立的一个新兴领域。

目前，国外数据挖掘发展趋势的研究方面主要有[11]：对知识发现方法的进一步发展，如近年来注重 Bayes 方法以及 Boosting 方法的研究和提高；传统的统计学回归法在 KDD 中的引用；KDD 与数据库的紧密结合。在应用方面包括：KDD 商业软件工具不断产生和完善，注重建立解决问题的整体系统，而不是孤立的过程。关于数据挖掘方面的应用研究作大概

的归纳如下。

购物篮分析，研究的是购物行为间关联的认识。类似的分析方法可用在主要目标是用交叉销售提高某个经济单位的商品的销售数量的问题上。可以将商品以最有效率的方式排列，将那些关联最大的商品放在一起以达到目的。

Web 点击流分析，是关于识别网站浏览模式的研究。在离散时序数据识别行为模型的问题中，可以用到相似的分析，如连续的商品交易或者研究人的职业行为。

网站用户分析，是关于根据网页行为进行用户分析的。分析的目的是将用户分类到未知数量的同质类中，这些类可以根据统计特征进行解释，分类是无指导的。

客户关系管理（CRM），研究的主要目的是提高客户的忠诚度，以从客户那里尽可能的盈利。CRM 应用非常广泛，通常可以处理任何公司客户数据库，将客户分类到不同的类，为将来公司的活动区分不同的目标。它的主要兴趣是客户分类。

信用评分，适用任何个人或公司对过去的行为进行评分的场合，目的是针对同样的个人或公司在 CRM 框架内计划将来的活动。分析的目的是建立评分规则，给每个顾客一个数值分数。

电视观众预测分析，目的是建立一条预测规则以使电视网尽量赢得最多的观众。它也可以应用于任何目的为预测总的个人偏好的情形。

我国从事数据挖掘的研究起步较晚，大约在 90 年代中期，初步形成了知识发现和数据挖掘的基本框架。近年来许多高校、科研院所在这一领域内开展研究，一般集中于学习算法的研究、数据挖掘的实际应用以及有关数据挖掘理论方面的研究。总体来说研究重点也正在从发现方法转向系统应用[12]，并且注重多种发现策略和技术的集成以及多种学科之间的相互渗透。目前进行的大多数研究项目是由政府资助进行的，但是基本上还是以学术研究为主，实际应用上仍然处于起步阶段。

1．2．2 图书馆系统数据挖掘的研究现状及应用前景

目前，个性化服务在电子商务领域的成功经验使得图书馆也开始研究以用户为中心的个性化信息服务。目前，图书馆的个性化服务主要有两种方式：

（1） 个性化信息定制。是指用户可以根据自己的需要来进行定制服务，图书馆根据用户的需求来提供特定的信息。

例如，康纳尔大学图书馆的 My Library Cornell 系统通过 My Links 和 My Update，由用户定制自己的需求，满足个性化服务的需求；北卡罗莱纳州立大学的 My Library 通过

服务定制，使用户可以得到图书馆的特殊服务[13]；清华大学图书馆也有相似的功能。

（2） 基于电子邮件的个性化信息服务方式。系统根据用户订阅的情况提供相应的栏目内容，定期或不定期的发送到其个人邮箱。目前国内一些高校图书馆已经开始尝试提供这种服务，根据用户的专业、研究方向，用电子邮件发送一些最新的资源和服务动态。如中科院文献情报中心的联合西文期刊篇名定题服务、中国科学院上海文献情报中心的新书信息推送和西文现期推送服务等。

国内对于数据挖掘在图书馆系统中的应用研究主要集中在高校的科研人员中。西南大学陈文文进行过图书馆使用者行为模式的数据挖掘研究，对读者社群关系进行了一定的探索，在对读者进行分类分析上给出较详细的分析；东南大学的周蓓重点进行了关联规则算法的改进研究，并开发了对应的 web 环境产品；还有很多研究人员重点对数字图书馆方面的数据挖掘应用进行了理论研究。总体上来说，对于图书馆而言，数据挖掘技术在图书馆系统中的应用还不广泛。但是，图书馆的信息服务模式与市场营销模式有很多的相似之处，通过收集、加工和处理设计读者行为的大量信息，确定特定的借阅群体或个体的兴趣、借阅习惯、倾向和需求，可以推断出其未来的借阅行为，可以在很大程度上提高图书馆服务的主动性及质量。

随着计算机技术和网络技术在大学图书馆的普及和应用，图书馆无论在服务内容、服务方式还是服务手段上都较以前有了很大的飞跃和提高。但是，图书馆的管理与服务中仍存在着与时代需求不相适应的地方。主要体现在：

① 服务重点总体上还停留在文献借阅的层次，在大学的教学与科研上没有能够起到向导的作用，对读者学习能力和素质的培养支撑不够，对高校培养有创新能力的高层次人才不利。

② 与学校的科研组织缺乏强有力的联系和沟通，从而对学校的总体科研活动支持力度不够。

③ 管理和服务相对封闭和僵化，造成与社会和公众的隔绝。

以上问题的根源在于很多大学的图书馆在新的时代、新的技术基础下工作重点依然是以借阅服务为主，解决问题的关键就是要把图书馆的工作中心从借阅服务转向到参考咨询服务上来。加强图书馆与各科研组织、读者、社会的联系，变目前的被动服务为主动服务，在大学发展的进程中能够发挥导航与风向标的作用。

那么，要建立图书馆、科研部门、读者、社会的互动系统，以及要求图书馆主动向各个部门和个体提供服务，就要求图书馆了解外部需求，掌握外部特性，才能真正的做到有的放矢。目前国内只有极少数的几个图书馆或情报中心能够提供以上所提到的服务，全国

众多的高校图书馆工作和服务仍然静如止水，可以说，在目前和未来几年，展开对读者行为范式的研究将是图书馆界与工业技术界需要来共同承担的重要课题。

1．3 论文的组织结构与研究方法

1．3．1 论文的主题思想

基于以上对高校图书馆的现状、面临的问题、未来的需求以及数据挖掘的意义、技术及其在图书馆领域的应用现状和前景的分析，本文重点应用数据挖掘技术来对目前现有高校图书馆管理系统中的数据进行分析和挖掘，试图找出读者的借阅特性和图书的利用特性，主要利用决策树和分类知识发现等技术来挖掘读者的行为范式，以了解图书馆所购资源与其所服务读者的特性及其相互间的关系，来为图书馆的工作、管理、服务、决策提供一些科学依据。

本文选择我校的图书馆管理系统中的数据来进行数据挖掘，目的包括：

（1） 为了了解读者的借阅特性，以及读者自身的兴趣特性，来为读者提供个性化的服务，包括：

① 图书陈列不再仅仅按照索书号排列，而抽出部分副本根据借阅的兴趣及时间特性来排列。

② 分析读者最有可能需要的信息，主动把信息推荐给读者。

（2） 找出图书被借阅的顺序及时间特性，开展图书阅读的特色服务，包括：

① 选出学科最受欢迎图书，开辟推荐书目栏块。

② 发现各学科的书籍的内在特性，由浅入深，由普教到专业的顺序特性，能够及时给读者进行读书引导。

我校在校学生2万多人，教职工近2000人，在全国高校中，我校在处于发展中的综合大学里具有一定的代表性。现以我校2002年到2007年的借阅记录来进行数据挖掘研究，试图找出读者的行为范式，不仅可以给我校的图书馆工作、教学科研带来一定的指引作用，同时也能为国内或国际同档次高校提供一定的借鉴作用，也可以作为决策者制定决策时的相关依据。

本文研究的内容包括：

（1） 了解图书馆现在的使用状况

（2） 以同一单位或部门的读者为对象探讨该读者群的借阅行为

（3） 以一学年的读者为对象探讨读者在整个学年中的借阅行为

（4） 以某一图书类别为对象，探讨分类图书的被利用特性

（5） 以低利用率读者群为对象，利用决策树算法分析其共同特征。

（6） 利用聚类算法，将读者进行细分研究。

本文研究的技术路线主要是，以图书馆管理系统中近 5 年的读者借阅历史数据以及图书馆的馆藏信息数据为研究对象，利用数据挖掘中的分类知识发现、关联规则发现、数据聚类、异常发现、序列模式发现等技术，进行图书馆系统数据挖掘研究，以期望从大量累计的图书馆数据中挖掘出有价值的、潜在的信息，进一步的了解读者以及馆藏图书的特性，为图书馆提供特色服务，缩短图书馆与读者，读者与图书、图书馆与图书的内在距离。让图书馆真正的做到由印度著名的图书馆学家阮冈纳赞所提出的“图书馆学五定律[14]”：书是供使用的；每个读者有其书；每本书有其读者；节省读者的时间：图书馆是一个生长中的有机体。

1．3．2 论文的组织结构与研究方法

本文首先从高校图书馆的发展和面临的问题角度介绍了课题研究的背景和意义，并分析了数据挖掘技术对改进高校图书馆服务水平的作用。然后对目前国内外相关领域的研究状况做了大致的介绍。第二部分主要介绍了数据仓库与数据挖掘的概念、体系结构、关键技术等，为课题研究打下了理论基础。第三部分进行了数据挖掘技术在高校图书馆中应用的理论研究；第四部分着手创建高校图书馆数据仓库模型体系，并对各个主题进行了分析。第五部分重点进行了数据挖掘技术在我校图书馆系统中的实例研究，并建立了高校图书馆的低利用率的读者特性模型以及读者细分模型，同时利用基于矩阵的数据挖掘算法对传统的 Apriori 算法进行了改进，较好避免了 Apriori 系列算法固有的缺陷，算法占用内存小，I/O 操作小，执行速度快，系统效率大大提高。最后总结了本文的研究成果，并指出了其中的不足以及对未来的展望。

本文采用数据挖掘技术对大学图书馆的读者管理、图书馆管理、工作服务等方面进行研究。主要采用了定性研究与定量研究、文献研究与实证研究相结合的方法。

文章第一、二、三部分是以文献研究、定性研究为主，对数据仓库、数据挖掘以及图书馆工作进行理论综述，分析了数据挖掘技术在大学图书馆系统中的可用性。第四部分在文献研究的基础上，对大学图书馆数据挖掘模型进行分析设计，分别建立了大学图书馆的物理模型和逻辑模型，第五部分利用数据挖掘算法对大学图书馆进行了实证研究，最后进

行专家分析总结并对未来进行了展望。

1．4 论文的创新之处

本文的创新之处在于将图书馆的管理、服务及未来发展与数据挖掘技术充分结合，来探讨如何利用数据挖掘技术来提高高校图书馆的服务质量，增强图书馆、读者、图书间的联系，强调数据挖掘技术在图书馆中的应用，不仅改变图书馆现有的服务方式，甚至将带来图书馆服务价值体系的转变，最终实现每位图书馆员都是知识的导航员，为每一位读者传递用户化的增值知识。

利用基于矩阵的数据挖掘算法[15]对传统的 Apriori 算法进行了改进，使得只需要扫描数据库 1 次，并且不用生成侯选项目集，同时节省了较大的存储空间。生成的频繁项集支持矩阵只需要对满足条件 i<j 的元素进行支持度计数运算等等都较好避免了 Apriori 系列算法固有的缺陷，算法占用内存小，I/O 操作小，执行速度快，系统效率大大提高。

数据挖掘技术的引入，将使图书馆的信息资源得以进一步的优化和丰富，信息服务的质量发生质的飞跃，业务服务的范围也将进一步拓展，从而为图书馆带来可观的经济效益和社会效益，但数据挖掘是一种新兴的智能信息处理技术，它的发展面临着许多的难题。同样，在图书馆的应用中如何将数据挖掘系统作为基本的数据分析模块与图书馆的原有业务系统、数据库系统和 WWW 资源有效集成？以实现集成的基于数据挖掘的信息处理环境，是面临的难题，有待进一步研究和探索。同时，除了图书馆业务系统中所产生的数据外，如何收集更多的读者相关的数据也是一个难题。

本章小结：本章首先分析了课题的研究背景及意义，深刻的分析了目前高校图书馆所面临的问题，并从数据挖掘技术已经在其他领域所取得的应用来分析数据挖掘在高校图书馆系统中应用的必要性、可行性，最后清楚的给出了本文的组织结构、研究方法及其创新之处。

第二章 数据挖掘技术和工具简介

2.1 数据挖掘概述

2.1.1 数据挖掘的概念、任务、功能

关于数据挖掘的定义[16-19]，有各种说法：

Hand et al(2000)的定义："数据挖掘是在庞大的数据库中找出有意义或有价值信息的方法。"

Bhavani(1999)的定义："数据挖掘是从存储在数据库中的大量数据资料中，设置盘问，提取以前未知的信息、模式和趋势的方法。""数据挖掘是从大量储存的数据中，利用模式识别、统计和数学的技术、筛选发现新的有意义的关系、模式和趋势的方法。"

Kovalerchuk & Evgenii Vityaev 的定义："这些技术现在用于发现潜藏在金融数据库中的趋势与模式。"

Berry and Linoff 对数据挖掘的理解更为深刻：分析报告给你的是后见之明；统计分析给你的是先机；数据挖掘给你见识。

综合以上的定义得出：数据挖掘就是从大量的数据中发现潜在规律、提取有用知识的方法和技术。数据挖掘用来探查大型数据库，发现先前未知的有用模式。数据挖掘还具有预测未来观测结果的能力，例如，预测一位新的顾客是否会在一家百货公司消费 1000RMB 以上。

数据挖掘的任务分为下面两大类：

（1） 预测任务。这些任务的目标是根据其他属性的值，预测特定属性的值。被预测的属性一般称为目标变量或因变量，而用来做预测的属性称为说明变量或自变量。

（2） 描述任务。目标是导出概括数据集中潜在的联系模式（相关、趋势、聚类、轨迹和异常）。本质上，描述性数据挖掘任务通常是探查性的，并且常常需要后处理技术验证和解释结果。图 2-1 展示了四种主要的数据挖掘任务。

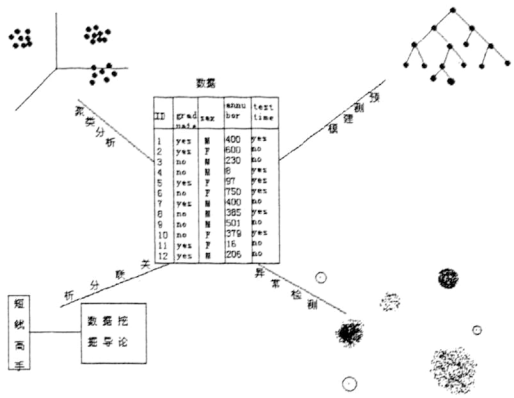

图 2-1 数据挖掘的主要任务

关联分析：是用来发现描述数据中强关联特征的模式。若两个或多个数据项的取值之间重复出现且概率很高时，就存在某种关联，可以建立起这些数据项的关联规则。在大型数据库中，关联规则很多，一般用“支持度”和“可信度”两个域值来淘汰那些无用的关联规则。

“支持度”表示该规则所代表的事例占全部事例的百分比。

“可信度”表示该规则所代表事例占满足前提条件事例的百分比。

聚类分析：面对海量的资料，首要的任务是将它合理地归类。如果已知要求，可以对资料设问，按照回答的不同给予分类，这叫做分类。如果事先没有任何要求，像全国各地环境监测的资料，就只能按资料反映的情况，比较接近的划归一类，这种归类的方法就是聚类。

聚类方法包括统计分析方法、机器学习方法和神经网络方法等。

预测建模：涉及以说明变量函数的方式为目标变量建立模型。有两类预测建模任务：分类，用于预测离散的目标变量；回归，用于预测连续的目标变量，是一种典型的方法，即利用大量的历史数据，以时间为变量建立线性或非线性回归方程。预测时，只要输入任意的时间值，通过回归方程就可求出该时间的状态[20]。

异常检测的任务是识别其特征显著不同于其他数据的观测值。这样的观测值称为异常点或离群点。异常检测算法的目标是发现真正的异常点，而避免错误地将正常的对象标注为异常点。也就是说，一个好的异常检测器必须具有高检测率和低误报率。

数据挖掘大体上有两种功能，即预测/验证功能和描述功能。前者指用数据库的若干已知属性预测或验证其他未知属性值；后者指找出描述数据的可理解模式。

2.1.2 数据挖掘的过程

数据挖掘一般包括5个步骤：

（1） 数据维护；

（2） 定义主题；

（3） 读入数据并建立模型；

（4） 理解模型；

（5） 预测。

事实上，数据挖掘是一个周而复始的过程，即从一个主题中产生的想法往往需要进一步分析从而导致新的主题，而新的主题又可以产生更新的主题。

（1） 数据准备

有些人喜欢将数据挖掘看作是一个不可思议的过程，认为它吞进的是原始数据，吐出来的则是钻石，为自己产生数百万元的资金。

实际上，数据挖掘是一个过程，而数据准备则为这个过程的核心。为一次数据挖掘研究准备数据时，需要考虑以下几方面的事情：

① 获取数据。数据挖掘过程中，获取数据的方式很多，可以通过访问数据仓库、通过基于事务的关系数据库或基于 PC 的数据库访问数据，可以通过数据转换工具访问数据；用查询工具访问数据或者从平面文件中访问数据。

② 提高数据质量。数据清洗，数据不总是“干净”的，可能会出现拼写或者表达的误会，还有数据失效和印刷错误的问题。数据缺失，数据的一些取值时常会发生缺失现象。这样有时会导致很难从数据中得到有用的信息，或依此为基础进行预测。

③ 数据导出。通常情况下，最有价值的数据是从已有的数据中导出的。要求首先限定数据的范围，再将数据进行分组。

（2） 确定主题

确定主题主要包括：了解研究主题的局限性，选择需要完成的良好的研究主题，确定待研究的合适的数据元素，以及决定如何进行数据抽样。

（3） 读入数据并建立模型

数据挖掘模型应该具有准确性和可理解性。就准确性而言，数据挖掘还不能完全代替统计分析，然而，统计分析又往往会漏掉被数据挖掘工具所发现的重要的关联关系。对于数据挖掘的可理解性[21]，可以从以下几个方面来理解。

① 模型是否可以使我们了解输入对结果会有什么作用？

② 模型是否可以使我们了解其预测为什么会成功或失败？

③ 模型是否可以使我们对复杂数据集产生预测结果？

④ 模型是否能对其产生的结果进行检测？

对于模型性能的评价可以分为两个方面：一是你可以什么速度构造出模型；二是你可以什么速度从模型中获得预测结果。

（4） 理解模型和预测

无论使用哪种模型，模型报告都会告诉你什么信息与特定结果具有关联关系。输入数据对特定结果具有影响并不意味着它们之间就一定具有因果关系。预测是一个相当直截了当的过程，即针对一组输入数据就是否会出现某一结果做出预测。对于真正的预测而言，预测所得出的都将是事先未知的结果。

2．1．3 数据挖掘策略

数据挖掘策略可以广义地分为有指导和无指导两类。有指导学习通过使用输入属性来预测输出属性值的方式建模。输出属性的结果依赖于一个或多个输入属性值，输出属性又称为因变量。当学习是无指导的时候，不存在输出属性。因此，所有用于建模的属性都是自变量。

图 2-2 表示了我们将要讨论的 5 种数据挖掘策略。

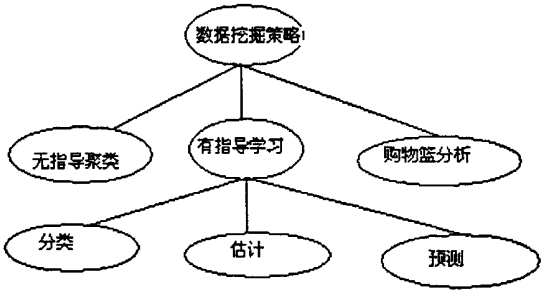

图 2-2 数据挖掘策略的层次结构

（1）分类。分类任务有 3 个一般特征：

① 学习是有指导的。

② 因变量是分类的。

③ 重点在于建立模型，将新的实例指派给一组定义明确的类中的一个。

分类任务处理的都是当前的而不是未来行为。而预测模型是被设计用于回答有关未来

行为的问题。

（2）估计。与分类模型相似，估计模型的目的在于确定一个未知输出属性的值，其输出属性的值是数值的而不是分类的。

（3）预测。要将预测同分类或估计相区分开来是不容易的。然而，与分类模型和估计模型不同，预测模型的目的在于确定未来的输出结果而不是当前的行为。预测模型的输出属性可以是分类的或数值型的。

（4）无指导聚类。对于无指导聚类是没有因变量来指导学习过程。学习过程通过使用聚类质量度量将实例分为两个或更多的类，来建立知识结构。主要目标在于发现数据中的概念结构。无指导聚类的一般作用包括：

① 确定能否在数据中发现概念形式的有意义的关系。

② 评估一个有指导学习模型的性能。

③ 确定有指导学习的最佳输入属性集合。

④ 侦测孤立点。

（5）购物篮分析。目的是要找到零售产品之间有趣的关系。购物篮分析的结果将帮助零售商设计推销方案、安排货架和商品目录，以及开发交叉销售策略。关联规则算法经常应用于对一组数据进行购物篮分析。

2．2 基本数据挖掘技术

（1） 关联规则方法

关联规则挖掘是由 Rakesh Apwal 等人首先提出的。两个或两个以上变量的取值之间存在某种规律性，就称为关联[22]。关联规则是形如“面包，黄油→牛奶”的一种规则，表示“在购买面包和黄油的顾客中，有 90%的人同时也买了牛奶”。

① 关联规则基本概念和问题描述

设 $I=\{i_1,i_2,\ldots,i_m\}$是项的集合。记 D 为事物 T 的集合，即事务数据库，其中，事务 T 是项的集合，并且$T \subseteq I$。每一个事务都有一个标识，记着 *TID*。设 X 是 I 中的一个项的集合，如果 $X \subseteq T$，那么称事务 T 包含 X。

关联规则是形如 $X \rightarrow Y$ 的蕴涵式，其中，$X \subset I, Y \subset I$，且 $X \cap Y = \Phi$。

规则 $X \rightarrow Y$ 在事务数据库 D 中的支持度是事务集中包含 X 和 Y 的事务数与所有事务数之比，记为 *support*($X \rightarrow Y$)，即 *support*($X \rightarrow Y$)=$\left|\{T : X \cup Y \subseteq T, T \in D\}\right| / |D|$。

规则 $X \rightarrow Y$ 在事务集中的可信度是指包含 X 和 Y 的事务数与包含 X 的事务数之比，记为 *confidence*($X \rightarrow Y$)，即 *confidence*($X \rightarrow Y$)=$|\{T : X \cup U \subseteq T, T \in D\}| / |\{T : X \subseteq T, T \in D\}|$。

给定一个事务集 D，挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度和最小可信度的关联规则。

②Apriori 算法

Apriori 算法是 Agrawal 等人于 1993 年提出了挖掘关联规则的一种重要方法，其基本思想是将关联规则挖掘算法的设计分解为两步：

1）找到所有支持度大于最小支持度的项集，这些项集称为频集。含有 k 个项的频集称为 k 项集。

2）使用第一步找到的频集产生所期望的规则。

Apriori 算法是一种最有影响的挖掘布尔关联规则频繁项集的算法。该算法为了生成所有频集，使用了递归的方法，其算法的伪代码可以表示为

L_1=|large1-itemsets|;
for(k=2;$L_{k-1} \neq \Phi$;k++) do begin
  C_k=apriori-gen(L_{k-1});
  for all transaction t∈D do begin
    C_t=subset(C_k,1);
    for all candidates c∈C_t do
      c.count++;
end
  L_k={c∈C_k|c.count≥minsup}
end
Answer=$\cup_k L_k$。

该算法首先产生频繁 1 项集 L_1，然后是频繁 2 项集 L_2，直到有某个 r 值使得 L_r 为空时，算法停止。在第 k 次循环中，先产生候选 k 项集的集合 C_k，C_k 中的每个项集是对两个只有一个项不同的数据 L_{k-1} 的频集做一个（k-2）连接来产生的。C_k 中的项集是用来产生频集的候选集，最后的频集 L_k 必须是 C_k 的一个子集。

（2） 决策树方法

决策树方法是利用实例集生成一个基于计算的测试函数[23]，根据测试函数值建立树的分支；在每个分支子集中重复建立下层结点和分支，从而生成一棵决策树。然后对决策树进行剪枝处理，最后把决策树转化为规则集，利用规则集可以对新实例进行分类。CART、

C4.5、CHAID、Quest 和 C5.0 是几种经典的基于决策树的分类算法。

（3） 神经网络方法

神经网络是模拟人脑的神经元结构而建立的多种神经网络模型。神经网络的知识体现在网络连接的权值上。神经网络方法用于非线性数据和含噪声的数据时具有更大的优越性，比较适合于市场数据库的分析和建模。它主要由三个部分组成：定义神经网络数据结构、如何进行学习的学习算法和获取信息的抽取技术。神经网络已经用于模式识别、语音识别、医疗应用、问题诊断、机器人控制等领域。

（4） 聚类算法

聚类（Clustering）是对物理的或抽象的对象集合分组的过程。聚类生成的组称为簇（Cluster），簇是数据对象的集合。簇内部的任意两个对象之间具有较高的相似度，而属于不同簇的两个对象间具有较高的相异度。相异度可以根据描述对象的属性值计算，对象间的距离是最常采用的度量指标。

聚类分析[24]主要针对的数据类型包括区间标度变量、二元变量、标称变量、序数型变量、比例标度型变量，以及由这些变量类型构成的复合类型。

一些基于内存的聚类算法通常采用数据矩阵和相异度矩阵两种典型的数据结构。

① 数据矩阵：设有 n 个对象，可用 p 个变量描述每个对象，则 n*p 矩阵 $\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$ 称为数据矩阵。数据矩阵是对象-变量结构的数据表达方式。

② 相异度矩阵：按 n 个对象两两间的相异度构建 n 阶矩阵 $\begin{pmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{pmatrix}$ 其中 d(i,j)表示对象 i 与 j 的相异度，它是一个非负的数值。当对象 i 和 j 越相似或“接近”时，d(i,j)值越接近 0；而对象 i 和 j 越不同或相距“越远”时，d(i,j)值越大。

k-均值聚类算法的核心思想是通过迭代把数据对象划分到不同的簇中，以求目标函数最小化，从而使生成的簇尽可能地紧凑和独立。首先，随机选取 k 个对象作为初始的 k 个簇的质心；然后，将其余对象根据其与各个簇质心的距离分配到最近的簇；再求新形成的

簇的质心。这个迭代重定位过程不断重复，直到目标函数最小化为止。设 p 表示数据对象，c_i 表示簇 C_i 的均值，通常采用的目标函数形式为平方误差准则函数：$E=\sum_{i=1}^{k}\sum_{p\in C_i}\|p-c_i\|^2$。这个目标函数中的距离度量是欧几里得距离，当然，也可以采用其他距离度量。

k-均值聚类算法：

输入 n 个对象的数据库，期望得到的簇的数目 k

输出 使得平方误差准则函数最小化的 k 个簇

方法

1）选择 k 个对象作为初始的簇的质心

2）repeat

3）　　计算对象与各个簇的质心的距离，将对象划分到距离其最近的簇

4）　　重新计算每个新簇的均值

5）until 簇的质心不再变化

面对大规模数据集，该算法是相对可扩展的，并且具有较高的效率。算法复杂度为 O(nkt)，其中 n 为数据集中对象的数目，k 为期望得到的簇的数目，t 为迭代的次数。算法通常终止于局部最优解。

k-均值的算法的缺点在于要事先给出期望生成簇的数目 k，这在某些应用中是不实际的。k-均值算法不适合于发现非凸面形状的簇和大小差异较大的簇。并且，该算法对“噪声”和孤立点数据敏感。

（5） 遗传算法

遗传算法[25-26]是一种全新的最佳化空间搜寻法，其最初概念是由 John Holland 于 1975 年提出，其主要目的如下：

① 以严密而抽象的科学方法解释自然界中“物竞天择、适者生存”的演化过程。

② 将生物界中基因演化重要机制以信息科学软件来做仿真。由于遗传算法的实用性、健壮性和优化的搜索方法，它已经应用到许多新领域，遗传算法提供了一种不同以往的思考模式，它可以在海量资料中快速搜寻、对比、演化出最佳点，并且具有学习机制，可以迅速有效地解决数据挖掘中的一些求解过程。

2．3　数据挖掘方法论及应用软件

目前比较流行的数据挖掘方法论主要有以 NCR， SPSS 大公司提出的跨行业标准数据

挖掘过程 CRISP-DM（Cross Industry Standard Process foe Data Mining），IBM 公司提出的通用数据挖掘方法，以及 SAS 公司提出的 SEMMA（Sample，Explore，Modify，Model，Access）方法论。实质上，它们的内容基本上都包括了数据准备、数据抽取、模型建立、模型评估和模型修正等过程。

1　SAS 的 SEMMA 方法

SAS 研究所提出数据挖掘方法勾画了使用其数据挖掘工具 Enterprise Miner 进行数据挖掘的大致过程。如图 2-4 所示，其中抽样涉及创建一个或多个数据表；探索即对数据进行深入探查以发现隐藏在数据中预期的或未被预期的关系及异常，从而获得对事物的理解和概念。然后通过建立、选择及转换变量对数据进行修改。建模即通过使用分析工具如回归分析、决策树和神经网络等从数据中发现那些能够对预期结果进行可靠预测的模式。评价即是对从数据挖掘过程中发现的信息的实用性和可靠性进行评估。

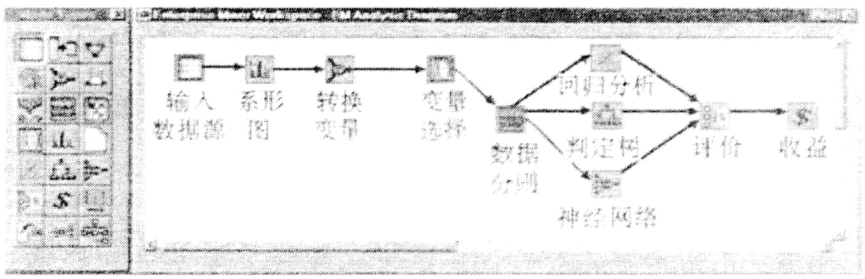

图 2-4 Enterprise Miner 数据挖掘的大致过程

SAS 将数据挖掘定义为对大量数据进行选择、探索、修改和建模的过程，其目的就是要从数据中揭示出以前未知的模式。可以适用于各种行业并且为解决诸如欺诈甄别、保留客户、消除摩擦、数据库营销、市场细分、风险分析、亲和力分析、客户满意度、破产预测、职务分析等业务问题提供了有效的方法。

2　IBM 的智能数据挖掘方法

Intelligent Miner 是由 IBM 公司开发的实用数据挖掘工具之一。它提供了专门在大型数据库上进行各种数据挖掘的功能，包括关联规则发现、序列模式发现、时间序列聚类、决策树分类和增量式挖掘等。其工具主要有数据智能挖掘机和文本智能挖掘机两种。

3　SPSS 的 CRISP-DM 方法[27]

NCR，SPSS 提出的跨行业标准数据挖掘过程 CRISP-DM 包括：商业理解、数据理解、数据准备、建立模型、模型评估以及结果发布 6 个步骤。

CRISP-DM 方法的流程图如下：

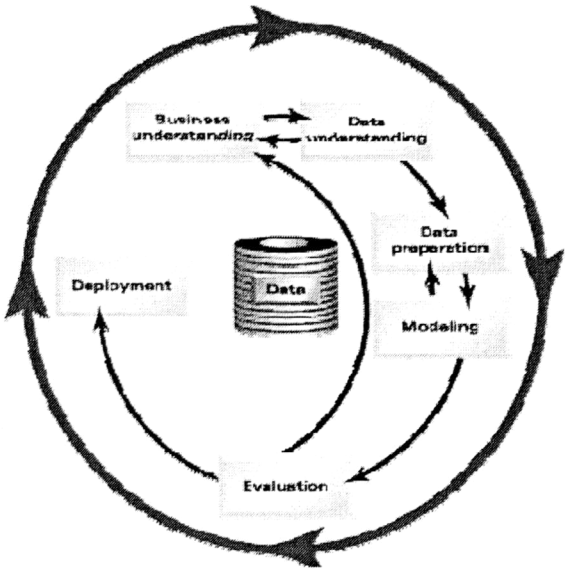

图 2-5 CRISP-DM 模型

如图 2-5 可知，CRISP-DM 模型包含了六个步骤，并用箭头指示了步骤间的执行顺序。这些顺序并不严格，用户可以根据实际的需要反向执行某个步骤，也可以跳过某些步骤不予执行。通过对这些步骤的执行，我们也涵盖了数据挖掘的关键部分。

Business understanding：商业理解阶段应算是数据挖掘中最重要的一个部分，在这个阶段里我们需要明确商业目标、评估商业环境、确定挖掘目标以及产生一个项目计划。

Data understanding：数据是我们挖掘过程的“原材料”，在数据理解过程中我们要知道都有些什么数据，这些数据的特征是什么，可以通过对数据的描述性分析得到数据的特点。

Date preparation：在数据准备阶段我们需要对数据做出选择、清洗、重建、合并等工作。选出要进行分析的数据，并对不符合模型输入要求的数据进行规范化操作。

Modeling：建模过程也是数据挖掘中一个比较重要的过程。我们需要根据分析目的选出适合的模型工具，通过样本建立模型并对模型进行评估。

Evaluation：并不是每一次建模都能与我们的目的吻合，评价阶段旨在对建模结果进行评估，对效果较差的结果我们需要分析原因，有时还需要返回前面的步骤对挖掘过程重新定义。

Deployment：这个阶段是用建立的模型去解决实际中遇到的问题，它还包括了监督、维持、产生最终报表、重新评估模型等过程。

有效的数据挖掘软件应当构造一个集成的数据挖掘系统，允许使用和比较不同的技术，还应该集成复杂的数据库管理软件。但这样的系统很少。

本文选择 SPSS 公司开发的数据挖掘软件 SPSS Clementine，它是一个数据挖掘工具平台，通过此平台可以采用商业技术快速建立预测性模型，并且将其应用于商业活动，从而改进决策过程。Clementine 是参照行业标准 CRISP-DM 模型设计而成，可支持从数据到更优商业成果的整个数据挖掘过程。

Clementine 的操作界面如图 2-6 所示：

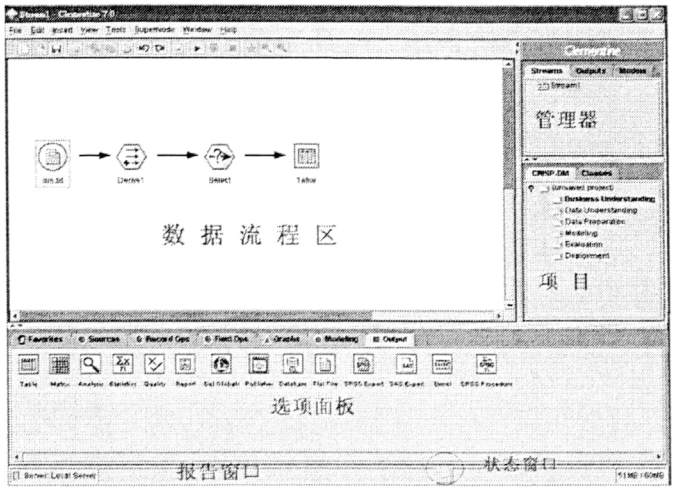

图 2-6 Clementine 的操作界面

Clementine 具有的优势及功能包括：

（1） 具有分类和预测、聚类、关联分析、时序分析等功能，提供神经网络、决策树与回归树、线性回归、Logistic 回归分析、自组织网络、快速聚类、二次聚类、主成分分析和因子分析等多种方法。

（2） 具有交互式可视化的界面，几乎所有的操作都可以在窗口下实现，而不需要编程来实现。

（3） 具备开放的数据库接口，支持定界或等宽格式文本文件、SPSS 文件、SAS 文件和多种类型的关系数据库。

（4） 提供两种建立模型的方式，在简单模式下，用户无需做任何设定，系统会按照默认的设置建立模型；在专家模式下，用户则可以根据自己的需要对模型中的各个参数进行适当的调节，从而使模型达到最佳的效果。

（5） 提供了强大的发布功能，可以将数据挖掘模型或者整个数据挖掘流程导出至嵌入系统。

（6） 提供了完善的数据流管理和项目管理的功能。前者可对工作区域内的数据流、数据挖掘模型、数据挖掘结果进行有效的管理。后者可对整个项目进行有效的管理，用户既可以按照数据挖掘的不同阶段对相关项目文件进行管理。

（7） 提供了 CEMI（Clementine external module interface）技术，可以把其他模型、数据准备、结果展示等功能集成到 Clementine 中。

本文利用 Clementine，对图书馆系统数据进行挖掘操作，试图建立图书馆读者行为范式以及文献资源的使用模式。

本章小结：本章详细的介绍了本文研究所需要的理论基础知识，包括数据仓库的概念、特征、体系结构以及其关键技术和系统结构的创建过程，数据挖掘的概念、任务、功能、过程以及数据挖掘的基本策略和技术，最后对目前流行的数据挖掘软件进行了大致的介绍，重点介绍了本文采用的 SPSS 公司开发的 SPSS CLEMENTINE 软件，为下面的数据挖掘和模型建立打下基础。

第三章 数据挖掘在高校图书馆系统中的应用分析

3.1 高校图书馆系统数据特点

1 高校图书馆的行业特点

（1） 高校图书馆的馆藏结构

高校图书馆的馆藏结构的特点是：图书、期刊根据学科分类，分别分布于不同的馆藏地。目前各大高校图书馆基本都是按照学科类别进行分布馆藏。通常分为自然科学藏书处、社会科学藏书处、外文文献藏书处。有的还特设文学藏书处、综合藏书处等。主要包括马克思主义、毛泽东思想、邓小平理论（A），哲学、宗教（B），社会科学总论（C），政治法律（D），军事（E），经济（F），文学、科学、教育、体育（G），语言、文字（H），文学类（I），艺术（J），历史、地理（k），自然科学总论（N），数理科学和化学(O)，天文学、地球科学（P），生物科学（Q），医药卫生（R），农业科学（S），工业技术（T），交通运输（U），航空、航天（V），环境科学（X），综合性图书（Z）22大类。

每个图书馆都是根据学校的发展状况及发展方向来确定自己的藏书特色。从多数高校图书馆的藏书结构来看，语言、文字（H），数理科学和化学(O)以及工业技术（T）三类图书占总图书比例较大，该类图书利用率也相对较高。

高校图书馆的藏书结构与社会公共图书馆和书店有着明显的区别。社会公共图书馆各类藏书量较为平均，这是由其服务读者层次丰富决定的。书店以追求最大的经济利益为其运作目标，其藏书结构也带有很强的利益性。高校图书馆的藏书结构有其独特的优点：

① 所藏图书资料范围广泛、全面，更加注重知识性。

② 所藏图书与读者的需求息息相关，与读者的专业兴趣联系紧密。

③ 不仅藏有最新的图书资料，而且也拥有丰富的历史图书资料。

因此高校图书馆必须充分发挥其作用与价值，不仅为高校输送知识养分，同时也要为社会整体素质的提高发挥重要作用。

（2） 高校图书馆的服务体系[28-30]

高校图书馆通常包括文献采购编目、流通借阅、期刊资料、读者服务和图书馆管理五个部门。高校图书馆的服务体系基本以服务读者为中心，文献建设与图书馆管理是服务读者的基本保障。其服务内容主要有：

① 每天为注册读者提供十多个小时的各类文献的借、阅、还免费服务。

② 丰富的电子资源可供注册读者 24 小时免费检索、下载。

③ 为有需求的读者提供信息咨询、原文传递、定题服务、代检代查服务。

④ 提供用户教育及培训，开设“文献检索与利用”课程以及各类专题讲座。

其特色是全流程计算管理，节省大量的人力，将图书馆员从大量的体力劳动中解放出来；同时读者可以自由取阅，比起以往的闭架借阅，拓宽了读者视野，扩大了阅读范围。另外，随着馆藏量的不断丰富，以及电子资源的发展，读者的借阅权限也越来越大，可供读者使用的文献范围也越来越广。

（3） 高校图书馆的服务方式

目前各大高校图书馆依然以被动服务为主，即主要为上门读者提供“借、阅、还”以及咨询服务。很少有能够主动提供服务的，即每个图书馆员都应该认为自己承担着信息积极提供者的角色，而不是采用一种“我们就在这儿，你们来向我们咨询吧”的态度。这就要求图书馆员不仅要懂得怎样能够快速获取信息及其细节，还要知道是谁正需要这些信息。这一服务方式对于国内的大部分图书馆来说都是很大的挑战。

2 高校图书馆的数据特点

对于一般的一所综合性大学，在校师生达 2 万余人，知识水平都是大学学历及以上的水平，读者的求知欲以及对图书馆的期待都较高，这些特点决定了高校图书馆的数据具有以下特点：

（1） 大量性。随着高校的不断发展，读者数量的不断增大，以及高校图书馆的硬件、软件环境不断的改善，到馆读者的数量不断增大，图书馆馆藏也不断加增，高校图书馆终端服务系统每天所累计的借阅数据、馆藏数据也以惊人的速度迅速增长，所面对的高校图书馆的借阅记录、馆藏数据都达到了海量的级别。

（2） 关联性。在高校图书馆系统所积累的大量借阅记录数据中，由于不同的读者同时借阅多种图书都有自己的原因，这就说明借阅记录数据之间存在有一定的关联性。例如：同种图书、同类图书在不同的时间被不同的读者所借阅；某几类图书总是被不同的读者同时借阅；同一读者在不同的时间借阅了不同的图书等等。高校图书馆所积累的借阅记录数据从某些角度上去看总会具有某种程度的相关性。

（3） 包含信息的潜在性。高校图书馆的大量数据中一定包含有许多非常有价值的信息，比如可能从中分析和挖掘出读者的现在及未来需求，让每个图书馆员都知道自己所拥有的信息被哪些读者所需要。当然，这些信息都是隐藏的，无法直接得到，必须通过各种原理和工具对数据进行分析和挖掘，才能使这些数据变成对每个馆员和决策者真正有用的数据。

事实上，数据挖掘技术在高校图书馆系统中的应用还存在很大的障碍。这主要表现在数据完整性的问题，重点要研究读者的行为模式，但是只能掌握读者所借阅图书的数据，无法与读者在校的其他活动相关联进行整体分析，另外图书馆自身所收集的读者信息数据内容也不完整，基本上只包含了读者的借阅证号和姓名、班级的信息，这限制了要进行的挖掘深度。因此，图书馆的管理应该进一步注重对读者信息库的建设。

虽然数据挖掘在高校图书馆系统中的应用存在障碍，但是根据以上分析的数据特点，还是可以看出，在高校图书馆的系统管理中应用数据挖掘技术具有很大的必要性和可行性。

3．2　数据挖掘技术在高校图书馆系统中的应用分析

3．2．1　数据挖掘技术在图书管理中的应用

1　文献采购和馆藏建设方面

高校图书馆为了能够获得更高的资源利用效率，提高图书馆馆藏建设的资金效益，需要根据读者的需求及变化，对文献采购和馆藏进行合理的规划。

通过数据挖掘系统，可以将借阅记录数据和文献典藏数据集中起来进行管理。应用数据挖掘技术，可以协助采购员对各类图书进行增减，确保正确的采购；协助编目员确定新书的最佳上架时间，从而能够让读者在最想要的时候很容易发现自己最需要的文献。

2　文献管理利用方面

图书馆的馆藏资源有数百万，但是并不是每一本图书都能为读者所利用。“二八”法则同样也适用于图书馆，即图书馆有20%的文献拥有80%的图书利用率。因此，从所有的馆藏中发现利用率最高的文献以及利用率最低的文献，来调整馆藏的结构；确定图书推荐栏目的内容；定时的对失去价值的图书进行淘汰或转赠等。

在图书馆系统的借阅数据库中，每天都积累了大量的数据，但这些数据本身并不能直接反映出读者的借阅倾向和研究趋势。数据挖掘技术通过对历史借阅记录数据的深层次分析，可以得出各种图书的借阅情况，从而可以更科学地制定各类图书的整体战略，对各类文献进行管理，为每一位读者提供科学的个性化服务。

3　文献排架管理方面

高校图书馆从以前的卡片目录借阅到如今的开架借阅，文献的排架一直是根据图书的分类号来排架的。这样排架的优势是能够让读者快速准确的找到所要的图书，而且在读者查找的同时也会帮助他们很方便地发现同类的其他图书，然而这种历史的延续方法并不表

明不存在其局限性。第一、这种方法要求读者在借阅之前必须已经查得该书的索书号。第二、因为图书馆每天都会有新书入库，如果严格按照图书分类号进行排架，就必须为每类图书预留足够的扩充空间。这样，预留空间太大，会浪费宝贵的空间资源，预留空间不足，将会引起更大的图书重新整架工作。第三，仅能够在一定程度上扩大读者对某一类图书的阅读范围，无法引起读者对其他类别图书的注意。例如，一个计算机科学专业的学生，每到图书馆就会直接到自然科学借书处的 TP 类书架寻找自己可能感兴趣的图书，他可能从未走进相隔并不太远的文学类图书借阅处。有的人或许认为这并不重要，但是，这对于培养德智体全面发展的综合性人才是不利的。

应用数据挖掘技术可以对图书馆系统中的数据进行分析，得出各类图书的利用率，挖掘出某一时期内最受欢迎图书，同时开设好书介绍栏目，甚至可以针对特定群体进行好书介绍和推荐。另外，还可以利用数据挖掘技术，了解读者的借阅习惯和偏好，考虑读者在图书馆的连贯借阅行为以及借阅图书的类别，根据挖掘出来的图书之间的关联性，来设立专门的图书展读处，并且将该处设立在每位进馆读者的必经之地，这样图书馆就能做到变等待读者上门借阅为主动吸引每一位到馆读者利用图书馆资源。另外，还可以根据各类图书的被借阅率以及时间特性，来科学地制定由图书馆主办的各类读书节活动，来有效地促进图书馆的工作效率和资源利用率。

3. 2. 2 数据挖掘技术在读者管理中的应用

如今的时代，是服务至上的时代。如今的市场，是买方的市场。而对于图书馆来说，读者就是上帝，所有的注册读者就是其服务的市场。那么数据挖掘中重要的客户关系管理对图书馆来说也是很重要的。

客户关系管理包括客户获得、客户保留、客户发展。对于图书馆来说，获得读者主要是每年新入学的新生以及新引进的教师职工，这一部分图书馆不需任何努力就可完成；相比而言吸引校外读者则是获得读者需要关注的重点工作，这包括两部分：已经办理通用借书证的读者和未办理注册的其他人员。这需要找出现有读者的文化水平、工作性质和兴趣爱好的特征，然后按照这些特征去寻找潜在的新的读者群。读者保留对于图书馆来说也有其特有的含义，通常读者不会主动注销其读者身份，除了因毕业、离校等原因外。因此，读者保留有两层含义：

（1） 图书馆可以取消因毕业或离校等原因而注销读者身份的制度，而可以采用预付押金等方法为已经离校的读者继续提供借阅服务。

（2） 挖掘出那些已经注册但基本不使用图书馆资源的读者——可以称之为“惰性读者”——的专业部门、学习工作状况等特性，可以为其提供他们可能感兴趣的信息。

读者发展，可以理解为不断的促进读者的阅读量，扩大读者的阅读兴趣和范围。这些要求都离不开数据挖掘知识，因为数据挖掘可以提供比较科学、可靠的依据，使图书馆制定的决策科学化，而不仅单纯依赖感性的经验。

3. 2. 3 数据挖掘技术在工作管理中的应用

利用时间特性和图书类别之间的关系进行聚类分析可以挖掘何时是业务工作繁忙期，何时是某类图书需求旺期，可以为安排人力和设定某类图书的流通周期提供一定的依据。

利用读者特性与图书类别的关系进行聚类分析找出读者的借阅特性，可以为主动向读者提供个性化服务提供有力的依据。

本章小结：本章详细分析了高校图书馆的工作服务特点以及数据特点，分别对数据挖掘在图书管理、读者管理、工作管理三个方面的应用进行理论阐述，为下文的模型设计奠定了基础。

第四章　高校图书馆数据库模型设计

数据库的应用，主要包括了两方面的内容，首先是数据库的建设内容，其次是数据分析的应用内容。数据库的创建基础是构建数据库的事实表和维表。在 Oracle 中创建数据库[31-33]，通常经历创建数据库、创建数据库表空间和创建数据库表等几个过程。

4.1 Oracle 数据库创建

Oracle 数据库创建[15]，采用其数据库构造助手（ODCA，Oracle Database Configuration Assistant）进行。主要包括：创建数据库、选择数据库模板、设置数据库标志名称为 libsys，SID 为 libsys、取消所有的“数据库特性”选项、选择专用服务器模式、设置数据库初始参数：用于 Oracle 的物理内存的百分比为 60%，数据库块大小为 8192 字节，设置表空间 system 为 1G，temp 为 512M，UNDOTBS1 为 2G、为每个重做日志组添加一个成员，即每个重做日志组由两个文件组成，并将两个文件分别放在不同的逻辑盘上，可以提高数据库的稳定性，然后开始数据库的创建过程，最后设置管理该数据库的 sys 用户和 system 用户的密码，均设为 libsys，最后完成数据库的创建。

然后以 libsys/libsys 的用户名、密码登陆 sql/plus，执行：

REM*****数据表空间 2G+2G=4G*****

Create tablespace LIB_DATA datafile 'D:\oracle\oradata\lib\lib_data1.ora' size 2048M extent management local autoallocate;

Alter tablespace LIB_DATA

add datafile 'D:\oracle\oradata\lib\lib_data2.ora' size 2048M;

Alter tablespace LIB_DATA

add datafile 'D:\oracle\oradata\lib\lib_data3.ora' size 2048M;

Alter tablespace LIB_DATA

add datafile 'D:\oracle\oradata\lib\lib_data4.ora' size 2048M;

Alter tablespace LIB_DATA

add datafile 'D:\oracle\oradata\lib\lib_data5.ora' size 2048M;

REM*****索引表空间 2G+2G=4G*****

Create tablespace LIB_IDX datafile 'D:\oracle\oradata\lib\lib_idx1.ora' size 2048M extent management local autoallocate;

```
Alter tablespace LIB_IDX add datafile 'D:\oracle\oradata\lib\lib_idx2.ora' size 2048M;
Alter tablespace LIB_IDX add datafile 'D:\oracle\oradata\lib\lib_idx3.ora' size 2048M;
Alter tablespace LIB_IDX add datafile 'D:\oracle\oradata\lib\lib_idx4.ora' size 2048M;
Alter tablespace LIB_IDX add datafile 'D:\oracle\oradata\lib\lib_idx5.ora' size 2048M;
REM*****临时表空间 2G+2G=4G*****
Create temporary tablespace LIB_TEMP tempfile
'D:\oracle\oradata\lib\lib_temp.ora' size 2048M extent management local uniform;
REM*****创建 libsys 用户*****
REM*****口令 libsys，默认表空间 lib_data，临时表空间 lib_temp*****
CREATE USER LIBSYS IDENTIFIED BY LIBSYS
DEFAULT TABLESPACE LIB_DATA
TEMPORARY TABLESPACE LIB_DATA
TEMPORARY TABLESPACE LIB_TEMP;
GRANT CONNECT,RESOURCE,DBA TO LIBSYS;
```

至此，完成了数据表空间、索引表空间、临时表空间和数据库表的创建，再用 imp 命令导入图书馆管理系统导出的备份文件即可。

4.2 数据库分析与模型开发

4.2.1 构建逻辑模型

根据第三章的数据挖掘在高校图书馆中的理论研究结果，这里主要建立读者资料的逻辑模型[34]。

首先，需要根据分析的需求定义读者资料所应包含的信息。时间信息：办证的时间、注销的时间、借阅量最大的时间等；读者类型信息：学生/研究生/教师、单位类型（工科/理科/文科）等；读者个人资料：借书证号、姓名、电话号码、性别、E-mail 等；读者总借阅量：年度量、平均量、学期量等。根据这些信息进一步细化，将得到具体的数据字段，然后按照维度建摸的规则将整个数据模型设计成星型结构，如图 4－1 所示。

4.2.2 构建物理模型

物理数据模型是依据中间层的逻辑数据模型创建的。它通过确定模型的键码属性和模

型的物理特性，扩展中间层数据模型而建立的。此时，物理数据模型就由一系列表所构成，其中最主要的是事实表模型和维表模型[16]。

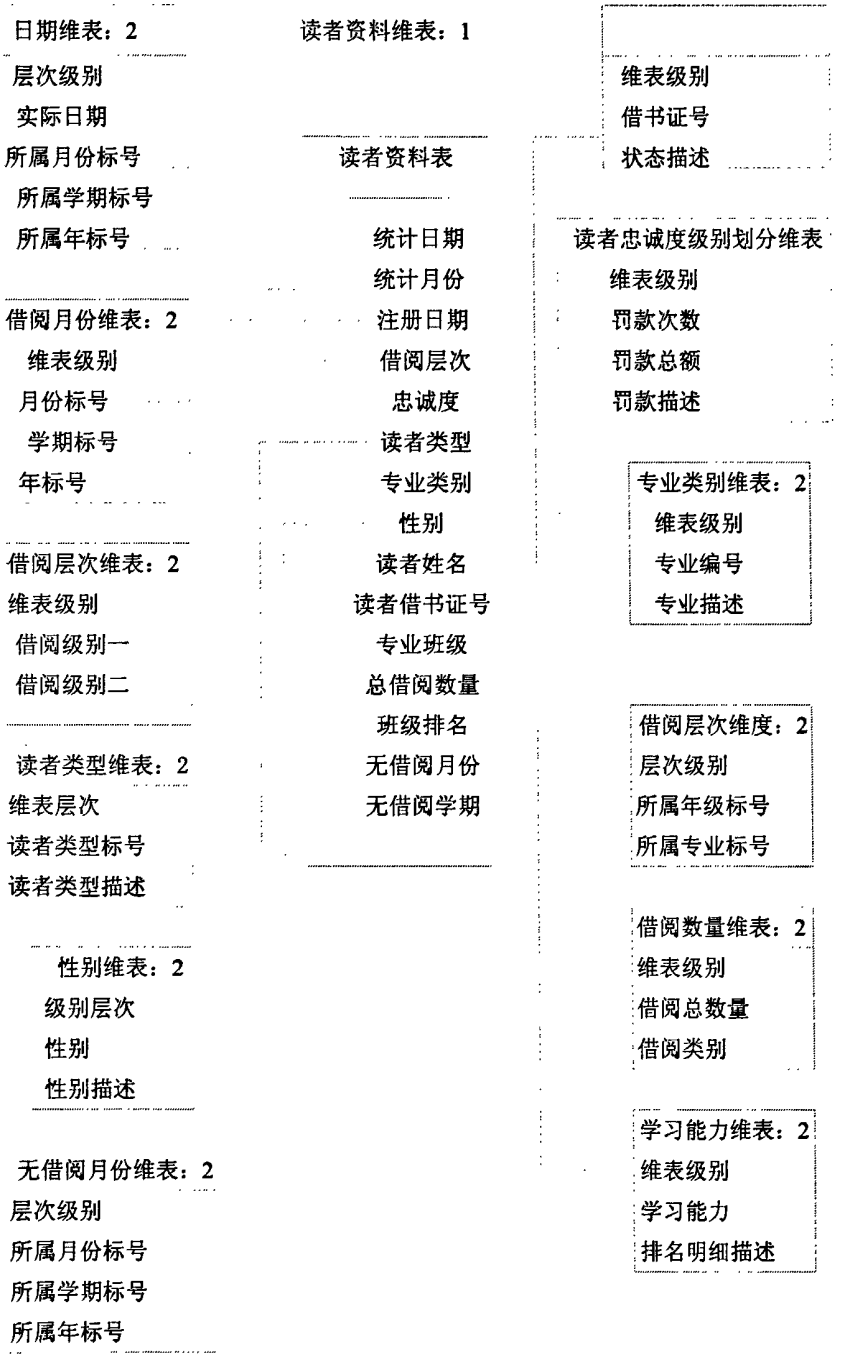

图 4—1 数据模型的星型结构图

（1） 事实表模型设计（列出三个主要的事实表）

① 读者信息事实表

Readerinfo(CERT_ID,ID_CARD,NAME,SEX,DEPT,OCCUPATION,DUTY,TOTAL_LEND_QTY,VOLT_FLAG,DEBT_FLAG,REDR_REG_DAY,REDR_DEL_DAY,REDR_TYPE_CODE,LEND_GRD)

② 读者借阅事实表

LENDHIST

（CERT_ID_F,PROP_NO_F,LOCATION_F,LEND_DATE,RET_DATE,RULE_NO_F,EXCD_DAYS,MARC_REC_NO_F,CALL_NO）

③ 图书信息事实表

BOOKITEM(PROP_NO,MARC_REC_NO,BOOK_STAT_CODE,BOOK_LEND_FLAG,IN_DATE,YEAR_CIRC_TIMES,TOTAL_CIRC_TIMES,FST_USE_DATE,LAST_USE_DATE,LOCATION,CALL_NO)

（2）维模型设计

维度表模型也是根据逻辑模型设计的，设计维度表的目的主要是把参考事实表的数据放置在一个单独的表中。其数据是直接参考事实表的数据。

本文采用 oracle enterprise manager console 维表创建向导来生成维表，其中日期维表和索引表结构如下：

日期维表：如表 1，对日期为分级及各个级别的中文描述。

表 1 日期维表

Name	Code	Type	P
层次级别	Agg_level	Smallint	No
实际日期	Lend_date	Date	No
所属月份标号	Lend_Month	Integer	No
所属学期标号	Lend_Term	Integer	No
所属年标号	Lend_Year	Integer	No

索引列表：如表 2 所示。

表 2 索引列表

Index code	P	F	U	Column code	sort
Idx1_d1	No	No	No	Month	ASC

本章小结：本章主要讲述了高校图书馆数据库模型设计的步骤和过程。首先详细讲述了本文建立 LIBSYS 数据仓库的过程与参数设置，其次从现实图书馆运营所产生的问题中提取分析主题，建立概念模型、逻辑模型以及物理模型，为下文的数据仓库的分析做准备。

第五章　我校图书馆管理系统数据挖掘实例研究

5．1　确定主题

本论文主要是对图书馆管理系统的数据进行挖掘，目的为了了解读者的行为特性、图书的被使用特性，来进一步提高图书馆的服务质量。根据这一主线，可以确定本文所要讨论的主题[35-37]：

(1) 了解图书馆现在的使用状况

① 各类图书被借阅利用情况；

② 所有读者的年度、月度利用图书馆资源的情况。

(2) 利用数据挖掘技术挖掘读者的行为特性

① 以同一单位的读者为对象探讨该单位或部门的读者的借阅行为；

② 以一学年的读者为对象探讨读者在整个学年中的借阅行为；

③ 以某类读者为对象探讨其注册有效期内的借阅行为。

(3) 利用数据挖掘技术挖掘图书的使用特性

① 以某一图书类别为对象，探讨分类图书的被利用特性；

② 根据读者借阅的图书类别，探讨图书之间的关联性。

(4) 利用数据挖掘技术找出图书馆被利用的时间特性

① 以年份为单位，找出每年图书馆资源建设、读者阅读趋向的变化特性；

② 以月份为单位，找出一年中每月读者利用图书馆的大致规律；

③ 以天为单位，找出读者使用图书馆的小时特性。

5．2　数据的选取

本文的研究对象是我校图书馆管理系统，所以所有数据均来源于该系统数据库。该系统于 2002 年 8 月投入使用，并且对我馆以前的所有馆藏进行了回溯建库，所以本文选取 2002 年 8 月到 2007 年 8 月的数据作为研究对象。我校图书馆管理系统是以 Oracle Database 作为数据库系统，主要包括 6 大模块：编目、流通、采访、典藏、统计、系统管理。本文选取其中的重点数据，包含我校所有的馆藏信息、２００２年以来我校所有读者的基本信息以及借阅信息，以及馆员的工作日志作为数据挖掘的研究对象。

(1)　我校馆藏信息数据表（bookitem）

包含财产号(PROP_NO)、MARK 记录号(MARC_REC)、条码号(BAR_CODE)、索书号、题名、责任者、出版社、ISBN、年代卷期、单价、分配地、载体类型、 Mark 状态、经手人、入藏时间等。在数据仓库中的存储格式如图 5-1 所示。

	PROP_NO	MARC_REC	BAR_CODE	BOC	E	YEA	PR	IN_DATE
1	2002010720	0000000001	2002010720	41	0	1998	14.00	2002-06-27
2	2002010721	0000000001	2002010721	41	0	1998	14.00	2002-06-27
3	2002010722	0000000001	2002010722	41	0	1998	14.00	2002-06-27
4	2002010723	0000000001	2002010723	41	0	1998	14.00	2002-06-27

图 5-1 馆藏信息表

(2) 读者借阅历史记录库(lendhst)

包含字段有：证件号(CERT_ID_F)、财产号(PROP_NO_F)、馆藏地(LOCATION)、借书日期(LEND_DATE)、还书日期(RET_DATE)、姓名、单位、书刊条码号、题名、作者、出版社、价格、索书号、借书经手人、借阅规则、借阅方式。 格式如图 5-2 所示。

	CERT_ID_F	PROP_NO_F	LOCAT	LEND_DATE	RET_DATE
1	00002755	9999010509	00001	2002-08-2616:05:13	2002-08-2616:52:46
2	90000080	2002012000	00003	2002-08-2609:48:01	2002-08-2709:02:03
3	90000048	2002012111	00003	2002-08-2709:03:09	2002-08-2709:03:34
4	90000048	2002029069	00003	2002-08-2709:11:21	2002-08-2709:12:03
5	90000048	2002012001	00003	2002-08-2709:12:51	2002-08-2709:13:57

图 5-2 历史借阅记录表

(3) 读者基本信息数据表(readerinfo)

包含读者条码号(CERT_ID)、有效证件号(ID)、姓名(NAME)、性别(SEX)、单位(DEPT)、读者类型、借阅等级、文化程度、职业、住址、邮编、电话、E-mail、违章次数、累计借书量、最大借阅册数、办证日期、证件状态。格式如图 5-3 所示。

	CERT_ID	ID	NAME	SEX	BIRTHDAY	DEPT	OCCUPATION	DU
1	10002377		戴 彦			989计本2		
2	10002378		陈 涓			989计本2		
3	50000404		徐文莉			社科系		
4	50000400		陈继华			社科系		
5	00002003		魏绍钧			985天动4		

图 5-3 读者基本信息表

(4) 图书分类表(call_no_lst)

包括类名(CALL_NO)、类号、MARK 号(MARK-REC_NO)、年总借阅次数(YEAR_CI)等。如图 5-4 所示。

	CALL_NO	C	LE	CIRC	MARC_REC_I	YEA	TOTA
1	I247.57/1	01	01	00	0000000001	54	54
2	TP316.86/1	01	01	00	0000000002	25	25
3	K712.03/1	01	01	00	0000000006	19	19

图 5-4 图书分类表

(5) 部门院系统计表

我校共包括教学单位 16 个学院和一个培训中心、8 个党群组织、18 个行政机构和 4 个直属部门。该表主要包括单位名称和代码。

(6) 阅览历史数据表

包含读者姓名、读者单位、读者类型、读者等级、读者证件号、到达时间。

(7) 读者借阅规则表包括读者类型、读者借阅等级，格式如图 5-5 所示。

	NO	LOCA_C(	CI	REDR_TYPE_COI	REMARK	RULE_NO
▶ 1	001	00001	*	01,08		0000000001
2	002	00003	*	09		0000000029
3	003	00016	*	09		0000000028

图 5-5 读者借阅规则表

(8) 读者违规代码，格式如图 5-6 所示。

	VOLT_CODE	VOLT_NAME
▶ 1	00	盗窃图书
2	01	损坏图书
3	02	外借复印超期
4	03	随书光盘借阅超期
5	04	随书光盘遗失
6	05	随书光盘损坏

图 5-6 违反规则表

5.3 数据的处理与转换

5.3.1 数据清洗

主要采用 SPSS 软件来对所有的数据进行整理、排序等操作。

（1） 读者信息的清洗

读者信息主要选取借书证号、单位、性别、读者类别、借阅规则 5 个字段。检查这 5 个字段，如果出现填充不完整、或缺失的现象，则删除该条记录。

（2） 馆藏信息数据的清理

这个表中的数据需要处理的多为人为输入错误，如小数点的问题、非法字符的问题或多一些空格的问题，对于这些错误可以人工修正，还有一些遗漏的数据，则采取删除记录的方式，因为本研究中选取的数据，都是以海量数据为单位进行研究，所以个别记录不会影响研究的结果。

（3） 读者借阅记录数据的清理

这个表包括我校 02 年 8 月到 08 年 1 月图书馆所有的借还数据。其中各个变量属性都存在缺失问题，数据表总计包含 2320770 条记录，利用 SPSS Clementine 的数据审核来进行分析结果如图 5-7 所示，得出借阅证号为空的记录有 253215 条，图书财产号的空值记录为 6513 条。在 SPSS 中以借阅证号和图书财产号为第一、第二排序变量进行排序。并将这两个变量值为空的实例进行文件分割。只对这两个变量有实际值的实例进行分析和操作。

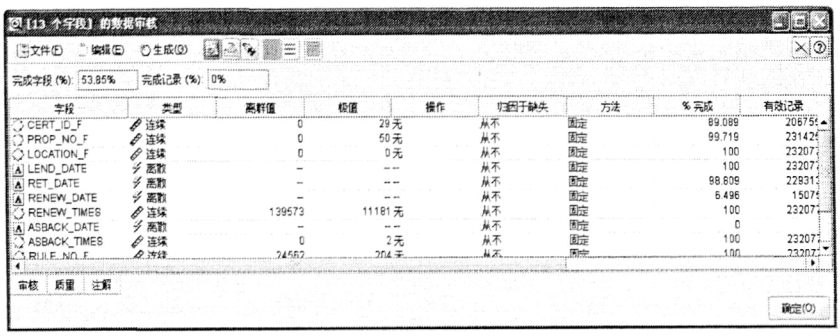

图 5-7 数据审核结果

5．3．2 数据转换

1 索书号转换

图书的索书号分类非常细致而且数值分散，为了可以进行聚类分析，在查询分析器中进行数据处理，分别选取第一大类和第二大类。

2 借阅时间转换

图书馆管理系统中产生的借阅时间格式为“年-月-日-小时-分钟-秒”，为了挖掘读者借阅的时间规律以及图书被利用的周期规律，在 SPSS 中利用其字符串读取函数 Substr(s,x1,x2) 将原数据表中的借阅时间字符串分别读取为借阅年，借阅月，借阅日，借阅小时，删除分钟、秒；对还书时间做同样的处理。

5．3．3 数据统计分析

1 图书被利用情况研究[38]

以 02 年图书阅读排行为研究对象，设定类型字段，题名为输出选项，其他均为输入选项，再添加神经网络节点，建成神经网络模型，点击浏览模型，用神经网络算法估计的准确性达到 98 .387，输入 51 个神经元，输出 1 个神经元，得到借阅次数为相对重要的输入

属性。因此再以题名和借阅次数为字段生成网络图形，流程图如图 5-8，运行结果如图 5-9 所示。

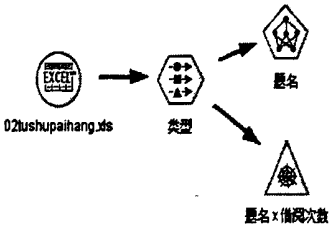

图 5-8 借阅次数与题名的网络图生成流程

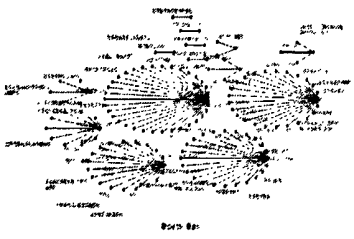

图 5-9 网络图运行结果

由图可见， 02 年的图书被借阅情况：

文学类书籍借阅排名第一，其中《亮剑》、《爱情从今晚开始》、《淑女味道：女性的世相》等为最受欢迎书籍；

排名第二的为《高等数学习题集》、《模拟电子技术基础学习与解题指南》、《大学英语六级考试真题详释》、《<高等数学>习题详解·释疑·指导》、《高等数学学习引导》等属于辅助教材类。

找出最受欢迎以及次受欢迎图书系列，就可以为这类图书开辟专门的、醒目的板块，主动吸引读者的眼球，进一步扩大此类图书的被阅读范围。同时可以将文学类图书和数学、工业技术图书的排架做关联，隐性的为读者补充及完整其知识体系。

图 5-9 是网络图形输出，显示的是分类变量值之间同时发生的频数图形，图中除了可以看出被利用率最高的图书外，还可以看出大部分图书被借阅次数为 2 到 6 次，这是具有相同的、不太高的利用率的很大一部分图书，可以进一步挖掘该类图书的特性，可以根据其特性进行某些促进阅读的手段，或者进行一些图书的淘汰。

按照借阅次数进行分布分析，得出借阅次数为 1、2、3、4 次的图书占总借书量的 77.78%，因此添加选择字段将借阅次数为 1 次、2 次、3 次和 4 次的图书记录导出，再用类名字段生成图书分布图，得出结果依然是工业技术、数理化、文学和语言文字占了总比例的 80%以上。数理科学化学类图书中数学类(01)占了近 89%，内容多为习题辅导类，物理类(04)占了 11%，内容为教材参考书类，没有化学类图书。工业技术中全属计算机类（TP3），内容包括网络、单片机、电脑故障以及软件系统的应用。语言文字中，英语类（H31）占了 87.5%，内容以等级考试和课外阅读类为主，日语类（H36）占了 12.5%，为初级学习类教材。

针对借阅次数为 1 到 4 次的图书，采取相应的推荐和导读策略，可以提高图书的利用效率。

2　以图书类名字段分布为研究对象

首先把 02 年所有图书按照类别进行统计，得到图书类别借阅量分布情况如图 5-10 所示。

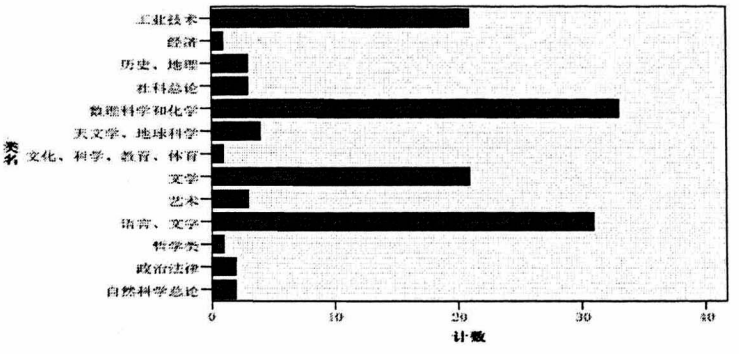

图 5-10 02 年图书类别分布

数理科学和化学类排名第一，语言文字类排名第二，工业技术类与文学类被借率相当，排名第三，天文学、地球科学与历史、地理，社科总论，艺术类相比，被借率稍高；经济类、科教文体、哲学类图书被借率最低。这是 02 年我校图书的借阅情况，说明基础类或公共类学科，倍受重视；工业技术与文学类图书受欢迎程度一致，比天文学、地球科学类图书的被借率要高，这与我校在 02 年的以大气科学、信息科学与技术、环境科学与工程为主，同时注重哲学社会科学学科建设体系的形成思想一致。

下面将 04 年、05 年、06 年、07 年图书被利用比例分布图进行对比，如图 5-11 所示。

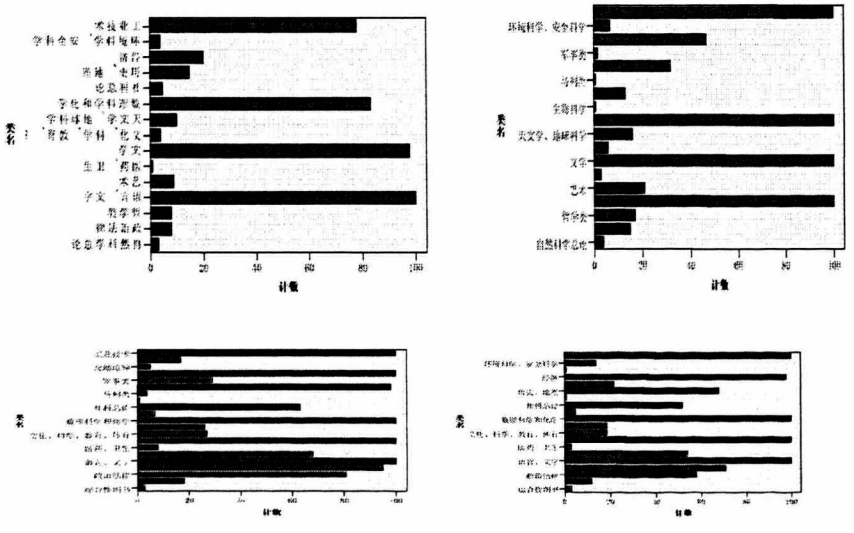

图 5-11 04~07 年图书类别分布情况

将02年到07年我校6年来图书借阅的分布图进行比较可以发现，我校各专业学科在近几年来得到了充足的发展，由02年、03年工业技术、数理科学和化学、语言文字和文学类图书借阅量领先到05年工业技术、经济、数理科学和化学、文学、语言文字、历史地理、哲学齐头并进。政治法律、艺术、社科总论、军事类、教育类等图书被利用率及被利用总量也在不断攀升。06年、07年，也只有马列类、交通运输类、医药卫生及综合类图书利用率较低，这与我校主体读者的成分有关。

从近5年来图书馆的图书借阅变化的趋势也可以看出我校办学的综合层次也在不断提高，读者的阅读需求也向全面化发展。

由以上分析可以得出：

（1） 数理科学、化学、语言文字类图书，应排在醒目、相近的位置，便于读者查找借阅。

（2） 工业技术类与文学类可利用他们的关联性，将其在网页查询中进行相关关联，主动引起读者的注意和兴趣，扩大读者的阅读范围，为提高读者的综合素质做出贡献。

（3） 哲学与科教文类图书应该引起重视，图书馆可以针对这类图书举办主题读书节等活动，来增进读者的阅读需求。

5．4 读者行为研究

1 以同一单位或部门的读者为对象

以02级博士读者群的借阅图书类别进行分析。

博士读者，02级借阅读者共12人，借阅总册数为54册，以图书类别为字段进行分布分析，如图5-12，可见图书主要集中在科图法的56.4大类和中图法的P4大类，对照分类表可知，均为大气科学（气象类）图书。

再考察时间特性，以年份和月份为字段生成网络图，如图5-13，可以发现该类读者在04年借阅量最大，03年稍低，02年、05年在校时间只有一个学期，所以借阅总量低。其中，02年11月、03年6月、9月、04年2月、6月、11月、05年1月借阅量最大。

再以同样的方法研究03级该类读者发现，除了大气科学类图书借阅量

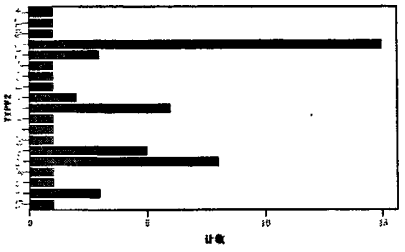

图 5-12 02 级博士借阅图书分类

图 5-13 02 级博士借阅时间网络图

很高外，另外中国文学和外语类图书借阅量也较高。借阅时间分布上 03 年借阅量最高，04 年、05 年借阅量相当，06 年借阅量稍低。其中 04 年 1 月、3 月、05 年 3 月、12 月、06 年 3 月借阅行为较多。

04 级博士读者借阅量年份分布为 04 年最高、05 年、06 年、07 年逐年递减，其中 04 年 9 月、10 月、11 月、05 年 1 月、9 月、11 月、12 月、06 年 6 月、12 月、07 年 3 月借阅行为较多。

05 级读者借阅年份分布 05 年、06 年借阅量较高，07 年、08 年开始递减，其中 05 年 9 月、11 月借阅量最高。

2 读者借阅的时间特性

在读者借阅历史记录数据中，时间特性可以反映读者借阅规律，时间内容包括年、月、日、时、分、秒。需要将该时间内容转换为单独的列数据，生成重点考察的新的字段包括年、月、日、时。

（1） 读者借阅月份与图书大类的分析

以最近的 07 年的读者数据为研究对象。

首先，以类别和所有的月份为输入属性，生成的多重散点图，可以直观地发现各类图书的借阅排名为：T、H、I、O、F、K、B、D、C、P、J、G、X、E、Q、R、N、A、Z、S、U、V。具体情况如下：

① T 类图书为全年借阅量最高,，受欢迎月份排名为 12、9、5、11、10、6，7、4、8，3、1、2。

② H 类排名第二，受欢迎月份顺序为 11、10、9、12、5、6、7、4、8、3、1、2。

③ I 类排名第三，受欢迎月份顺序为 11、10、12、5、9、6、4、7、8、3、1、2。

④ O 类排名第四，受欢迎月份顺序为 10、12、11、9、6、5、4、7、8、3、1、2。

⑤ F 类排名第五，受欢迎月份顺序为 12、6、11、9、5、10、4、7、8、3、1、2。

从所有读者中随机抽样选取一组数据建立模型，得到读者借阅量按照月份顺序排列如

图 5-14 所示，9、11、6、10、8、4、5、1、3、12，2 月与 7 月为假期，借阅量为零。然后再以 07 年的读者借阅数据为学习样本，从中选取 2000 条数据进行 100 次分析比较，最终得出一致的结果如图 5-15 所示，9、11、10 三个月借阅量最大，5、6 两月次之，7、12 月借阅量相当，2、3、1 三个月借阅量最低。

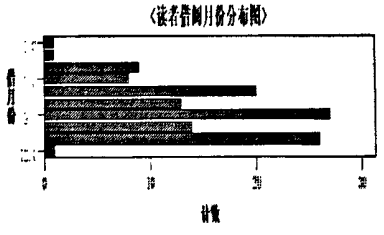

图 5-14 读者个例月份分布

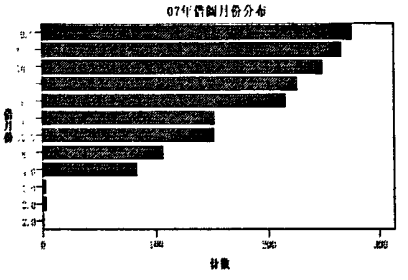

图 5-15 07 年抽样数据月份分布

可以发现，无论是个例还是大量数据的统计，9 月为全年借阅量最大，1 月为全年借阅量最小。

这说明 9 月为新学期的开始，经过一个较长的暑假的休息，同时也开始了新学年的学习，读者学习动力充足，且精力充沛，借阅量最高；1 月为学年上学期的结尾，新年和即将而来的春节假期，都让读者无心阅读，同时大部分读者也都在为期末考试做准备，而无暇阅读除了考试科目以外的书籍。

针对这一情况，图书馆可以在 9 月，充分利用读者学习的热情，将各学科的代表性书籍在显眼的地方进行集中的展示，让读者一走进图书馆就感觉到浓厚的学习氛围，同时也对读者的专业学习进行科学的引导。

1 月，读者的心思较为涣散，对专业知识的学习已经出现疲倦，此时图书馆可以加强读者知识互补的学习，对工学、理学的读者加强文化、艺术类学习的引导，而对文学、艺术类读者可适当进行一些技术基础知识的学习引导。

10、11 月借阅量仍较高，5、6 月借阅量中等，7、12、8、4 四个月借阅量开始下降，1、2、3 月借阅量极低。这说明学年中读者上学期学习情绪高涨，一直维持到 12 月开始减弱，到了 1 月借阅量降低，这可能与一月的假期以及学期末的考试有关，随着寒假的来临，图书馆开放时间的缩短，2 月借阅量依然很低。一直到 3 月，新学期的开始，学科老师的阅读指导以及读者自身的学习欲望，使得图书馆的借阅量总体开始上升，4、5、6 三个月读者需求平稳增长，7、8 月为暑假，读者需求依然保持。

（2） 图书借阅月份分布

首先来分析一下全年读者的图书借阅分布情况。T、H、I、O、F 五类图书借阅量高居榜首， K、B、D 三类图书次之，U、V 图书借阅量最低。现在以 07 年读者借阅记录数据为样本，进行数据审核，选出其中重要性高的前 2000 条数据进行分析。结果如图 5-16，H 类图书排名最前，TP 类、O 类图书也较高；K、B、C、J 类图书也较受欢迎。另外，利用选择节点，对 9 月的图书借阅分布（如图 5-17）进行分析发现：图书分类与全年图书借阅分布完全一致。说明 9 月的借阅数据决定着整个数据集中数据的分布。

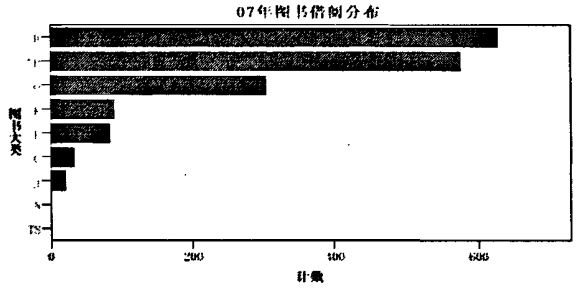

图 5-16 07 年读者阅读图书分布图

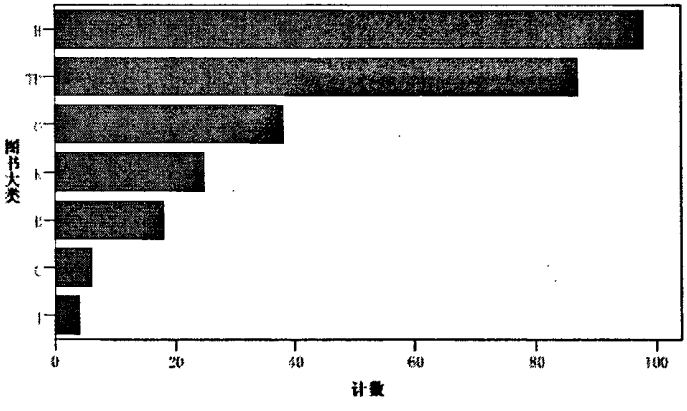

图 5-17 07 年 9 月读者借阅图书分布

再看看 07 年全年图书借阅数据的月份图书类别分布。将每个月份的数据都构造成借阅量与图书大类的网络图，如图 5-18 所示，可以清晰的看出各类图书的图书借阅率（借阅量/总借阅量）排序。

1 月、2 月，3 月 I 类借阅量最高，占总借阅量的 30.4%，T、H、O 类借阅相当，约占 16%到 18%左右，J、B、K、F、C、D、E 类占 1%到 4%左右，其他各类图书借阅量几乎为零。可以说明在 1 月读者对文学类图书感兴趣，图书馆可以举办一些文学类图书读后感征文活

动，进一步提高读者的阅读质量以及其他读者对文学类图书的兴趣；工业技术、语言文字、数学类图书的借阅者主要集中在某些专业读者群，可针对读者的专业情况向其推荐相关书籍；马列、医药、农业、交通等类借阅量极低，可以很容易的发现其借阅者一直为某些部门的个别读者，图书馆可以将其列为专门服务对象，定期向其推荐相关图书。

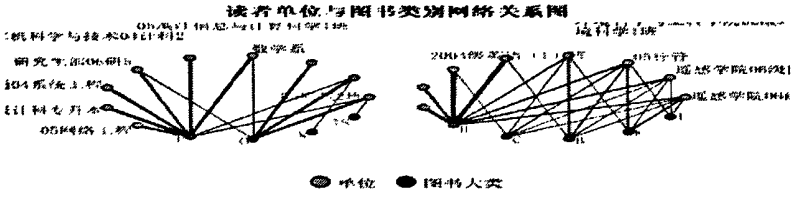

图 5-18 07 年图书借阅量与图书大类的网络图

图书馆可利用这种情况，安排 2、3 月份举行适当的主题读书活动，扩大读者的阅读范围，4、5 月份，将与各类考试相关的图书进行醒目排架，方便和激励读者。

（3） 借阅记录数据表分析

以读者借阅记录表为研究对象，分别进行下列分析：

① 以年份和馆藏地为频数分析对象，SPSS 生成的频数分布表（如表 5-1）和分布图如 5-19 所示。

频数分布表的第一列显示频数分析变量的变量值，第二列是相应变量值的频数，第三列是百分比，第四列是有效百分比，第五列是累计百分比。

柱形图的纵坐标表示频数，可以通过柱形的高低来比较各年的频数特征。

由以频数表和柱形图得到的分析结论如下：

首先，本次统计的总数为2067555条记录，借阅年份的状况是2005年、2006年借阅数量相当，位列第一，由02年到06年借阅数量基本递增，07年有所下降。百分比与有效百分比相同，说明所选取的变量中无缺失值。

其次，馆藏位置借阅量分布如图 5-20 所示，对照馆藏地代码可以看出社会科学处 5 年内借出总量排第一，自然科学借书处第二，文学借书处与综合借书处相当，排第三。而 08 年新开设的新书借书处也有一定的借阅量，这说明此借书处与其他借阅处相比很受欢迎，外文借书处和期刊、报刊借阅处借阅量相当低。

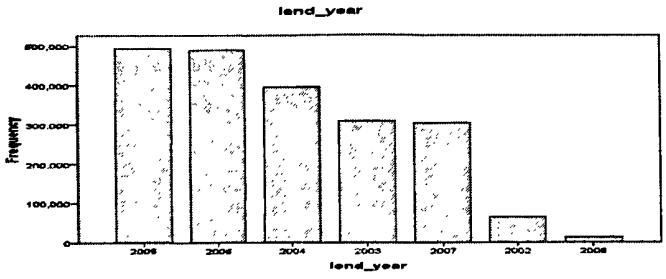

图5-19 年借阅总量分布图

表5-1 03年到08年图书借阅频数分布表

		Frequency（频数）	Percent（百分比）	Valid Percent（有效百分比）
年份	2003	308331	14.9	14.9
	2004	395545	19.1	19.1
	2005	493595	23.9	23.9
	2006	488944	23.6	23.6
	2007	303435	14.7	14.7
	2008	13485	.7	.7
	Total	2067555	100.0	100.0

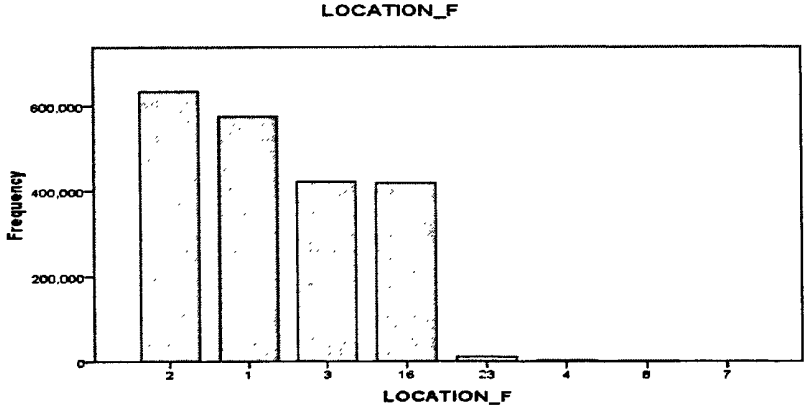

图5-20 各借阅处借出图书总量分布图

② 以月份、日期、小时为变量进行统计分析

此时选取02年9月到07年8月五年的数据为研究对象，这样才能保证所研究的时间规律具有一定的准确性。得到分布图如图5-21、5-22、5-23所示：

可以看出全年中，9~12月借阅总量大于3~6月，1、2月借阅总量大于7、8两个月，其中11月借阅总量最大。说明每个学年的上学期读者借阅兴趣浓厚，下学期总体有所下降；同时节假日对读者的阅读兴趣影响较大。

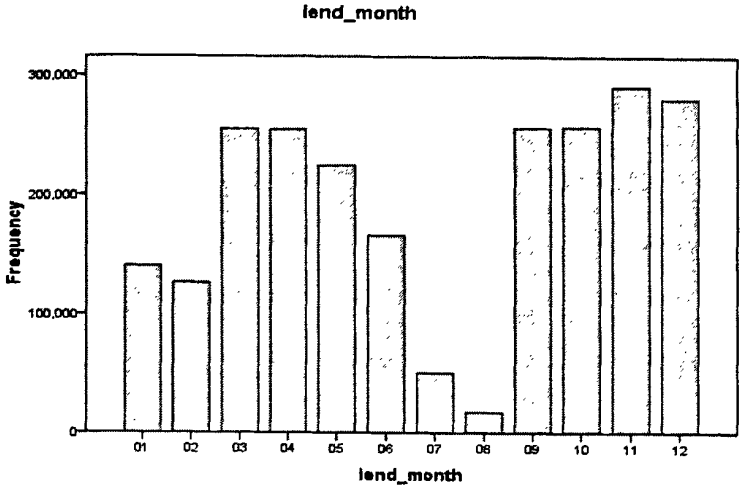

图5-21 5年内平均每月借阅量分布图

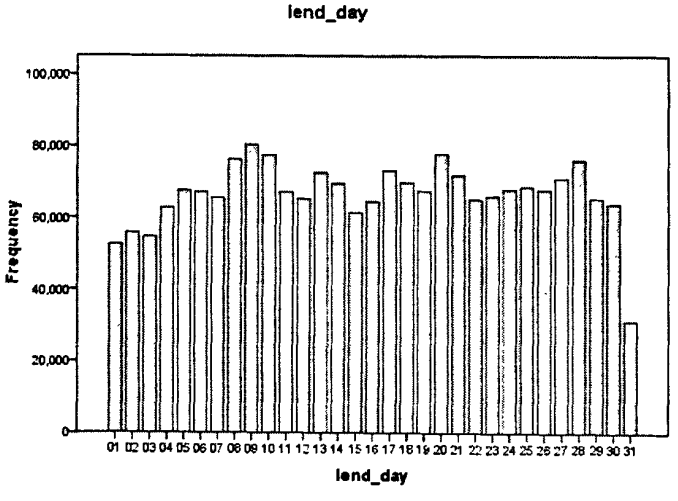

图5-22 5年内平均每日借阅量分布图

在每月中，借阅量呈波形分布，9、20、28日出现借阅最高峰，1、2、3、31日即月首

和月尾出现最低峰，其他日期借阅量相当。这是由于每月初几天出现的节假日较多，而一年中有31日的月份也仅有7个，所以其借阅量较低；10日、20日、30日左右出现借还高峰应该与图书馆设定的可借阅周期有关，研究的数据中大部分读者借阅周期为40天，如果希望能平均每日的工作量，可以采取为不同的读者设定不同的借阅周期来进行调节。

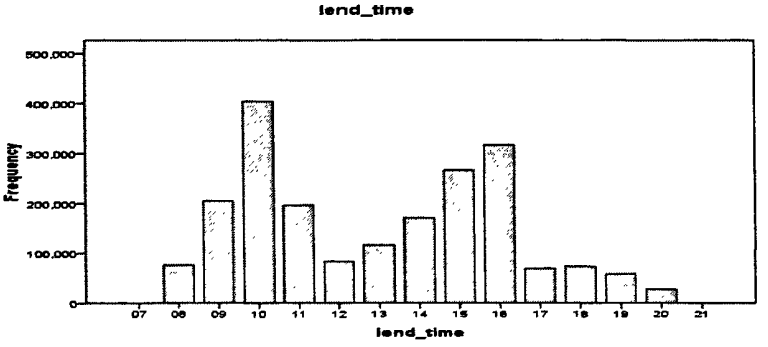

图5-23 5年内平均每时借阅量分布图

在每天图书馆开放时间中，借阅量呈较大波形分布，上午10点和下午4点为借阅最高峰；上午9点、11点和下午2点、3点借阅量为次高峰，其余时间借阅量相当。这主要与图书馆开放时间有关，选取的数据中，图书馆很多时间都是8点开放，16：30借阅处闭馆。同时，上午10点和下午15：45为我校下课时间，这是借阅量达到高峰的主要原因。

③ 以年份、月份为变量进行二维列联表分析

选择月份为列变量、年份为行变量，并输出包含行百分比、列百分比、总百分比的两维交叉列表（如表5-2）和频数分布柱形图如下图5-24所示：

该表为观测频数，依然是包含所有的有效记录数。

表5-2 年份月份变量生成的二位列联表

	Cases（实例）					
	Valid（有效）		Missing（缺失）		Total（总数）	
	N	Percent	N	Percent	N	Percent
lend_year（借阅年份）* lend_month（借阅月份）	2320770	100.0%	0	.0%	2320770	100.0%

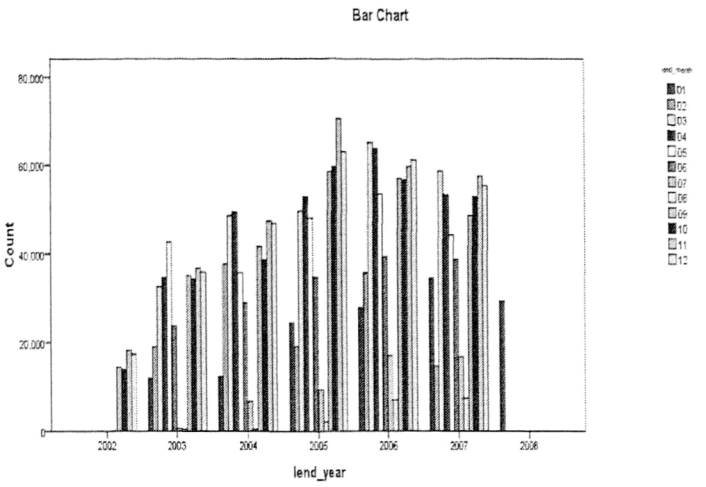

图5-24 以年份、月份为变量进行二维列联表分析频数分布柱形图

以年份来细分发现，每年11月份借阅量最大的发生频率较高，8月借阅量最低的发生频率也较高。7、8月为暑假期间，也是每年图书馆资源利用的分水岭，随着06年、07年我馆暑假期间连续开放，图书资源利用律也有所上升。1、2月的借阅量的高低变换主要取决于寒假的起止日期，3月到6月、9月到12月的每年借阅量稳步增长，这个结果与所有数据汇总研究的结果一致。

（4） 读者信息数据表分析

由于我校读者的信息录入一直采用人工输入的方法，而不是批量数据的导入，这直接导致读者信息数据具有不完整、不一致、错误、缺失等缺点。首先利用SPSS Clementine对该表数据进行审核，去除所有的空白字段及无效字段，然后在SPSS中利用COMPUTE生成我们分析需要的新字段，包括读者年级、读者专业、读者工作性质等。因为我馆从02年开始使用汇文系统进行数字化管理的，选用99级到目前的注册读者数据为研究对象，删除该范围以外的读者。

① 按照年级对读者的借阅总量进行分类汇总，得到的结果如表5-3。

由表可以看出，03、04级读者平均阅读量相当，且最大；99级阅读量最低，因为系统里只有读者在02年8月到目前在校期间的数据，那么99、00、01级的读者借阅数据均不完整，所以下面当以年级为研究对象时将重点选取03、04级读者的借阅数据为研究对象。

表5-3 各年级读者借阅量表

Grade（年级）	TOTAL_LEND_QTY_MEAN（平均年总借阅量）	N_BREAK（实例样本数）
99	12．4	1915
07	16．91	135
06	42．71	108
00	43．4	3409
教师	61．96	4497
01	62．51	3883
05	63．67	5480
02	74．84	4388
03	101．52	5066
04	103．47	2907

② 按照专业、部门对读者的借阅总量进行分类平均，得到部门与平均总借阅量的顺序如表5-3。

表5-3 各部门读者借阅总量均值表

博士后	外部	信控	计管	防雷	应电	博士	机关	滨江	经管
0.0	2.3	3.5	5.20	8.6	9.6	14.9	33.3	34.8	35.8
会计	测控	职工	经贸	信工	数学	网络	声像	转本	电子
57.8	60.7	63.5	64.3	66.3	68.9	69.7	69.8	73.3	73.9
行管	环境	通信	英语	教学	法学	电信	防雷	商务	气象
74.0	75.8	79.4	79.8	81.1	84.0	84.7	86.0	86.7	88.1
信息	研究	公管	应气	环境	物理	人力	信管	旅游	单招
89.0	89.0	89.1	91.7	94.2	94.3	94.5	94.5	94.7	95.0
自动化	市场	化学	系统	农资	应用	高职	地理	日语	中文
95.4	96.0	96.9	98.3	102.4	103.8	106.2	110.6	111.6	113.4
海洋	遥感	生态							
115.5	119.0	119.3							

由表可以看出五年内读者平均借阅次数最高的为中文、海洋、遥感、生态类读者；平均借阅次数最低的为博士后、外部读者、信控、计管的读者；平均借阅次数越低，说明该类读者中惰性读者数目越大。该数据因为其读者部门的来源数据不精确，所以这里显示的

只是大致的借阅分布。

③ 对读者借阅总次数进行频数分析，分布图如图5-25。

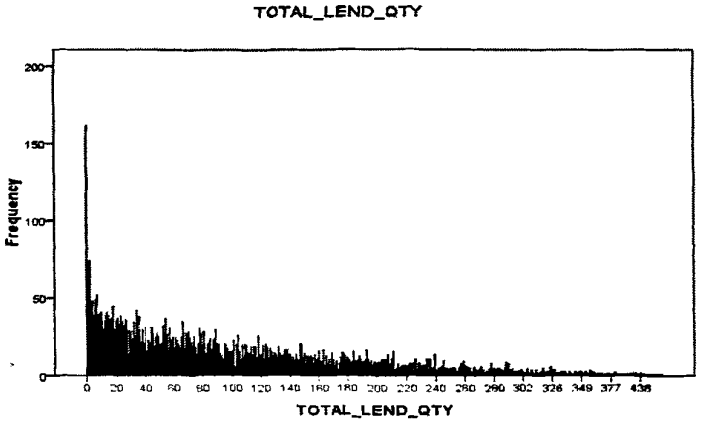

图5-25 读者借阅总次数频数分析图

由图可以发现，借阅为0次的读者数量竟然远高于其他借阅次数的读者数量，大多数读者的借阅次数集中在60到200之间，根据这一分布，可以由读者的借阅次数从59以下、60~200，200以上3个等级将读者进行分组来研究，分别为惰性读者，忠实读者，活跃读者三个群组。

（5） 图书借阅信息分析[24~28]

将图书信息表与读者借阅信息表相关联，得到每本图书借阅次数信息表。主要包括图书MAC记录号、条码号、分类号、总借阅次数等。

将同一本书的不同复本的借阅总量进行汇总，得到每本不同数的五年内总借阅量，根据总借阅量进行升序排序，可以清楚地得到每种图书利用率的高低。

借阅量最大的为2515次，书名《动力气象学-修订本》，作者杨大升等，气象出版社，1980年版，分类号为56.43/11-2；次之为2503次，书名《高等数学（同济第四版）》，分类号为013/32:3；第三位为《高等数学(同济五版)考点精析习题全解》2236次，分类号为013-44/80:1；第四位为《鹿鼎记》2109次，分类号为I247.58/52；第五位为《楚留香传奇系列》2006次，分类号为I247.58/26；再看借阅量总次数为零的图书种类竟然很多，利用总借阅量作为变量进行频次分析，结果如图5-26所示：

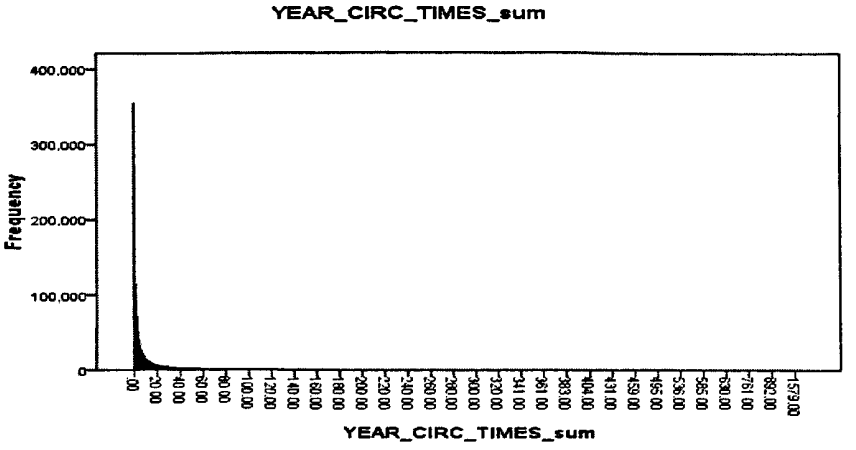

图5-26 图书年总借阅次数频次分析图

从图中可以明显看出，借阅量低的图书占总图书的比例非常大，按照借阅次数小于20作为“惰性图书”来重点进行分析。

以上对读者的信息数据与借阅数据以及图书的利用数据进行了大致分析，了解了目前我校图书馆的总体利用状况，下面在以上分析的基础上，来建立相关模型进行进一步的数据挖掘。

5.5 数据挖掘实例研究

5.5.1 应用决策树算法对图书馆的低利用率的读者特性进行分析

1 背景

读者在图书馆注册以后，很少到图书馆或者甚至从未走进图书馆，这是图书馆资源利用效率低的主要因素。对于普通的高校图书馆来说，每年新增读者人数约 5000 到 10000 人左右，真正能够充分利用图书馆资源的读者不到 20%。为了吸引读者到馆阅读所需要的成本，要比为读者增设一门课程的平均成本低的多。如果能够大大地调动读者的积极性，这会对于高校培养综合性高级人才做出很大的贡献。

在我国的一些著名学府，如北京大学、清华大学、南京大学、东南大学等高校，图书馆的学习气氛相当浓厚，每日到馆读者人数也非常多，图书馆所藏的各种资源的利用效率也较高。相比一些地方普通高校，在图书馆的建设方面也投入了大量的人力、物力、财力，但是有很多资源几乎无人问津。如一些期刊资料阅览室每日的到馆读者不到 30 人，这不仅

浪费了图书馆为之付出的大量资金，还有人力、空间，而且，这也从侧面反映了高校的科研氛围不足，这对于从根本上提高学校的办学水平是很大的障碍。

2 分析目的

根据借阅率高的读者和借阅率低的读者的性质和借阅行为，进行挖掘分析，建立低利用率读者的预测模型，分析哪些读者的惰性概率较大，他们的借阅行为如何，造成其惰性的其他相关因素是什么等[33]。

可以为图书馆管理与决策人员制定相应的策略和吸引读者到馆提供决策依据，并预测在该策略下读者利用图书馆的情况。

3 挖掘过程

在本次分析中，综合采用了 CRISP-DM 和 SEMMA 的步骤。整个挖掘过程如图 5-27 所示。

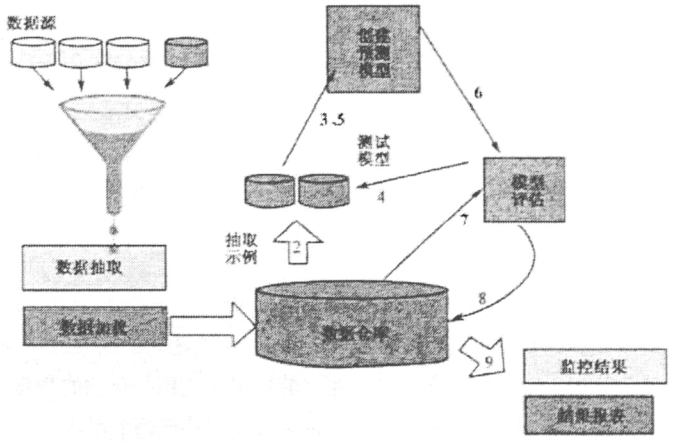

图 5-27 数据挖掘过程图

整个数据挖掘过程的详细描述如下：

① 数据准备。首先从现有的数据库中提取出低借阅率读者借阅记录数据表，我们以总借阅次数等于 0 的读者借阅记录为研究对象，确定读者证件号、姓名、性别、单位、职业为所要建立的模型的主题相关的数据项，抽取数据，生成衍生变量，不断反复，直至得到全部“满意”的数据变量。最后，将这些衍生变量集合成一个集合文件，包括每个读者的各种资料，存入“惰性读者数据集市”中。

② 取样。从所得到的衍生变量集合文件中抽取样本，包括不同借阅次数，以及不同大小的多份样本，作为训练模型及测试模型使用。抽取出来的样本直接传送到数据挖掘服务器上。

③ 建立模型。选择决策树技术，利用训练数据集来训练并建立模型。

④ 验证模型。验证数据对于已经建立的模型来说是全新的数据，但是该数据集要经

过不断的检查处理，只到可以与已建立模型具有一个大致相同的模型准确性。

⑤ 模型评分、监测。将数据挖掘建立的模型输出成为一个 C 语言的子程序，并传回数据挖掘主机，由主机上的模型评分主程序来调用。

利用模型计算读者的各种指标，还要计算出与该模型相关的性能度量值，来比较不同的模型，从而选出最好的。

4 数据源的描述

（1） “惰性读者”信息表，定义总借阅次数等于 0 次的为“惰性读者”，从读者借阅信息表中导出需要的数据，其中包含的字段有：读者证件号、姓名、年级、单位、专业、总借阅量、读者注册日期、读者注销日期、读者类型、读者借阅规则、平均借阅量、总样本数等。

（2） “惰性读者”借阅历史表、借阅规则表、读者类型表，所含字段包括读者所借图书的条码号、分类号、学生等。

5 使用决策树方法来建立低利用率读者模型[39-41]

（1） 利用借阅次数等于 0 次的读者信息表，共包含 2334 条记录，我们利用 C5. 0 算法构造预测决策树或规则集，选择单位作为目标字段，年级输入字段，生成的模型结果如下：

grade in [""] [众数: 教学] => 教学

grade in ["00"] [众数: 计管] => 计管

grade in ["01" "99"] [众数: 计应] => 计应

grade in ["02" "03" "05"] [众数: 滨江] => 滨江

grade in ["04" "06" "07"] [众数: 成教] => 成教

该模型将年级字段分为5个节点集合，在每个集合中都得出其中借阅量为零最多的单位。可以发现教学类读者中惰性读者较多，在目前在校学生中，大一、大二、大四学生中，成教读者的惰性成分较大，大三的读者中滨江的读者惰性成分较大。

（2） 对借阅次数等于零的读者，将年级作为输入字段，单位作为目标字段，利用QUEST算法构造预测决策树，结果如图5-28。

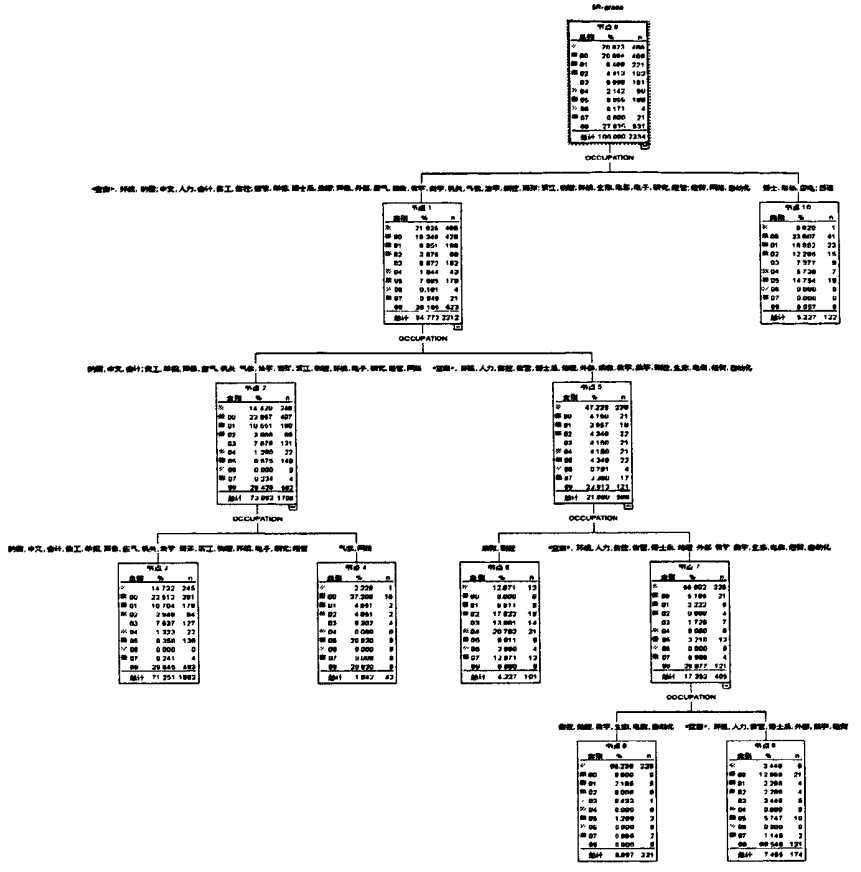

图5-28 QUEST算法构造预测决策树结果

生成的对应规则如下：

OCCUPATION in ["" " 环境" " 防雷" "中文" "人力" "会计" "信工" "信控" "信管" "单招" "博士后" "地理" "声像" "外部" "应气" "成教" "教学" "数学" "机关" "气象" "法学" "测控" "海洋" "滨江" "物理" "环境" "生态" "电信" "电子" "研究" "经管" "经贸" "网络" "自动化"] [众数: 99]

OCCUPATION in [" 防雷" "中文" "会计" "信工" "单招" "声像" "应气" "机关" "气象" "法学" "海洋" "滨江" "物理" "环境" "电子" "研究" "经管" "网络"] [众数: 99]

OCCUPATION in [" 防雷" "中文" "会计" "信工" "单招" "声像" "应气" "机关" "法学" "海洋" "滨江" "物理" "环境" " 电子" "研究" "经管"] [众数: 99] => 99

OCCUPATION in ["气象" "网络"] [众数: 00] => 00

OCCUPATION in ["" " 环境" "人力" "信控" "信管" "博士后" "地理" "外部" "成教" "教学" "数学" "测控" "生态" "电信" " 经贸" "自动化"] [众数:]

OCCUPATION in ["成教" "测控"] [众数: 04] => 04

OCCUPATION in ["" " 环境" "人力" "信控" "信管" "博士后" "地理" "外部" "教学" "数学" "生态" "电信" "经贸" "自动化"] [众数:]

OCCUPATION in ["信控" "地理" "教学" "生态" "电信" "自动化"] [众数:] =>

OCCUPATION in ["" " 环境" "人力" "信管" "博士后" "外部" "数学" "经贸"] [众数: 99] => 99

OCCUPATION in ["博士" "市场" "应电" "日语"] [众数: 00] => 00

决策树可以区分不同的读者群组以及每一群组的潜在懒惰因素。图中就指出了某个懒惰群组的共同特性。在这个群组中，有80%的惰性读者。这些共同特性包括：

* 博士、市场、应电、日语类的学生中00级读者较为懒惰；
* 环境、人力、信管、博士后、外部读者、数学、经贸类读者中99级较为懒惰；
* 成教、测控专业04级读者较为懒惰；
* 气象、网络类专业00级读者较为懒惰。

6 结果分析

读者惰性模型的分析，以及决策树的输出结果，会产生多个惰性读者的群组。对于每一个群组特性所隐含的意义以及合理与否，需要与读者的其他情况结合来共同检查与诊断。

该模型因为掌握的读者资料相当有限，这个严重限制了研究的深度。但是作为各部门可能出现的惰性读者群体还是可以与其他的分析维度结合一起分析，来预测其他读者中可能成为惰性读者的群体。

5. 5. 2 应用聚类分析对高校图书馆读者进行细分

现代的高校图书馆服务已迫切要求由过去的被动服务要改变为主动服务，读者中心论要求对读者进行全方位的认识，进行一对一的准确的个性化服务。它的前提就是对读者群进行细分。

（1） 认识了解读者

主要包括：

① 读者分为哪几种类型？

② 每一类读者有什么不同的借阅特征？

③ 读者群有什么流动趋势？包括阅读图书类别的转移和由勤奋转为惰性等等。

④ 一个类新的图书或服务应该向哪一个群体的读者推荐？

⑤ 如何鼓励一个特定读者群的阅读行为，使他们能最高效的利用图书馆资源。

⑥ 不同类型的读者在某个特定时刻的需求是什么？

⑦ 同一个读者在不同时刻的需求是什么？

（2） 传统的读者分类

大多数的高校图书馆对读者的分类是按照读者的身份来进行的，分为：教师、职工、学生、研究生、校外读者 5 类。但是这种划分只能说明到馆读者的大致成分，没有更进一步的信息，这样的划分的缺点如下：

① 过于笼统、粗糙 这样划分，同样是教师类别，他们之间还存在很大的差别，学生与学生之间的区别也很大。

② 没有与读者的借阅行为特征相结合，很难进行准确的图书推广活动以及个性化服务。

（3） 进行读者细分所依赖的数据

在读者细分时，读者的行为特征是主要的特征，读者的身份类别为辅助特征。读者的行为特征主要包括读者借还特征、阅览特征，借还特征包括：借书时间、还书时间、借书类别、借书数量等。阅览特征包括：所到阅览室类别、进入时间、离开时间等。

读者的身份特征主要包括读者的类别、专业、部门、年级等。

从表中选择读者信息表（READER）、借阅历史表（LEND_HIS）、图书记录表（ITEM）中需要的字段创建视图（DUZHEXIFEN），生成包含研究所需的所有数据项的表。SQL 执行语句如下：

```
CREATE OR REPLACE VIEW DUZHEXIFEN AS
SELECT LIBSYS.ITEM.CALL_NO, LIBSYS.ITEM.PROP_NO,
   LIBSYS.ITEM.TOTAL_CIRC_TIMES, LIBSYS.LEND_HIST.CERT_ID_F,
   LIBSYS.LEND_HIST.LEND_DATE, LIBSYS.LEND_HIST.LOCATION_F,
   LIBSYS.LEND_HIST.RET_DATE,
   LIBSYS.READER.DEPT, LIBSYS.READER.NAME,
   LIBSYS.READER.SEX, LIBSYS.READER.TOTAL_LEND_QTY
   FROM LIBSYS.ITEM, LIBSYS.LEND_HIST, LIBSYS.READER
   WHERE LIBSYS.ITEM.PROP_NO = LIBSYS.LEND_HIST.PROP_NO_F
   AND LIBSYS.LEND_HIST.CERT_ID_F =LIBSYS.READER.CERT_ID
       WITH READ ONLY;
```

收集到的数据量很大，达 300 多万条，我们从中随机抽取 10%的数据进行研究。

（4） 进行读者细分所采用的数据挖掘技术[42]

进行读者细分采用的是聚类分析方法。

聚类分析是一种统计分析方法，它通过将实体的特征进行归一化处理，对具有多特征的实体进行分组，使每一组之间的差异最大化，组内的差异最小化。

将抽取出的数据选取图书总借阅次数、借阅读者证号、图书馆藏地、读者总借阅次数、借阅年、月、日作为输入字段构造 K-Means 聚类模型，结果如图 5-29。

总共将读者细分为 5 个聚类。

图 5-29 读者特征的聚类分析

聚类 1 为 在 9 月的 8 日左右在 4 借阅处借阅的图书，所借图书累计使用次数达 20 次。

聚类 2 为在 4 月 16 日左右下午 1 点借阅的读者，所借图书累计使用次数达 29 次。且借阅地点集中在 16 借阅处。

聚类 3 为在 3 月 14 日左右下午 1 点借阅的读者，所借图书累计使用次数达 18 次。且借阅地点集中在 1 借阅处。

聚类 4 为在 7 月 21 日左右下午 1 点借阅的读者，所借图书累计使用次数达 17 次。且借阅地点集中在 2 借阅处。

聚类 5 为在 10 月 22 日左右下午 1 点借阅的读者，所借图书累计使用次数达 20 次。且借阅地点集中在 6 借阅处。

将读者根据其自身借阅特性进行细分，可以为图书馆工作者有针对性地为各类不同的读者提供个性化服务。

这里还可以选择其他不同字段按照不同指标将读者群作进一步细分。

5．5．3 基于矩阵的数据挖掘算法在高校图书馆系统中的应用

1 基于矩阵的数据挖掘算法基本思想[43-467]

检测侯选项目集是 Apriori 系列算法的性能瓶颈，针对 Apriori 算法的不足，根据频繁项集的性质和二进制逻辑运算的基本思想，这里提出基于矩阵的数据挖掘算法。

（1） 选择读者借阅信息表进行预处理

重点选取读者借阅证号作为事务编号 Tid ，读者借阅图书类名信息作为项 Items。设最小支持度为 33%，则由支持度 Sup(A→B)=P(A∪B)得出频繁项目集至少要出现 3 次。

以 05 级博士的借阅记录来进行分析，将其读者借阅信息进行如下处理如表 5-4。

表 5-4 读者编号与借阅信息表

事务编号 Tid（读者）	项 Items（借阅信息）
T1	P4，TP7
T2	TP3
T3	I2，I2
T4	P4
T5	56
T6	I2
T7	56，H19，O1，P4，TP3，TP7
T8	I5，TP3
T9	I2，P2
T10	P4
T11	O1
T12	P3，P4，TP3
T13	39，H31，X8
T14	TP3
T15	P4

首先为每个项目建立一个比特向量 BV_i，事务 Tid 的长度决定比特向量 BV_i 的维数，每一事务根据编号顺序在 BV_i 中对应一个位置，如果一个项目 i 在第 j 个事务出现，那么对应的比特向量 BV_i 的第 j 位置“1”，反之则置“0”，BV_i 中“1”的数目即是包含项目 i 的事务数。表中 BV_1(P4)=（100100100101001），BV_2(TP7)=(100000100000000)，BV_3(TP3)=(010000110000010)，BV_4(I2)=(001001001000000)，BV_5(56)=(000010100000000)，BV_6(H19)=(000000100000000)，BV_7(O1)=(000000100010000)，BV_8(I5)=(000000010000000)，

BV_9(P2)=(000000001000000)，BV_{10}(P3)=(000000000001000)，BV_{11}(39)=(000000000000100)=BV_{12}(H31)= BV_{13}(X8)。项目 1 为 6 次，项目 2 为 2 次，项目 3 为 4 次，项目 4 为 3 次，项目 5 为 2 次，项目 6 为 1 次，项目 7 为 2 次，项目 8、9、10、11、12、13 均为 1 次，根据所设最小支持度频繁项目集至少要出现 3 次，则频繁 1-项目集 L_1={1，3，4}。

（2） 数据挖掘在图书馆中的应用

如果 BV_i 与 BV_j 中“1”的个数不小于最小支持度阈值，则项集{i,j}是一个频繁 2-项集。由此，构造一个（m-1）Xm 行的频繁项集支持矩阵 ARM，其中 m 是事务数据库中频繁项目的个数，矩阵初值为 0。对于上面得到的比特向量，当 i<j 时，两两进行逻辑与运算，N_{ij} 存放统计运算结果中“1”的个数，如果 N_{ij} 的数目不小于最小支持数，则 ARM[i,j]= N_{ij}，同时将 ij 所代表的项目元素放入 L_2。表 5-5 为频繁项集支持矩阵，得到的频繁 2-项集 L_2={{1,3}}。

表 5-5 频繁项集支持矩阵

与运算结果	BV_1	BV_3	BV_4
BV_1	0	1	0
BV_3	0	0	0

基于频繁项集支持的矩阵 ARM，我们用频繁 K-项集的最后一项来扩展项目集为（k+1）-项目集。设{$i_1,i_2,\ldots,i_k$}∈L_k，如果在矩阵图中 ARM[i_k,i_u]≥min_sup(i_k<i_u)，并且 ARM[i1,iu]、ARM[i2,iu],…,ARM[ik-1,iu]>=min-sup ,那么{i1,i2,…,ik}被扩展为(k+1)-项集{i1,i2,…,ik,iu}，如果 BVi1∧BVi2∧…∧BVik∧BVu 中“1”的数目不小于最小支持数，则{i1,i2,…,ik,iu}是频繁(k+1)-项集。算法迭代执行直到不能产生新的频繁项集为止。这样我们最终得到频繁 2-项集 L_2={{1,3}}。

I1，I3 分别代表图书分类号 P4，TP3，L_2={I1,I3}是通过矩阵数据挖掘算法得到的频繁项目集。再给出置信度为 65%对频繁项集产生强关联规则，并以期望置信度 Coverage(A→B)=P(A)和作用度 lift（A→B）=P(B|A)/P(B)=Confidence(A→B)/Coverage(A→B)加以判断分析。经筛选可以得出强关联规则如表 5-6。

表 5-6 强关联规则表

关联规则	置信度	期望置信度	作用度
I1→I3	33%	2/5	0.83
I3→I1	50%	4/15	1.875

2 结果分析

（1） 规则{I1→I3}，置信度为 33%，作用度为 0.83，作用度小于 1 说明 I1 和 I3 是负相

关的，这样的关联规则即使置信度和支持度都很高，通常也是无效的；

（2） 规则{ I3→I1}，置信度为 50%，作用度为 1.875，说明 I3 和 I1 是正相关的，意味着博士生读者借阅计算机类图书时含有再借阅气象类图书的趋势。

这样，系统可以根据分析结果，生成推荐页面，将原先被动的服务模式转向可根据用户兴趣提供主动的个性化服务，当用户登陆系统时，该用户就可以得到相关的书目推荐。同时，图书馆也可以根据这一结果在现实中进行个性化推荐服务。

这一算法可以对各个部门的读者进行同样的分析，挖掘其兴趣，为图书馆个性化服务提供科学依据。

本章小结：本章重点从数据挖掘的 CRISP-DM 思想对图书馆系统数据主要从图书、读者、服务三个角度进行数据分析和挖掘，对图书馆的读者借阅行为、图书资源的利用情况分别按照单位、时间等因素进行了探索性数据分析，利用决策树算法对低利用率读者的借阅特性进行了研究，利用聚类算法对读者进行了细分，并对传统的 Apriori 算法进行了改进，利用基于矩阵的数据挖掘算法对某一单位一定时间内的读者借阅行为进行了分析；最后得出了一些有益的结论。

第六章　结论及展望

6.1　结论

本文通过数据仓库和数据挖掘技术对高校图书馆系统数据进行了分析，主要从图书、读者、服务三个角度分别进行了分析，得出的主要结论如下：

（1）　图书管理

①　找出了语言、文字，文学、工业技术、数学 4 类最受欢迎图书，经济，历史、地理，哲学宗教，政治法律类图书受欢迎程度次之。

②　5年来借阅量最大的为2515次，书名《动力气象学-修订本》，作者杨大升等，气象出版社，1980年版，分类号为56.43/11-2；次之为2503次，书名《高等数学（同济第四版）》，分类号为013/32:3；第三位为《高等数学(同济五版)考点精析习题全解》2236次，分类号为013-44/80:1；第四位为《鹿鼎记》2109次，分类号为I247.58/52；第五位为《楚留香传奇系列》2006次，分类号为I247.58/26；

③　2社会科学处5年内借出总量排第一，1自然科学借书处第二，16文学借书处与3综合借书处相当，排第三。而08年新开设的新书借书处也有一定的借阅量，这说明此借书处与其他借阅处相比很受欢迎，外文借书处和期刊、报刊借阅处借阅量相当底。

④　哲学与科教文类图书应该引起重视，图书馆可以针对这类图书举办主题读书节等活动，来增进读者的阅读需求。

（2）　读者管理

①　五年内读者平均借阅次数最高的为中文、海洋、遥感、生态类读者；平均借阅次数最低的为博士后、外部读者、信控、计管的读者；平均借阅次数越低，说明该类读者中惰性读者数目越大。

②　所有的惰性读者分布如下：

* 博士、市场、应电、日语类的学生中 00 级读者较为懒惰；
* 环境、人力、信管、博士后、外部读者、数学、经贸类读者中 99 级较为懒惰；
* 成教、测控专业 04 级读者较为懒惰；
* 气象、网络类专业 00 级读者较为懒惰。

③　根据读者的习惯借阅时间和借阅图书类别将读者细分为 5 类。

（3）　服务管理

① 9、11、10三个月借阅量最大，5、6两月次之，7、12月借阅量相当，2、3、1三个月借阅量最低全学年中，每个学年的上学期读者借阅兴趣浓厚，下学期总体有所下降；同时节假日对读者的阅读兴趣影响较大。

1 月，读者的心思较为涣散，对专业知识的学习已经出现疲倦，此时图书馆可以加强读者知识互补的学习，对工学、理学的读者加强文化、艺术类学习的引导，而对文学、艺术类读者可适当进行一些技术基础知识的学习引导。

10、11 月借阅量仍较高，5、6 月借阅量中等，7、12、8、4 四个月借阅量开始下降，1、2、3 月借阅量极低。这说明学年中读者上学期学习情绪高涨，一直维持到 12 月开始减弱，到了 1 月开始厌学，一直到 3 月，读者才渐渐有了学习的欲望，但是依然没有新学年开始时的充足动力和强烈需求，4、5、6 三个月读者需求平稳增长，7、8 月为暑假，读者需求依然保持。

② 1 月、2 月，3 月 I 类借阅量最高，占总借阅量的 30.4%，T、H、O 类借阅相当，约占 16%到 18%左右，J、B、K、F、C、D、E 类占 1%到 4%左右，其他各类图书借阅量几乎为零。可以说明在 1 月读者对文学类图书感兴趣，图书馆可以举办一些文学类图书读后感征文活动，进一步提高读者的阅读质量以及其他读者对文学类图书的兴趣；工业技术、语言文字、数学类图书的借阅者主要集中在某些专业读者群，可针对读者的专业情况向其推荐相关书籍；马列、医药、农业、交通等类借阅量极低，可以很容易的发现其借阅者一直为某些部门的个别读者，图书馆可以将其列为专门服务对象，定期向其推荐相关图书。

③ 在每月中，借阅量呈波形分布，9、20、28日出现借阅最高峰，1、2、3、31日即月首和月尾出现最低峰，其他日期借阅量相当。

在每天图书馆开放时间中，借阅量呈较大波形分布，上午10点和下午4点为借阅最高峰；上午9点、11点和下午2点、3点借阅量为次高峰，其余时间借阅量相当。

6．2 论文不足及展望

本论文虽然从图书馆主体三要素中得出一些有益的结论，但是还是存在一些不足：

（1） 对读者数据收集不够，仅限于图书馆系统数据，未来可与读者管理部门沟通，能够获得更多的数据。

（2） 我校近 5 年发展较快、变化较大，结果个例色彩较重，缺少其他高校的数据做参考，使得结论的普适性降低。

（3） 只针对图书馆工作展开了部分的讨论，工作中还有更多的内容需要开拓。

(4) 仅限于探讨研究，还未形成最终的产品。

希望未来能够克服以上的不足，继续展开进一步研究。

参考文献

[1] 谢培树,肖冬荣,朱国强. 基于神经网络和遗传算法的变型设计研究. 机械设计与制造,2006（3）,p144-146.

[2] 毛雅琳. 新世纪的图书馆. 中国科学院第十二次图书馆学情报学科学讨论会文集, p34-36.

[3] 孙秋燕, 温尚明. 新世纪的图书馆和情报服务. p37-40.

[4] 张力. 数据挖掘在图文信息系统中的应用. 华东师范大学硕士学位论文 ,2006.

[5] 刘嘉 等编译. 国外图书馆学重要著作选译, 华艺出版社,2002,北京.

[6] 中国科学院文献情报中心 编. 中国科学院第十二次图书馆学情报学科学讨论会文集. 北京图书馆出版社, 2002, 北京.

[7] 陈京民 等编著. 数据仓库与数据挖掘技术. 电子工业出版社,北京, 2002.

[8] [美]Pang-Ning Tan 等著. 数据挖掘导论. 人民邮电出版社, 2006, 北京.

[9] 麦永浩. 数据仓库和数据挖掘方法研究及其在公安信息建设中的应用. 博士学位论文, 2000.

[10] 周蓓. 数据挖掘技术在图书馆系统中的应用研究. 东南大学工程硕士学位论文, 2006.

[11] 任世锦. 基于区间数的不确定性数据挖掘及其应用研究. 浙江大学博士学位论文, 2006.

[12] 邓小梅. 基于数据挖掘的电信客户细分模型研究. 大连理工大学硕士学位论文, 2006.

[13] 陈文文. 图书馆使用者行为模式的数据挖掘研究. 西南大学硕士学位论文, 2007.

[14] Vrushali Khilari IMR-Pune [India]. *Pricing Strategy of KM-Related Services in Nonprofit Organizations(Specifically Library Services).* 2006 IEEE International Conference on Management of Innovation and Technology, P275-279.

[15] Renhao Huang. *Study and analysis of information on the reader's potential dis content in an academic library.* Library Management 2007,28(1/2) p27-35.

[16] 孙卫祥. 基于数据挖掘与信息融合的故障诊断方法研究. 上海交通大学工学博士学位论文, 2006.

[17] 范中磊,潘龙法. 一种基于呼叫中心和数据挖掘的客户数据库模型. 计算机应用研究, 2002（1）,p84-85.

[18] Wendir Bukowitz Ruth L. *Williams.Knowledge Managemen*t field book.

[19] Shearer C. The CRISP-DM model: The New Blueprint for Data Mining. Journal of Data

Warehousing，2000, 5(4), p13~20.

[20] 肖冬荣．自动控制系统可靠性提高及其计算方法．交通与计算机，1991 年 03 期,p46-48.

[21] 韩耀，张春法，刘宁．零售业客户关系管理及数据挖掘的应用研究．情报杂志,2005(11), p55~5.

[22] 肖冬荣，黄静．基于均值、方差和偏度的投资组合模糊优化模型．统计与决策,2006(14), P37-38.

[23] 肖冬荣，朱京，张辉．基于决策指标分类的供需链合作伙伴选择算法．辽宁工程技术大学学报，2002(5),P664-666.

[24] Tracy A. Hurley Carolyn W. Green. *Knowledge Management and the nonprofit Industry, A within and between approach*, Journal of Knowledge management practice, 2005.

[25] 肖冬荣．城市控制论初探．衡阳师范学院学报,1993 年 06 期,p 16-19.

[26] 肖冬荣．系统科学及其当前存在的问题．系统工程理论与实践，1990(5), p1-5.

[27] Shearer C. The CRISP-DM model: The New Blueprint for Data Mining. Journal of Data Warehousing，2000, 5(4),p13~2.

[28] Kalpana Das Gupta, *Library practices for effective management,* Delhi, ILA, 2001.

[29] Madanmohan Rao, *Leading with Knowledge.Km Chronicles:Travelogue I*, New Delhi, Tata McGraw-Hill, 2003.

[30] Chong Siong Choy, *Critical factors in the successful implementation of knowledge management.* Journal of Knowledge management practice,2005.

[31] 飞思科技产品研发中心 编著．Oracle 9i 数据仓库构建技术．电子工业出版社，北京，2003.

[32] [美]Paulraj Ponniah 著．数据仓库基础．电子工业出版社, 2004，北京.

[33] 飞思科技产品研发中心 编著. Oracle 9i 基础与提高．电子工业出版社，北京, 2003.

[34] 段云峰 等编著．数据仓库及其在电信领域中的应用．电子工业出版社, 2003，北京.

[35] V NagaLakshmi, I Rameshbabu *A Security Mechanism for library management system using low cost RFID tags.*

[36] Giannis Tsakonas, Christos Papatheodorou. *Exploring usefulness and usability in the evaluation of open access digital libraries.* Information Processing & Management, Volume 44, Issue 3, May 2008, Pages 1234-1250.

[37] Valerie J Gillet. *New directions in library design and analysis,* Current Opinion in Chemical Biology, In Press, Corrected Proof, Available online 18 April 2008.

[38] Jernej Trnkoczy, Vlado Stankovski. *Improving the performance of Federated Digital Library services,* Future Generation Computer Systems, In Press, Accepted Manuscript, Available online

18 April 2008.

[39] R.Agrawal, R.Sricant. *Fast Algorithms for Mining Association Rules.* Very large Databases, Santiago, 1994, p487~499.

[40] Bleyberg, M.2., Zhu，D.Cole，K.，Bates，D.，Zhan，W.(1999). *Developing an integrate library decision support warehouse.* IEEE international conference on Systems, man, and cybernetics. Vol2, p546-551.

[41] Hinnerburg A, Keim D A. *An efficient approach to clustering to clustering in large multimedia databases with noise.* Proc of 1998 Int Conf Knowledge Discovery and Data Mining(KDD'98), 1998(8): 58~65.

[42] 吕巍,蒋波,陈洁.基于 K-means 算法的中国移动市场顾客行为细分策略研究.管理学报，2005,2(1), p80~8.

[43] George Gigli, Éloi Bossé, George A. Lampropoulos. *An optimized architecture for classification combining data fusion and data-mining,* Information Fusion, Volume 8, Issue 4, October 2007, Pages 366-378.

[44] Jochen Hollmann, Anders Ardö, Per Stenström. *Effectiveness of caching in a distributed digital library system,* Journal of Systems Architecture, Volume 53, Issue 7, July 2007, p403-416.

[45] Fengrong Gao, Chunxiao Xing, Xiaoyong Du, Shan Wang. *Personalized Service System Based on Hybrid Filtering for Digital Library,* Tsinghua Science & Technology, Volume 12, Issue 1, February 2007, p1-8.

[46] Manjunath Lohar, Mallinath Kumbar. Teachers' Attitudes Towards Library Facilities and Information Resources in First Grade Colleges in Shimoga District: A Survey. SRELS Journal of Information Management 07,2007,44(2), p179-206.

[47] Gadi Rothenberg. *Data mining in catalysis: Separating knowledge from garbage,* Catalysis Today, In Press, Corrected Proof, Available online 28 March 2008.

研究生期间发表的论文

[1]. 张海燕，肖冬荣，李诗平. 计算机病毒入侵及对抗技术. 微计算机信息,2008,03(9): p77-78.

[2]. 潘文婵，周杰，张海燕. 非理想功率控制和多用户检测在 WCDMA 高空平台通信系统中的联合应用. 电子技术应用,2008,34(5):p101-103.

研究生期间承担的主要课题项目

[1]. 应用耗散结构理论构建高校电子阅览室管理体系. 校内课题 主持

致 谢

本文是在我的导师肖冬荣教授的悉心指导和严格要求下完成的。肖教授严谨的治学态度和对科学孜孜以求、实事求是的钻研创新精神，使我受益匪浅；肖教授热心真诚、平易近人、虚怀若谷的高尚品德，将在我以后的学习、生活和工作中产生积极而深远的影响。在此，我向我的导师肖冬荣教授表示崇高的敬意和深深的感谢！

在我攻读硕士学位的过程中，张颖超教授、罗琦教授、马杰良副教授、赵英教授、赵远东副教授、郭伟副教授等在我专业学习过程中，给我提供了很多的帮助和指导，谨此表示感谢！

在论文完成过程中，图书馆庞新国馆长、吴德冈副馆长、信息部的孙明杰主任、李诗平老师、毕硕本副教授、王红林老师、高超老师、王兴同学都给了我很多的宝贵意见，在此一同表示谢意！另外，感谢南京信息工程大学图书馆的全体同事在工作上、生活上为我提供的帮助！

感谢我的爱人和家人，因为有了你们的支持、鼓励和帮助，我的论文才得以顺利完成。

最后，向所有帮助过我的人致以最美好的祝愿！

张海燕

2008．5